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

Abstract. Video reasoning typically operates within the Video Question-Answering (VQA) paradigm, which demands that the models understand and reason about video content from temporal and causal perspectives. Traditional supervised VQA methods gain this capability through meticulously annotated QA datasets, while advanced visual-language models exhibit remarkable performance due to large-scale visual-text pretraining data. Nevertheless, due to potential language bias and spurious visual-text correlations in cross-modal learning, concerns about the reliability of their answers persist in real-world applications. In this paper, we focus on the grounded VQA task, which necessitates models to provide answers along with explicit visual evidence, i.e., certain video segments. As temporal annotation is not available during training, we propose a novel bi-directional reasoning framework to perform grounded VQA in a weakly-supervised setting. Specifically, our framework consists of two parallel but dual reasoning paths. They conduct temporal grounding and answering based on the video content, approaching it from two dual directions that are symmetrical in terms of temporal order or causal relationships. By constructing a cycle-consistency relationship between these two branches, the model is prompted to provide self-guidance supervision for both temporal grounding and answering. Experiments conducted on the Next-GQA and Env-QA datasets demonstrate that our framework achieves superior performance in grounded VQA and can provide reasonable temporal locations that validate the answers.

Keywords: Grounded Video Question Answering · Video Reasoning · Weakly-supervised Learning · Video Grounding

1 Introduction

Video Question Answering (VQA) has emerged as a critical method for assessing the capabilities of multi-modal models [15, 24, 29], aiming to understand

✉ Corresponding author.

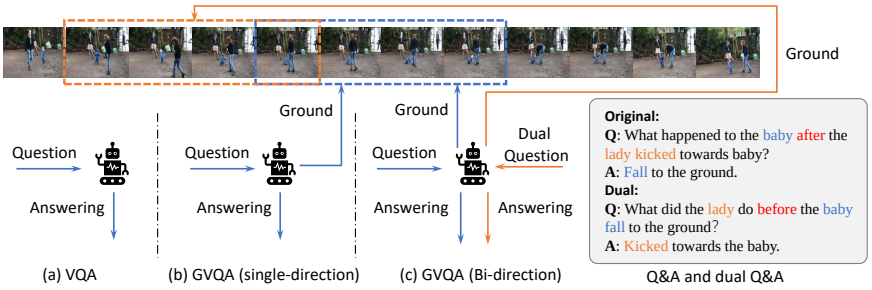

Fig. 1: Comparison between the idea of (a) standard VQA (b) Grounded VQA (c) Our proposed grounded VQA via bi-directional reasoning.

and reason about video content. Thanks to large-scale pretraining on visual-text data [19] and advancements in visual backbones [4], visual-language models have achieved remarkable performance across various VQA tasks. Despite these achievements, it remains an open question whether these models genuinely derive their answers from the relevant video content or if they predominantly rely on *language biases* inherited from large language models [3, 21–23], or the *spurious visual-text correlations* [28] learned during multi-modal pretraining [15, 19, 24].

While the reliability of multi-modality models has been a subject of investigation in image understanding for some time, exploring these concerns within the context of video remains less examined. To this issue, recent developments have introduced the task of visually Grounded Video Question Answering (GVQA) as outlined in NExT-GQA [27]. This task requires models not only to answer questions but also to identify and provide relevant video moments as visual evidence supporting their answers, as illustrated in Fig. 1(b). This benchmark is notably executed under a weakly-supervised setting, where only the validation and test sets come with ground-truth temporal annotations for evaluation.

There are some VQA methods have attempted to offer implicit grounding by applying temporal or spatial attention mechanisms across video frames [7,11–13]. VGT [28] attempted to explicitly ground key frames from videos from a novel causal perspective. Recently, NExT-GQA presented a simple solution, which learning a single Gaussian distribution for answer generation. However, this method falls short in providing directed supervision for accurate moment localization. SeViLA [30] explored the use of BLIP-2 [15] models to generate pseudo-labels for both grounding and answering, enhancing the grounding aspect of VQA. Despite this innovative approach, it necessitates additional, costly pre-training on the TSG video grounding dataset and employs a multi-stage training process. Therefore, how to build up a reliable GVQA model in a weakly-supervised setting still remains challenging.

In this paper, we introduce a novel framework, *TimeCraft*, designed to address the task of weakly-supervised GVQA through bi-directional reasoning (Fig. 1(c)). This framework enables accurate self-supervision for both grounding and answering without the need for additional grounding annotations. Specif-

ically, by leveraging advanced Large Language Models (LLMs), we generate a dual question and answer (Q&A) pair for each ground-truth Q&A in the training set. This dual pair reverses the original Q&A in causal or temporal aspects. Utilizing the original and its dual Q&A pair, we develop a bi-directional reasoning framework comprising an efficient Transformer-based GVQA model and a ***cycle-consistency*** training pipeline. The model first identifies key temporal moments using learnable Gaussian functions. It then grounds these moments to provide answers and reconstructs the dual question from the input question. This process is executed concurrently from two reasoning paths during training, starting from the original and the dual questions, respectively. All intermediate outputs (e.g., grounded moments, reconstructed questions) related to the same concept from both paths are encouraged to be consistent. This self-guided supervision enables our GVQA model to achieve precise grounding and enhanced, more dependable QA performance. We conduct comprehensive experiments on the new GVQA benchmark, NExT-GQA [27], and the traditional VQA dataset, Env-QA. The results highlight the significant effectiveness of our proposed framework.

