What You Ask is What You Ground: Bridging Question Intent to Temporal Evidence for Grounded VideoQA

Jinhwan Seo[1], Kyubeom Han[1], Jumin Lee[1], Junhyug Noh[2†], and Sung-eui Yoon[1†]

[1] KAIST
[2] Ewha Womans University
{jinhwan.seo,qbhan,jmlee}@kaist.ac.kr, junhyug@ewha.ac.kr, sungeui@kaist.ac.kr

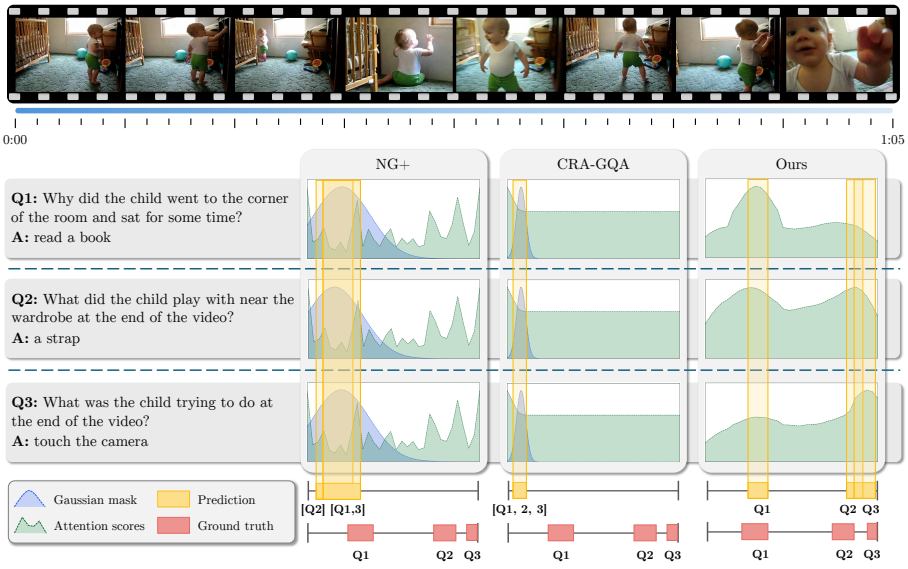

Fig. 1: Given a single video and three distinct questions targeting different temporal moments, prior methods yield nearly identical temporal predictions regardless of question intent. In contrast, GroundFormer produces three distinct segments, each aligned with the event required to reason toward the corresponding answer.

Abstract. We study a critical yet overlooked failure mode in Grounded Video Question Answering: *question-invariant grounding*, where models predict nearly identical temporal segments for different questions about the same video. We trace this behavior to two structural limitations in prior common designs: (i) *modality isolation* that fixes video representations before they receive question semantics, and (ii) *weak question injection* inside the grounding module. To address this, we propose

†Corresponding authors.

GroundFormer, which conditions video features on question intent *before* localization via learnable communication tokens that mediate directed visuo-lingual interaction. On top of the question-conditioned features, a factorized MIL cross-attention couples answer selection with temporal evidence under candidate-level supervision, while Gaussian smoothing converts peaked attention into temporally coherent segments. We further introduce a hierarchical multi-modal contrastive loss that aligns video, question, and answer embeddings across a two-pass training pipeline. GroundFormer achieves state-of-the-art grounded VideoQA performance on NExT-GQA and STAR, substantially improving question-discriminative temporal grounding.

Keywords: Grounded VideoQA · Question-Conditioned Grounding · Weakly Supervised Temporal Localization

1 Introduction

Video Question Answering (VideoQA) [2, 4, 12, 28] has emerged as a primary benchmark for evaluating the multi-modal reasoning capabilities of vision-language models (VLMs) [1, 5, 14, 15, 22, 26]. Recent advances in large-scale pretraining of large language models [6,10,17] have driven remarkable progress on the VideoQA task, yet the BlindQA experiment [29] reveals that models without video input achieve performance on par with state-of-the-art methods, suggesting that predictions may reflect linguistic priors rather than visual understanding. This experiment motivates Grounded VideoQA [29], which requires models to produce not only a correct answer but also the temporal segment from which it is derived, ensuring that answers are grounded in visual evidence. However, we identify a tendency that persists across existing methods [3, 16, 29]: *question-invariant grounding*, where models produce nearly identical temporal predictions for different questions within the same video.