2 Related work

2.1 Video Question Answering

Video Question Answering (VQA) stands as a cornerstone in the realm of multimodal video understanding. It requires models to deliver precise answers drawn from video content. Traditionally, VQA benchmarks concentrated on fundamental aspects of video content recognition and understanding, including binary selection, counting, and scene recognition tasks. However, the landscape has evolved with the introduction of new benchmarks aimed at evaluating the reasoning capabilities of models, particularly in areas of causal and temporal reasoning [6, 9, 25, 26]. In response, a variety of fully-supervised VQA methodologies have been developed, seeking to improve performance through enhanced context learning, temporal modeling, the implementation of attention mechanisms [7, 11, 14], and the learning of causal representations [28]. The recent breakthroughs achieved by large language models [21, 23] have significantly propelled vision-language models (VLMs) [5, 15, 19, 24, 29] forward, notably improving VQA performance, especially in the context of zero-shot QA settings. Diverging from traditional supervised training methods, these models hone their video reasoning and understanding abilities through the analysis of extensive text-video datasets. Nonetheless, a critical inquiry persists regarding the extent to which the responses generated by such techniques are genuinely grounded in the pertinent video content, rather than being influenced by the linguistic biases inherent in large language models or the spurious vision-language (VL) correlations that may emerge through cross-modal pretraining.

2.2 Weakly-supervised Video Grounding

As annotating temporal boundaries is costly, weakly-supervised video grounding (WSVG) has garnered considerable attention. It focuses on localizing the refer-

ent with only video-level annotations (i.e., language queries). Current WSVG methods fall into two categories: multiple instance learning (MIL) based methods [8, 10, 17, 18] and reconstruction-based methods [16, 20, 31]. MIL-based algorithms align visual-language pairs by attracting matched pairs and repelling mismatched ones. On the other hand, reconstruction-based methods rank proposals based on a reconstruction distance or loss, considering the proposal that best reconstructs the language query as the matching one. Early WSVG approaches depended on sliding windows across the temporal dimension to generate multi-scale proposals, which proved to be computationally expensive. Consequently, more advanced methods have predominantly employed learnable Gaussian functions to generate proposals.

2.3 Weakly-supervised Grounded Video Question Answering

Some VQA [7,11,14] approaches implicitly analyze the reliability of VQA models by employing temporal or spatial attention mechanisms. For instance, MIST [7] attempted to identify key frames at the feature level by framing this process as a spatial-temporal attention mechanism. Similarly, VGT [28] explicitly learned to ground key frames from a causal perspective. Recently, SeViLA [30] integrated BLIP-2 to develop a grounded VQA model. This involves a two-stage training process: the first stage localizes keyframes based on questions and provides answers based on these frames, while the second stage refines the frame localizer with generated pseudo labels for the frames. Moreover, SeViLA undergoes pre-training on large-scale video grounding datasets to acquire strong prior knowledge for grounding. Recently, Xiao et al. [27] expanded the original NExT-QA [26] video dataset into a weakly supervised grounded VQA benchmark, NExT-GQA. This enhanced dataset includes additional temporal labels (start and end timestamps) tied to the questions in the QA pairs for the validation and test sets. Furthermore, it introduces a straightforward solution for grounded VQA, employing MIL learning between constructed positive and negative questions to learn single Gaussian weights. In this paper, we primarily conduct our experiments on the NExT-GQA dataset.

3 Method

3.1 Beyond single-directional Q&A

Standard Q&A pairs typically follow a *single-directional* reasoning process (as shown in Fig. 1(a)(b)), from the question to the answer. However, due to the limited variety of Q&A text combinations, this singular mapping often overlaps significantly with language biases or common sense knowledge [27]. Consequently, VLMs might exploit these shortcuts, relying on biases or leveraging pre-existing knowledge from LLMs to furnish answers. Motivated by this observation, we introduce a novel transformation strategy for Q&A pairs, aiming to expand the reasoning pathways beyond the conventional question-to-answer direction.

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 the boy prevent himself from falling? Touch the front wheel after looking up.*
<Example-2>
Given Q&A: *What did the man in black shirt do as he stood behind the baby? Smile into the camera.*
Generated Q&A: *What did the man in black shirt do as he smile into the camera? Stood behind the baby.*
<Example-3>
Given Q&A: *Where is the man playing this game? Living room.*
Generated Q&A: *What are the man doing in the living room? Playing game.*

Fig. 2: Prompt for LLMs to generation dual Q&A. 3-shot examples are presented here.