As illustrated in Fig. 1, when multiple questions target different temporal events within the same video, existing methods such as NG+ [29] and CRA-GQA [3] collapse all predictions onto the same segment regardless of question intent. Each question asks about a distinct moment, yet the predicted grounding intervals are nearly indistinguishable across questions, revealing that previous methods [3, 29] fail to condition on what each question asks. We attribute such phenomena to two structural limitations: visual representations are fixed before any question signal reaches them, and the question is reduced to a single token at the grounding stage that is immediately absorbed by the dominant visual stream. As a result, the grounding module operates on representations that are blind to what each question asks.

In this work, we propose **GroundFormer**, a framework that addresses *question-invariant grounding* by establishing a visuo-lingual communication pathway through which question intent is propagated into visual representations. To counter modality isolation, we introduce learnable communication tokens that act as a bottleneck between the visual and linguistic branches: the tokens extract fine-grained

question semantics from the linguistic stream and inject the captured intent into visual encoding. To overcome the insufficient language signal in the grounding module, we formulate Grounded VideoQA as a Multiple Instance Learning (MIL) problem via cross-attention, where question-conditioned communication tokens serve as queries and video tokens as keys and values, such that temporal evidence is derived from question intent under candidate-level supervision alone.

In summary, our contributions are:

- We introduce learnable *communication tokens* as an explicit bottleneck between the visual and linguistic branches, enabling question intent to shape video representations *before* grounding and alleviating modality isolation.
- We cast Grounded VideoQA as a *multiple instance learning (MIL)* problem and propose a MIL cross-attention that couples answer prediction with temporal evidence, producing question-conditioned localization from candidate-level QA supervision alone.
- We achieve state-of-the-art Acc@GQA and grounding performance on NExT-GQA and STAR, and PIoU and GT-Pred Correlation analyses confirm that our grounding is more question-discriminative than prior methods.

2 Related Work

VideoQA. With the success of vision-language models (VLMs) [5, 13–15, 22], VideoQA has advanced from simple action recognition to tasks requiring reasoning over event sequences, causal relationships, and temporal dependencies across video [4, 12, 20, 24, 28]. Early approaches employed RNNs and CNNs [12, 23, 33], while recent methods adopted transformer-based architectures [23, 30] and Q-Former frameworks [13, 34] that couple frozen visual encoders with large language models (LLMs) via learnable queries. Each generation improves QA capability, yet whether these gains reflect visual understanding or linguistic priors inherited in the LLM backbone remains unresolved [29]. This ambiguity motivates Grounded VideoQA, which requires not only a correct answer but also the temporal evidence from which it is derived.

Grounded VideoQA. NExT-GQA [29] formalizes visually grounded VideoQA by introducing temporal grounding as an explicit evaluation criterion under weak supervision, where no grounding annotations are available during training. Subsequent methods tackle this challenge from complementary perspectives. TimeCraft [16] enforces cycle consistency between causally reversed Q&A pairs as a self-supervised grounding signal, while CRA-GQA [3] applies causal intervention to disentangle visual and linguistic confounders. QGAC-TR [31] calibrates temporal attention via contrastive objectives over visual and textual context; on the generative side, Grounded-VideoLLM [25] augments the vocabulary with discrete temporal tokens, and TOGA [9] couples timestamp prediction with open-ended answer generation under weak supervision. Despite these efforts, our analysis shows that existing methods often produce identical temporal grounding for different questions within the same video, which we attribute to late fusion of independently encoded modalities. GroundFormer addresses this question-invariant

Table 1: Question-invariant grounding analysis on NExT-GQA. Lower PIoU indicates less collapse across questions; higher GT–Pred Corr. indicates better agreement with the ground-truth pairwise overlap structure.