Dual Q&A We define the **dual Q&A** pair as a Q&A pair that mirrors the original Q&A pair in causality and temporal structure. For example, the following two Q&A pairs are dual, where both of them focus on the same events while their reasoning path is symmetrical.

```
Q1: Why did the adult move his spoon to the girl's bowl?
A1: Pretend to take her food.
Q2: How did the adult pretend to take the girl's food?
A2: Move his spoon to the girl's bowl.
```

Promoting LLMs for dual Q&A transformation Facilitated by the great text understanding ability of large language models (LLMs), we can easily generate dual Q&A pairs by prompting LLMs. To achieve this, we first employ a *circle-prompting* strategy, initially providing a small set of original pairs and their dual pairs as hints and querying an LLM (e.g., Llama-2) for prompts that can generate such dual Q&A results. We subsequently conduct iterative tests of prompts for better results and manually correct LLM errors in the loop to ensure that our prompts generate accurate dual Q&A pairs. This iterative process seeks to achieve a balance between precision and conciseness. An example prompt we adopted is specified in Fig. 2, where the standard few-shot prompting strategy is utilized. It's worth noting that the process described is not particularly tricky in practice, given that most modern large language models can handle this task with relative ease. In this way, each original ground truth Q and A in training set can be transformed into a dual Q_d and A_d. In this way, we augment the VQA training set into a more complete set in both causality and temporal structure. More results of our generated dual Q&A can be found in the supplementary.

3.2 Bi-directional reasoning

Transformer-based VQA We adopt the widely-used Transformer-based architecture for Video Question Answering (VQA) to construct our framework. Specifically, for a given video v and question q, the model aims to predict the

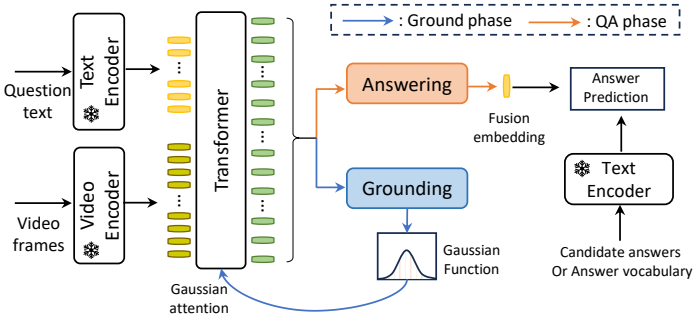

Fig. 3: The architecture of Transformer-based VGQA model.

correct answer a^* from a set of candidate answers A. Initially, the visual and textual inputs are tokenized using pre-trained video and text encoders, respectively. These tokens are then concatenated and input into a multi-modal Transformer network, which extracts essential information from both visual and textual inputs to predict the final answer. This process can be formulated as follows:

$$a^* = \texttt{argmax}_{a \in A} \Omega(a|(v, q), A) \tag{1}$$

As suggested by [27], we adopt a dual-style Transformer as Ω for efficiency.

Ground, then answer We formulate the grounded VQA as two sub-tasks: grounding and QA answering. As for temporal grounding, we adopt the learnable Gaussian function to explicitly learn the moment location t based on the cues provided by the video content v and question q, which can be formulated as:

$$t^* = \mathbf{G}(t|(v, q)) \tag{2}$$

where $\mathbf{G}$ denotes the grounding module. To enable a differentiable end-to-end training, t is represented by a learnable Gaussian function $N(\mu, \sigma^2)$ in the temporal dimension, where μ and σ stands for the mean and standard deviation. In practice, based on the Transformer-based VQA model, we append a Gaussian predictor, which consists of linear transformation layers for efficiency. It receives the fusion multi-modal outputs to predict these two Gaussian parameters. Then, given the predicted Gaussian function, the question and video tokens tied with Gaussian attention weights $t \sim N(\mu, \sigma^2)$ are fed back to a QA round. Specifically, a Gaussian-conditioned attention mechanism [32] is applied to the Transformer for final answering, which plays its role by aggregating contextual video information within the frames highlighted by the Gaussian distribution. In this round, Ω takes a more localized moment to predict the final answer embedding.

Given the above process, our grounded VQA model can be formulated as:

$$a^*, t^* = \texttt{argmax}_{a \in A} \Omega(a|(v_t), q, A) \tag{3}$$

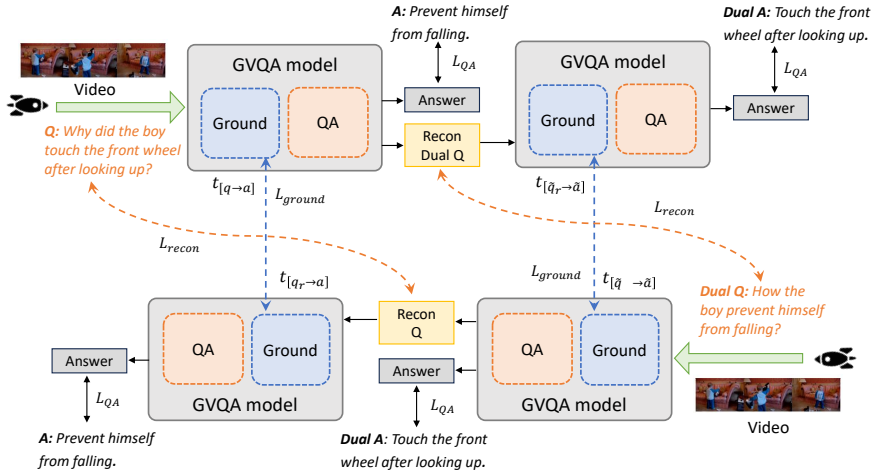

Fig. 4: The overall training pipeline of our bi-directional reasoning framework.

An illustration of the above grounded VQA model is presented in Fig. 3.

Cycle-consistency between bi-reasoning path The most critical problem for weakly-supervised grounded VQA is how to supervise the grounding phase without any temporal labels during training.

In this paper, we address this issue by devising a bi-directional reasoning framework, which achieves a cycle consistency between two dual reasoning paths to obtain reliable supervision for both grounding and QA.

Illustrated in Fig. 4, our training pipeline is structured around two parallel yet counter-directed reasoning paths for each video. The forward path begins with the original question q, and the inverse path with its dual $\widetilde{q}$. Within each path, the GVQA model sequentially executes grounding and then answering processes. Building on the foundational GVQA model, we introduce an auxiliary task named **dual question reconstruction**. This task compels the model to reconstruct the dual question utilizing both the original question and the grounded key moment. This approach not only necessitates the inclusion of cues present in the original answer but also demands an understanding of the deep causal or temporal relationships between the question and answer, moving beyond mere correlation. In practice, we deploy a lightweight linear transformer layer as the reconstruction head, positioned after the final projection layers in Ω. Then, the same GVQA model further receives the reconstructed dual question as input, performing the subsequent GQA process, which outputs grounded moments and a predicted answer for the dual question.

The same "GQA-reconstruction-GQA" process proceeds simultaneously in another reasoning path, where all inputs and intermediate outputs are dual with the former path. Given all intermediate outputs from two reasoning paths, we can set a cycle consistency between the outputs belonging to the same con-

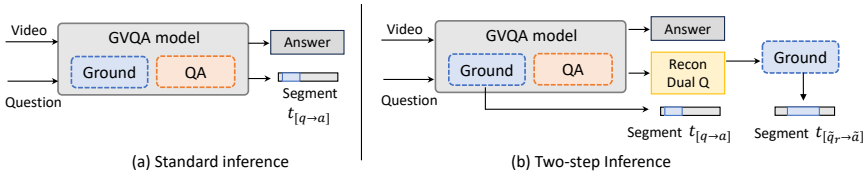

Fig. 5: The illustration of inference phase of our framework.

cept. For example, the moment grounded from the original question to answer and the moment grounded from the reconstructed original question to answer. Specifically, we collect all the intermediate outputs from the forward reasoning path: $(t_{[q\to a]}, a, \widetilde{q}_r, t_{[\widetilde{q}_r\to\widetilde{a}]})$ and from the inverse path: $(t_{[\widetilde{q}\to\widetilde{a}]}, \widetilde{a}, q_r, t_{[q_r\to a]})$, where $t_{q\to a}$ indicates the grounded moment for answering a given the question q, q_r and $\widetilde{q}_r$ indicate the reconstructed question (and dual question), $\sim$ indicates the dual question and answer. The cycle consistency in the bi-directional reasoning framework can be achieved in the following goals.

First, for grounded moment t parameterized by Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, we encourage them to be consistent by minimizing the KL divergence between two Gaussian distributions:

$$\mathcal{L}_{ground} = \mathtt{KL}(t_{[q\to a]} || t_{[q_r\to a]}) + \mathtt{KL}(t_{[\widetilde{q}\to\widetilde{a}]} || t_{[\widetilde{q}_r\to\widetilde{a}]}) \tag{4}$$

where the KL divergence between two Gaussian distributions is calculated as:

$$\mathtt{KL}(t_1 \parallel t_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{5}$$

Note that Eq. (5) is not symmetry and we set the moments grounded from the real q and $\hat{q}$ as the target distributions, i.e., the former item in Eq. (5).

Second, we build the consistency between reconstructed questions and the original one from two paths. It is achieved by minimizing the L2 distance between the text embeddings $\mathbf{f}$ of the real question (or real dual question) and the reconstructed question embedding:

$$\mathcal{L}_{recon} = \mathtt{L2}(\mathbf{f}_{\widetilde{q}}, \mathbf{f}_{\widetilde{q}_r}) + \mathtt{L2}(\mathbf{f}_q, \mathbf{f}_{q_r}) \tag{6}$$

Finally, we performed the standard QA losses $\mathcal{L}_{QA}$ between all predicted answers and their corresponding answer target. For multiple-choice QA, the cross-entropy loss is selected as $\mathcal{L}_{QA}$.

In this way, we could achieve a ***cycle-consistency*** between two dual reasoning paths. Leveraging the generated dual Q&A, the cycle-consistency criterion ensures that both grounding and QA learning processes are self-supervised and self-verified, with reconstruction serving as a pivotal bridge.