Metric	NG+ [29]	CRA-GQA [3]	**GroundFormer**	Ground Truth
PIoU ↓	85.7	88.9	**28.2**	21.3
GT–Pred Corr. ↑	0.005	0.006	**0.121**	–

grounding by introducing learnable tokens for stronger visuo-lingual alignment under QA weak supervision.

Weakly Supervised Learning for Temporal Grounding. In the absence of frame-level annotations, Multiple Instance Learning (MIL) [7] treats each video as a bag of temporal instances and identifies relevant segments from video-level supervision alone. Prior MIL-based methods [8, 11, 19, 21] learn segment-level representations by contrasting matched and mismatched visual-language pairs, achieving competitive performance in video moment retrieval [19, 21] and video grounding [8, 11]. However, these methods operate on declarative queries (*e.g.*, "A person throws a ball") that explicitly describe the target event, whereas Grounded VideoQA poses interrogative queries (*e.g.*, "Why did the person throw the ball?") whose supporting evidence is only implicitly defined by the question–answer relationship. Our approach is the first to adopt a MIL-based formulation for Grounded VideoQA, coupling answer selection with temporal localization through a cross-attention mechanism.

3 Motivation

Observation. Most existing Grounded VideoQA methods [3,16,29] share a common grounding structure built upon a Gaussian-based localization module. Yet, as Fig. 1 illustrates, the previous methods often show *question-invariant grounding*, where models produce nearly identical temporal predictions regardless of the questions. To quantify this behavior, we introduce two diagnostic metrics. Pairwise IoU (PIoU) measures the average temporal IoU between predictions for different questions within the same video; high PIoU indicates invariant grounding, while low PIoU indicates question-discriminative grounding. GT-Pred Correlation (GT-Pred Corr.) complements PIoU by measuring whether predicted overlaps vary in the same direction as ground-truth overlaps, computed as the Pearson correlation between ground-truth and predicted pairwise IoU across the dataset.

We evaluate PIoU and GT-Pred Corr. on NG+ [29] and CRA-GQA [3] in Tab. 1. Both methods record PIoU of 85.7 and 88.9, far above the ground-truth PIoU of 21.3, while their GT-Pred Corr. remains near zero (0.005 and 0.006). These results indicate that their grounding predictions are nearly indistinguishable across questions, motivating our analysis of the common grounding structure underlying this failure.

Preliminary. Given an untrimmed video v, a question q, and answer candidates $A = \{a_1, \ldots, a_{A_n}\}$, a vision encoder extracts frame-level features $x_v \in \mathbb{R}^{B\times T\times D_V}$ from T sampled frames, where B is the batch size and D_V the visual feature dimension. A language encoder produces embeddings $x_l \in \mathbb{R}^{B\times L_n\times L_l\times D_L}$, where L_n denotes the number of candidates (question or answer) and L_l is the maximum token length, and D_L the language feature dimension. The pooled question and answer embeddings, denoted x_q and x_a, respectively, are projected to a shared dimension D. The methods then solve:

$$a^*, t^* = \arg\max_{a\in A} \; \Psi(a \mid v_t, q, A) \;\; \Phi(t \mid v, q)\,, \tag{1}$$

where Φ is a *grounding module* that estimates a temporal grounding distribution t given (v, q), and Ψ predicts the answer from the localized evidence v_t.