3.3 Optimization and inference

Optimization The overall loss for training is formulated as:

$$\mathcal{L} = \mathcal{L}_{QA} + \alpha\mathcal{L}_{recon} + \beta\mathcal{L}_{ground} \tag{7}$$

where α, β are the hyper-parameters that balance the priority of different loss terms. Ablation studies on these hyper-parameters can be found in our supplementary. The overall framework is trained in an end-to-end manner.

Inference During inference, the bi-directional reasoning process is not necessary and we directly feed the given question and video into a single GVQA model Ω to generate results (Fig. 5(a)). Therefore, our proposed training framework will not impact the overall inference efficiency. The grounding is achieved by the confidence interval $t = (\mu - \gamma\sigma, \mu + \gamma\sigma) * d$, where hyper-parameters γ control the width of the confidence interval, d is the total duration of the video. Moreover, by virtue of our modular pipeline, besides the standard inference, we could obtain different styles of grounded results by appending a grounded module (sharing the same weight) to the standard inference pipeline (as shown in Fig. 5(b)). Please see Sec. 4.3 for more experiments and discussion.

4 Experiments

4.1 Experimental setting

Implementation details In theory, our framework can be built upon most of the pre-trained multi-modal Transformers. For fair comparisons, we adopt the Temp[CLIP] [27] and FrozenBiLM [29] as our backbones following [27]. For each video, we sample 32 frames uniformly for VQA models. AdamW is utilized to optimize the model training, with an initial learning rate 1e-5. Early stopping is also adopted if the results on the validation set do not increase in 5 epochs. The overall training process contains 30 epochs and is conducted on A6000 GPUs. The details of the dual Q&A pairs generation can be found in our supplementary.

Dataset We mainly evaluate our framework on the recently proposed grounded VQA benchmark NExT-GQA [27]. Due to the lack of more available GVQA datasets, we further evaluate our method on VQA dataset Env-QA [6].

Evaluation metrics We adopt four metrics for GVQA evaluation following [27]: (1)**mIoP**: It estimates whether the predicted temporal window lies inside the ground truth. (2)**mIoU**: It calculates the temporal overlap between the predicted moment and the ground truth. (3)**Acc@GQA**: A new metric proposed in [27] for grounded GQA task, which inspects the percentages of questions that are correctly answered and also visually grounded (i.e., IoP $\geq$0.5). (4)**Acc@QA**: Standard metric for QA task, which denotes the percentage of correctly answered questions. For both mIoP and mIoU, results are reported with the mean values and values with thresholds of 0.3 and 0.5.

Table 1: Grounded VQA results on NExT-GQA test set. "CM" indicates whether the backbone model is pre-training on cross-modal data. BT: BERT. RBT: RoBERTa. FT5: FLAN-T5-XL [2], DBT: DeBERTa-V2-XL. For a fair comparison, we *de-emphasize* the method that were pre-trained on large-scale video grounding annotations. "Model" indicates the multi-modal Transformer model.

Method	Model	T Enc	V Enc	Acc@GQA	Acc@QA	mIoP	IoP@0.3	IoP@0.5	mIoU	IoU@0.3	IoU@0.5
Human	-	-	-	82.1	93.3	72.1	91.7	86.2	61.2	86.9	70.3
Random	-	-	-	1.7	20.0	21.1	20.6	8.7	21.1	20.6	8.7
SeViLA	BLIP-2	FT-5	ViT-G	16.6	68.1	29.5	34.7	22.9	21.7	29.2	13.8
IGV	-	BT	ResNet	10.2	51.3	21.4	26.9	18.9	14.0	19.8	9.6
MIST	Temp[CLIP]	BT	ViT-L	12.1	60.5	23.6	29.3	20.7	11.4	16.3	7.0
PH	VGT	RBT	RCNN	14.4	57.7	25.3	26.4	25.3	3.0	3.6	1.7
	VLOLETv2	BT	VSWT	12.8	57.2	23.6	25.1	23.3	3.1	4.3	1.3
	Temp[CLIP]	RBT	ViT-L	15.2	62.5	25.4	28.2	25.5	6.6	9.3	4.1
	FrozenBiLM	DBT	ViT-L	15.8	71.8	22.7	25.8	22.1	7.1	10.0	4.4
NG+	Temp[CLIP]	RBT	ViT-L	16.0	63.3	25.7	31.4	25.5	12.1	17.5	8.9
	FrozenBiLM	DBT	ViT-L	17.5	73.1	24.2	28.5	23.7	9.6	13.5	6.1
TimeCraft	Temp[CLIP]	RBT	ViT-L	18.2	65.6	**28.1**	**35.1**	**27.8**	**15.6**	**21.2**	**9.6**
(Ours)	FrozenBiLM	DBT	ViT-L	**18.5**	**74.7**	26.3	32.7	24.9	13.2	18.6	8.4

4.2 Main results

Competitors We compare our framework with three GVQA approaches: (1) SeViLA [30], (2) Post-hoc(PH) [27], (3) NG+ [27]. Additionally, to provide a comprehensive comparison, we further implement the advanced VQA method MIST [7] into the GVQA setting, we consider its selected segment with the highest score as the grounded moment.