A representative grounding module $\Phi(t \mid v, q)$ injects the question into the visual stream as a *single* token. Let x_{q_n} denote the question token; it is concatenated with T visual tokens to form $[x_v;\ x_{q_n}] \in \mathbb{R}^{B\times(T+1)\times D}$, which is then sum-pooled to $\mathbb{R}^{B\times D}$ and passed through Φ to regress Gaussian parameters for temporal grounding:

$$\begin{aligned} (\mu, \sigma^2) &= \Phi(\texttt{pool}([x_v;\ x_{q_n}]))\,, \qquad & \mathbf{g}_{\text{gauss}} &= \mathcal{N}(\mu, \sigma^2), \\ \mathbf{g}_{\text{attn}} &= \texttt{Linear}(\mathbf{g}_{\text{gauss}} \odot x_v)\,, \qquad & t^* &= (\mathbf{g}_{\text{attn}} \cap \mathbf{g}_{\text{gauss}})\,. \end{aligned} \tag{2}$$

Discussion. Based on our formulation of the common grounding structure in Eq. (2), we identify two structural limitations that make the grounding module Φ prone to question-invariant behavior.

1. **Modality isolation before Φ.** Visual and linguistic branches encode independently, and cross-modal interaction occurs only at a late stage. Consequently, frame features carry only limited information about *what the question asks*, leaving the grounding module Φ to operate on question-agnostic video features.
2. **Insufficient linguistic signal inside Φ.** The question is injected as a single token among T visual tokens in $[x_v;\ x_{q_n}] \in \mathbb{R}^{B\times(T+1)\times D}$ and is immediately collapsed by pooling. With a 1:T imbalance, the question signal is easily absorbed by the dominant visual stream, making the input to Φ weakly dependent on the question and falling back to question-invariant grounding.

These two limitations compound: modality isolation removes question semantics from the visual features, while the weak linguistic signal prevents Φ from recovering question dependence later. As quantified in Tab. 1, this drives existing methods toward question-invariant grounding. GroundFormer breaks this pattern by conditioning visual representations on question semantics before Φ, thereby restoring question-discriminative grounding (Fig. 2).

4 Approach

We propose GroundFormer, a framework that conditions visual encoding on question semantics prior to the grounding module. The framework operates as

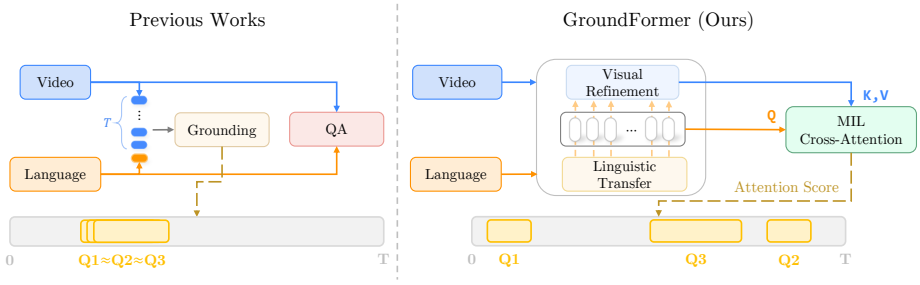

Fig. 2: Overview of GroundFormer. Left: Prior methods [3, 16, 29] weakly inject question information into the grounding module, often collapsing different questions to the same salient interval. **Right:** GroundFormer conditions video features on question semantics via communication tokens (*Linguistic Transfer* → *Visual Refinement*) and couples answer selection with temporal grounding using MIL Cross-Attention (token queries, video keys/values), producing question-discriminative segments.

a two-pass pipeline that separates question understanding from answer prediction (Sec. 4.1). Within each pass, learnable communication tokens bridge the two modalities via Linguistic Transfer and Visual Refinement, and MIL cross-attention derives a question-conditioned grounding signal from answer supervision alone; Gaussian smoothing then refines sparse attention peaks into temporally coherent segments (Sec. 4.2). A two-pass pipeline further encourages question-discriminative visual representations during training (Sec. 4.3), and the model is optimized with classification and hierarchical multi-modal contrastive objectives (Sec. 4.4).

4.1 Two-Pass Pipeline

GroundFormer performs two passes over shared weights. The question pass takes question candidates as input, constructed by pairing the ground-truth question with questions sampled from other videos in the mini-batch as negatives, and identifies the most video-relevant question, encoding visual discriminativity with respect to question semantics. The second pass takes answer candidates as input and produces the answer prediction $\hat{a}$ along with the grounding signal. This disentanglement allows the hierarchical multi-modal contrastive loss to encourage question-discriminative visual representations. Note that despite the two-pass pipeline, the design introduces no additional computational cost at inference, as the two passes share weights and only the answer pass executes at test time.

4.2 Architecture

GroundFormer consists of two core components: learnable communication tokens that transfer question semantics into visual representations, and MIL cross-attention that derives a question-conditioned grounding signal from answer supervision. Fig. 3 illustrates the overall architecture and attention flow.

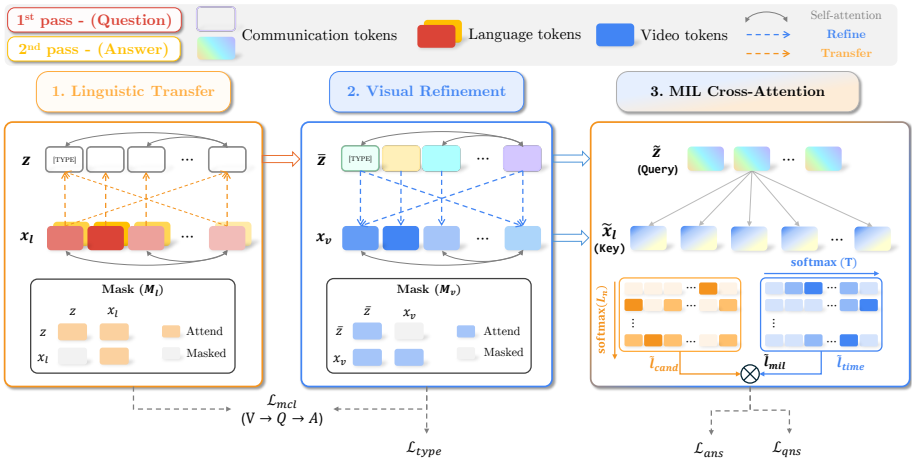

Fig. 3: Details of GroundFormer. (1) **Linguistic Transfer:** communication tokens z attend to language tokens x_l under an asymmetric mask M_l to absorb candidate semantics. (2) **Visual Refinement:** video tokens x_v attend to the distilled tokens $\bar{z}$ under M_v, producing question-conditioned features $\tilde{x}_v$ while keeping tokens uncontaminated. (3) **MIL Cross-Attention:** token queries attend to video keys/values to obtain attention logits, which are factorized into candidate and temporal distributions ($\tilde{\ell}_{cand}$, $\tilde{\ell}_{time}$) whose product yields MIL scores and an implicit grounding signal. The model is trained with losses for question matching (1st pass) and answer prediction (2nd pass), together with type and multi-modal contrastive objectives.

Communication Tokens. GroundFormer introduces a set of communication tokens $z = [z_{\texttt{type}};\ z_{\texttt{query}}] \in \mathbb{R}^{B\times(1+N)\times D}$ as a bottleneck between two modalities. The [TYPE] token $z_{\texttt{type}} \in \mathbb{R}^{B\times 1\times D}$ encodes a coarse question category and is supervised by an auxiliary loss (Sec. 4.4); this coarse prior guides the remaining tokens to specialize across reasoning patterns. The N query tokens $z_{\texttt{query}} \in \mathbb{R}^{B\times N\times D}$ capture fine-grained question semantics and propagate the captured intent to the visual representation through two steps: *Linguistic Transfer* and *Visual Refinement*.

Linguistic Transfer first transfers candidate semantics into the communication tokens via asymmetric masked self-attention. A transformer layer concatenates z with x_l along the token dimension, allowing z to attend to both z and x_l while x_l attends only to itself:

$$\bar{z} = \texttt{SA}_1([z;\ x_l], M_l)\big|_z. \tag{3}$$

The asymmetric mask ensures that $\bar{z}$ captures candidate semantics while preventing the communication tokens from corrupting the language representations.