Comparison results and analysis Table 1 presents the GVQA results on the NExT-GQA dataset. Analysis of the performance highlights that while advanced pre-trained Visual Language Models (VLMs), such as FrozenBiLM and VIOLETv2, demonstrate commendable QA performance, their grounding capabilities (e.g., mIoU, Acc@GQA) are notably lacking. This underscores a preference for relying on common sense knowledge derived from pretraining data rather than engaging with the relevant visual content tied to the Q&A, aligning with our initial hypothesis and motivation.

Our model, TimeCraft, sets a new benchmark by achieving superior GQA performance and elevating QA accuracy to a new state-of-the-art. A particularly notable advancement is observed in the mIoU metric, signifying that our bi-directional reasoning framework effectively directs the QA model to prioritize the pertinent temporal moments providing visual evidence. However, the extent of improvement varies across different multimodal-transformer models. For instance, Temp[CLIP], a dual-transformer model trained from scratch on the NExT-GQA dataset, exhibits significant gains in QA accuracy parallel to its grounding performance enhancements. Conversely, FrozenBiLM, which largely retains weights from a pre-trained LLM and has been pre-trained on cross-modal data, shows a modest improvement in QA, especially when compared to Temp[CLIP]. It is also critical to note that although FrozenBiLM secures the

Table 2: Performance on NExT-GQA dataset of different types of questions. Note that NExT-GQA has removed the descriptive question in its test set.

Method		Causal				Temporal			
		Acc@GQA	Acc@QA	mIoP	mIoU	Acc@GQA	Acc@QA	mIoP	mIoU
NG+	Temp[CLIP]	17.3	64.7	27.8	12.2	14.1	61.8	22.9	11.8
	FrozenBiLM	19.5	78.4	25.6	9.1	15.3	68.0	22.8	10.4
TimeCraft	Temp[CLIP]	19.0	66.5	**30.2**	**17.4**	**18.0**	64.7	**26.2**	13.8
	FrozenBiLM	**19.7**	**79.5**	27.2	11.3	17.3	**70.1**	25.4	**15.1**

Table 3: QA accuracies on Env-QA test set. Question types that require temporal understanding are emphasized. Results of TimeCraft is reported with Temp[CLIP].

Method	Attribute	State	Event	Order	Number	All
ST-VQA [11]	41.66	48.98	33.87	54.09	38.54	41.97
STAGE [14]	39.49	49.93	34.52	55,32	37.98	42.53
AIO [24]	41.78	52.98	37.57	55.16	38.50	44.86
Temp[ATP] [1]	42.87	53.49	38.35	55.25	38.65	45.43
TSEA [6]	42.96	56.73	39.84	55.53	39.35	47.06
MIST[CLIP] [7]	44.05	58.13	42.54	56.83	40.32	48.97
TimeCraft (Ours)	43.74	**58.20**	**43.86**	**58.37**	39.94	**49.02**

highest Acc@GQA, this achievement is primarily attributed to its QA capabilities and does not necessarily indicate superior performance in VGQA.

Per-category results To evaluate the efficacy of our proposed framework comprehensively, we conducted performance tests across different types of Q&A, with the results summarized in Tab. 2. It is noteworthy that our framework outperforms NG+ with a significant margin in both causal and temporal reasoning, which inherently demands precise temporal content for accurate responses. This performance not only showcases the superior capabilities of our framework but also reinforces its effectiveness in guiding the model to identify and leverage key moments crucial for providing correct answers.

Effectiveness on standard VQA Given the scarcity of VGQA datasets, we extended the application of our framework to a standard VQA dataset to assess its adaptability and effectiveness. The results on the Env-QA dataset, as detailed in Tab. 3, demonstrate that TimeCraft facilitates improvements across various question types, with particularly notable enhancements in event and ordering reasoning. This underscores the value of integrating moment grounding into the VQA model, showcasing its potential to elevate QA performance. Additionally, it's observed that our framework contributes to incremental improvements even in question types less reliant on grounding (e.g., attribute), highlighting its broad applicability and the nuanced benefits it offers to general VQA tasks.

4.3 Ablation study

Breakdown ablation To study the effectiveness of different components within our bi-directional reasoning framework, we trained our model using various com-

Table 4: The breakdown analysis of each part in our framework. "Dual Q&A" indicates whether we generate dual Q&A pairs for original Q&A.

Setting			Acc@GQA	Acc@QA	mIoP	mIoU
Dual Q&A	$\mathcal{L}_{recon}$	$\mathcal{L}_{ground}$				
			15.5	59.4	25.8	7.7
✓			15.8	60.8	26.1	9.2
✓	✓		16.7	62.5	26.8	12.1
✓		✓	15.6	59.8	25.5	8.4
✓	✓	✓	**18.2**	**65.6**	**28.1**	**15.6**

Table 5: Performance on NExT-GQA dataset using different grounded outputs.

Setting	Causal			Temporal		
	Acc@GQA	mIoP	mIoU	Acc@GQA	mIoP	mIoU
(I) Answer-critical	**66.5**	**30.2**	17.4	**64.7**	**26.2**	**13.8**
(II) Question-critical	64.2	28.7	15.9	60.2	21.7	10.0
(III)Union	66.3	29.1	**18.6**	63.8	26.5	12.9

binations of losses and detailed the results in Tab. 4. Notably, even when disabling all optimization objectives for grounding and reconstruction, merely integrating the dual Q&A into standard QA training enhances the baseline. This suggests that the generation of dual Q&A questions can serve as an effective augmentation strategy for VQA tasks, mitigating the spurious correlations introduced by static QA pairs. Regarding the reconstruction and grounding losses, we observed that the former plays a more critical role in improving performance. This aligns with our expectations, as the alignment of grounded moments from both reasoning paths heavily depends on the accurate reconstruction of dual questions. By implementing all optimization objectives, our framework demonstrated superior performance, underscoring the importance of complete cycle consistency within our bi-directional reasoning framework.

Grounding question or answer? As discussed in Sec. 3.3 and Fig. 5, we introduce variations to the standard inference path by incorporating a grounding module, yielding three distinct and plausible grounded outcomes:(i) The standard results, identified as answer-critical moments, (ii) Moments determined by the second grounder, tagged as question-critical moments, (iii) The union of areas from both (i) and (ii), which synthesizes information from both the question and answer. We evaluated the efficacy of these grounded outcomes on the test set, categorizing the results by question types in Tab. 5. The evaluation reveals notable differences across various metrics. Interestingly, outcome (iii) achieves the best mIoU in the causal reasoning category, which often requires comprehensive information from both the question and answer for more accurate moment grounding. This finding underscores the potential of our framework in effectively grounding both question text and answer evidence. Further visualization results related to this observation are available in our supplementary materials.

Table 6: Results on NExT-GQA with leveraging different LLMs to generate dual QA. All results are reported with TimeCraft (Temp[CLIP]).

LLM	Acc@GQA	Acc@QA	mIoP	mIoU
Llama-2 7B [23]	17.3	64.0	26.4	34.2
Llama-2 13B [23]	**18.2**	**65.6**	**28.1**	35.1
Gemini [21]	17.8	65.0	26.9	**35.3**
InternLM [22]	17.5	64.8	27.6	34.8

Table 7: The results of incorporating the dual QA as augmented data for training. All results are reported with Temp[CLIP].

Setting	Acc@GQA	Acc@QA	mIoP	mIoU
PH	15.2	62.5	25.4	7.1
+ Dual QA	15.8	63.2	25.8	8.7
NG+	16.0	63.3	25.7	12.1
+ Dual QA	16.3	64.0	26.1	12.9
TimeCraft	18.2	65.6	28.1	15.6

Impact of LLMs for dual Q&A generation We generate dual Q&A pairs by prompting advanced LLMs. In this section, we explore how the choice of LLM influences our final results. Specifically, we considered Llama-2, Gemini, and InternLM as candidates. Results are summarized in Tab. 6. Contrary to expectations, incorporating dual Q&A generated by Llama-2 achieves the best, despite Llama-2 being the smallest LLM among candidates. Upon analyzing the outputs from different LLMs, we observed that Gemini and InternLM possess a broader base of common sense knowledge than Llama-2. This often leads them to introduce potentially irrelevant content into the generated dual Q&A, which could be detrimental to our cycle-consistency training. Therefore, we recommend Llama-2 as a sufficiently effective and more accessible choice for this task.

Benefit of dual Q&A We further test the effectiveness of generated dual Q&A pairs by directly adding them to the original training set and conducting standard QA training. Experiments in Tab. 7 prove that the NG+ can be further improved by incorporating our generated dual Q&A. It shows that the dual Q&A generation can also be considered an effective augmentation strategy for VQA.

How does the sampled frame number impact GQA? We study the impact of input frame length and present the results in Fig. 7. We can see that most metrics grow with the increase of the frame number, while mIoP and mIoU decrease when increasing the frame number to 128.

4.4 Optimization and inference

Efficiency analysis Our TimeCraft involves bi-directional reasoning during training, which introduces twice-forward passes. Therefore, one of the limitations of our proposed framework is the increase in training time. To quantify this issue, we report our model efficiency in Fig. 8. It can be seen that the training time of our framework is still acceptable considering the improvements. Moreover, most of the other metrics still remain efficient as the backbone model.

Visualization We visualize some good and failed cases predicted by TimeCraft in Fig. 6. We can see that the Temp[CLIP] tends to present a better performance than FrozenBiLm, which is consistent with the quantization result.