Visual Refinement propagates the captured intent $\bar{z}$ into the visual stream via a second asymmetric masked self-attention layer. Concatenating x_v with $\bar{z}$,

the layer allows x_v to attend to both x_v and $\bar{z}$ while $\bar{z}$ attends only to itself

$$\tilde{x}_v = \texttt{SA}_2([x_v;\ \bar{z}], M_v)\big|_{x_v}, \qquad \tilde{z} = \texttt{SA}_2([x_v;\ \bar{z}], M_v)\big|_{z}. \tag{4}$$

Each frame in $\tilde{x}_v$ now encodes candidate-specific information, resolving the modality isolation. The asymmetric mask keeps $\tilde{z}$ independent of visual content, and the updated $\tilde{z}_{\texttt{query}}$ serves as queries for the subsequent grounding step.

Weakly-Supervised Grounding via MIL Cross-Attention. We generate a grounding signal from QA supervision alone using MIL cross-attention mechanism. We stack candidates along the L_n dimension, yielding intent queries $\tilde{z}_{\texttt{query}} \in \mathbb{R}^{B\times L_n\times N\times D}$ and question-conditioned visual tokens $\tilde{x}_v \in \mathbb{R}^{B\times L_n\times T\times D}$. We compute cross-attention logits:

$$\begin{aligned} \tilde{\ell} &= \frac{(\tilde{z}_{\texttt{query}} W_Q)\cdot(\tilde{x}_v W_K)^{\intercal}}{\sqrt{D_k}} \in \mathbb{R}^{B\times L_n\times N\times T}, \\ \tilde{\ell}_{time} &= \texttt{softmax}_T(\tilde{\ell}), \qquad \tilde{\ell}_{cand} = \texttt{softmax}_{L_n}(\tilde{\ell}), \\ \tilde{\ell}_{mil} &= \sum_{t=1}^{T}\Big(\tilde{\ell}_{time}\odot\tilde{\ell}_{cand}\Big) \in \mathbb{R}^{B\times L_n\times N}. \end{aligned} \tag{5}$$

Here, $\tilde{\ell}_{time}$ distributes probability over frames for each candidate, acting as the temporal evidence signal, while $\tilde{\ell}_{cand}$ distributes probability over candidates at each frame. Their element-wise product assigns a high score to a candidate only when it is supported by frames that also receive high temporal attention; averaging over time yields the MIL logit $\tilde{\ell}_{mil}$. This factorization binds the candidate score to its supporting frames, so $\tilde{\ell}_{time}$ serves as a grounding signal derived from answer supervision. The cross-attention output aggregates visual evidence weighted by $\tilde{\ell}_{time}$, and average pooling over the N query tokens produces $\hat{z} \in \mathbb{R}^{B\times L_n\times D}$, which is fed to the classifier heads.

Grounding Refinement via Gaussian Kernel. MIL Cross-Attention produces a candidate-conditioned temporal distribution $\tilde{\ell}_{time}$. In the answer pass, we take the temporal attention associated with the predicted answer $\hat{a}$ as the raw grounding signal $\mathbf{g}_{attn} \in \mathbb{R}^{B\times T}$. MIL factorization often concentrates attention on the most discriminative frames, resulting in sharply peaked signals that can fragment the predicted temporal segment. We therefore apply a 1-D Gaussian kernel to smooth $\mathbf{g}_{attn}$ into a temporally continuous segment.

$$\mathcal{G}_\sigma(k) = \exp\!\left(-\frac{k^2}{2\sigma^2}\right), \qquad \mathbf{g}_{gauss} = \mathcal{G}_\sigma * \mathbf{g}_{attn}, \tag{6}$$

where k indexes the kernel position, $*$ denotes 1-D convolution over the T frames, and σ controls the smoothing extent. We use $\mathbf{g}_{gauss}$ to derive the final grounded segment.

4.3 Hierarchical Multi-modal Contrastive Learning

Unlike declarative queries that directly describe a target event, interrogative queries in Grounded VideoQA require an intermediate reasoning step: the model

must first identify the relevant visual evidence before arriving at the answer. This two-step reasoning naturally motivates a hierarchical alignment structure across the two-pass pipeline. In the first pass, video embeddings are aligned with question embeddings (V→Q), grounding visual representations in question semantics. In the second pass, question embeddings are aligned with answer embeddings (Q→A), binding the answer to the question-conditioned visual evidence. Together, V→Q→A alignment pulls video, question, and answer embeddings into a coherent embedding space progressively, using in-batch candidates as negative samples.

4.4 Training Objectives

GroundFormer is trained end-to-end with candidate-level supervision only. Each pass is supervised by standard cross-entropy over its candidates using the pooled representation $\hat{z}$. We additionally apply cross-entropy on the MIL logits $\tilde{\ell}_{mil}$ to ensure that the factorized scoring aligns with the ground-truth candidate:

$$\begin{aligned} \mathcal{L}_{qns} &= \texttt{CE}(f_Q(\hat{z}_Q), y_q) + \texttt{CE}\Big(\tilde{\ell}^Q_{mil}, y_q\Big)\,, \\ \mathcal{L}_{ans} &= \texttt{CE}(f_A(\hat{z}_A), y_a) + \texttt{CE}\Big(\tilde{\ell}^A_{mil}, y_a\Big)\,, \end{aligned} \tag{7}$$

where f_Q and f_A are the question and answer classifier heads, y_q and y_a are the ground-truth indices.

The hierarchical multi-modal contrastive loss aligns embeddings progressively across the two passes. Let $\tilde{x}^i_v \in \mathbb{R}^D$ denote the pooled video embedding from the question pass for the i-th sample in a mini-batch of size B, $\bar{x}^{+,i}_q \in \mathbb{R}^D$ its ground-truth question embedding, and τ a temperature parameter. All $B \times Q_n$ question embeddings in the mini-batch form the contrastive pool, where $\bar{x}^{j,m}_q$ is the m-th question-candidate embedding of the j-th sample. So the (V→Q) alignment is:

$$\mathcal{L}_{vq} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\text{sim}(\tilde{x}^i_v,\, \bar{x}^{+,i}_q)\,/\,\tau\big)}{\sum_{j=1}^{B}\sum_{m=1}^{Q_n} \exp\big(\text{sim}(\tilde{x}^i_v,\, \bar{x}^{j,m}_q)\,/\,\tau\big)}. \tag{8}$$

Using the same anchor $\bar{x}^{+,i}_q$, the (Q→A) alignment is defined against all $B \times A_n$ answer embeddings:

$$\mathcal{L}_{qa} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\text{sim}(\bar{x}^{+,i}_q,\, \bar{x}^{+,i}_a)\,/\,\tau\big)}{\sum_{j=1}^{B}\sum_{m=1}^{A_n} \exp\big(\text{sim}(\bar{x}^{+,i}_q,\, \bar{x}^{j,m}_a)\,/\,\tau\big)} \tag{9}$$

$$\mathcal{L}_{mcl} = \mathcal{L}_{vq} + \mathcal{L}_{qa} \tag{10}$$

The [TYPE] is supervised with an auxiliary classification loss:

$$\mathcal{L}_{type} = \texttt{CE}(f_T(\tilde{z}_{type}), y_{type})\,, \tag{11}$$

where f_T is a linear classifier and y_{type} is the question category label provided by each benchmark [27, 28].

The total loss combines all objectives:

$$\mathcal{L} = \mathcal{L}_{ans} + \mathcal{L}_{qns} + \mathcal{L}_{mcl} + \mathcal{L}_{type}. \tag{12}$$

5 Experiment

5.1 Experimental Setup

Datasets. We evaluate GroundFormer on two popular grounded VideoQA benchmarks. NExT-GQA [29] targets causal and temporal reasoning through human-annotated questions, while STAR [27] targets situated reasoning through programmatically generated questions.