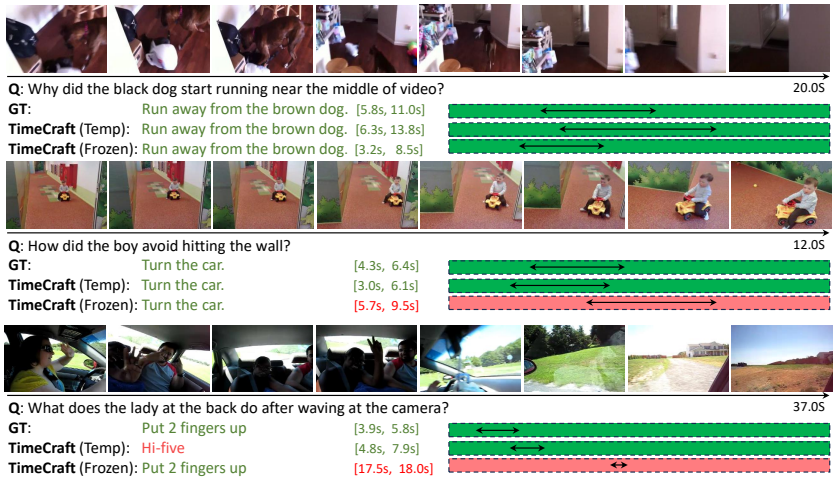

Fig. 6: Visualization results on NExT-GQA test set. Green indicates the correct answer or grounding, while wrong predictions are in red.

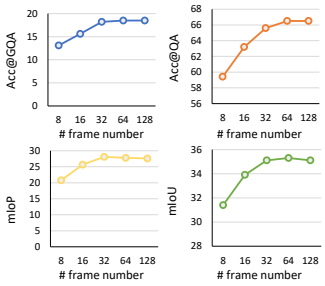

Fig. 7: Impact of sampled frame number (from 8 to 128).

Fig. 8: The efficiency analysis of our proposed framework. Value of time is based on 1 epoch in training&test, reported with 2× A6000 GPU.

Model	Train Param.	Infer Param.	Model Size	Train Time	Infer Time
TEMP[CLIP]	130.3M	130.3M	0.5G	1.8m	9.0s
+TimeCraft	130.6M	130.6M	0.5G	3.0m	9.6s
FrozenBiLM	29.7M	1.2B	3.8G	0.3h	1.0m
+TimeCraft	43.9M	1.2B	3.8G	1.7h	1.8m

5 Conclusion

This paper addresses the challenge of grounded VQA by proposing a bi-directional reasoning framework *TimeCraft*. Our framework consists of dual reasoning paths for temporal grounding and answering, and establishing cycle consistency for self-supervision. It presents superior grounded VQA performance on the Next-GQA and Env-QA datasets, accurately providing temporal evidence for its answers.

Acknowledgements

The paper is supported in part by the National Natural Science Foundation of China (No. 62325109, U21B2013) and the Lenovo Academic Collaboration Project.

References

1. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the" video" in video-language understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2917–2927 (2022)
2. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: An empirical study of end-to-end video-language transformers with masked visual modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22898–22909 (2023)
6. Gao, D., Wang, R., Bai, Z., Chen, X.: Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1675–1685 (2021)
7. Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., Shou, M.Z.: Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14773–14783 (2023)
8. Gao, M., Davis, L.S., Socher, R., Xiong, C.: Wslln: Weakly supervised natural language localization networks. arXiv preprint arXiv:1909.00239 (2019)
9. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11287–11297 (2021)
10. Huang, J., Liu, Y., Gong, S., Jin, H.: Cross-sentence temporal and semantic relations in video activity localisation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7199–7208 (2021)
11. Jang, Y., Song, Y., Kim, C.D., Yu, Y., Kim, Y., Kim, G.: Video question answering with spatio-temporal reasoning. International Journal of Computer Vision **127**, 1385–1412 (2019)
12. Jiang, J., Chen, Z., Lin, H., Zhao, X., Gao, Y.: Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11101–11108 (2020)
13. Jin, W., Zhao, Z., Li, Y., Li, J., Xiao, J., Zhuang, Y.: Video question answering via knowledge-based progressive spatial-temporal attention network. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **15**(2s), 1–22 (2019)
14. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvqa+: Spatio-temporal grounding for video question answering. arXiv preprint arXiv:1904.11574 (2019)
15. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

16. Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11539–11546 (2020)
17. Ma, M., Yoon, S., Kim, J., Lee, Y., Kang, S., Yoo, C.D.: Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. pp. 156–171. Springer (2020)
18. Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11592–11601 (2019)
19. 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)
20. Song, Y., Wang, J., Ma, L., Yu, Z., Yu, J.: Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. arXiv preprint arXiv:2003.07048 (2020)
21. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
22. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM` (2023)
23. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
24. Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K.Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., et al.: All in one: Exploring unified video-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6598–6608 (2023)
25. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
26. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
27. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. arXiv preprint arXiv:2309.01327 (2023)
28. Xiao, J., Zhou, P., Chua, T.S., Yan, S.: Video graph transformer for video question answering. In: European Conference on Computer Vision. pp. 39–58. Springer (2022)
29. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems **35**, 124–141 (2022)
30. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. Advances in Neural Information Processing Systems **36** (2024)
31. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3517–3525 (2022)

32. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3517–3525 (2022)