- **NExT-GQA** [29] extends NExT-QA [28] with temporal grounding annotations for the validation and test splits (10.5K annotated QA pairs), enabling evaluation of both answer correctness and evidence localization. Compared to short-clip VideoQA benchmarks [4, 12, 20] (∼10 s), it contains longer untrimmed videos (avg. ∼44 s) and emphasizes causal and temporal reasoning under weak supervision, spanning five question types: Causal-Why, Causal-How, Temporal-Before&After, Temporal-When, and Temporal-Present.
- **STAR** [27] is a situated reasoning benchmark with 4,901 videos and 60,206 QA pairs, each annotated with a temporal segment. Videos average ∼30 s with a segment-to-video ratio of ∼0.4, roughly twice that of NExT-GQA. Questions are categorized into four types: Interaction, Sequence, Prediction, and Feasibility.

Evaluation Metrics. We report metrics for both answering and grounding. For QA, we report Acc@VQA, the percentage of correctly answered questions. For localization, we report mIoU and mIoP between the predicted and ground-truth temporal segments. Since Grounded VideoQA requires both a correct answer *and* faithful grounding, we use Acc@GQA as the primary metric. Following NExT-GQA, Acc@GQA counts a prediction as correct only if the answer is correct and the grounded segment achieves IoP ≥ 0.5.

Implementation Details. We implement GroundFormer in PyTorch. For video representation, we uniformly sample T=32 frames per video and extract features using a frozen CLIP ViT-L/14 vision encoder [22]. For text, we use a frozen RoBERTa-base encoder [17] to embed questions and answer candidates. The GroundFormer block consists of 4 transformer layers. We set the number of learnable query tokens to N=32 and the Gaussian kernel to σ=2. We train for 30 epochs with batch size 16 using AdamW [18], an initial learning rate of 10^{-5}, and a cosine decay schedule. All experiments are conducted on RTX 3090 GPUs.

Baselines. We compare our method against grounding methods: PH [29], NG+ [29], TimeCraft [16], and CRA-GQA [3], while varying their base architectures with different capacities: Temp[CLIP] [29] (130–145M) and FrozenBiLM [32] (1.2B). We also compare our method with QGAC-TR [31] using its own architecture for Grounded VideoQA.

Table 2: Comparison with state-of-the-art methods on NExT-GQA [29].

Arch.	Method	Param.	**Acc@GQA**	Acc@VQA	mIoP	TIoP@0.3	TIoP@0.5	mIoU	TIoU@0.3	TIoU@0.5
Temp[CLIP] [29]	PH [29]	130M	15.2	59.4	25.4	28.2	25.5	6.6	9.3	4.1
	NG+ [29]	130M	16.0	60.2	25.7	31.4	25.5	12.1	17.5	8.9
	TimeCraft [16]	130M	18.2	**65.6**	28.1	35.1	27.8	15.6	21.2	9.6
	CRA-GQA [3]	145M	18.2	61.1	28.6	34.3	28.5	14.2	21.4	10.6
FrozenBiLM [32]	PH [29]	1.2B	15.8	69.1	22.7	25.8	22.1	7.1	10.0	4.4
	NG+ [29]	1.2B	17.5	70.8	24.2	28.5	23.7	9.6	13.5	6.1
	TimeCraft [16]	1.2B	18.5	74.7	26.3	32.7	24.9	13.2	18.6	8.4
	CRA-GQA [3]	1.2B	18.8	70.2	26.5	32.6	25.9	13.5	20.1	9.6
QGAC-TR [31]		–	18.3	63.6	28.3	32.8	27.7	15.7	18.6	11.7
GroundFormer		211M	**21.5**	61.7	**34.0**	**41.5**	**33.7**	**17.5**	**26.5**	**12.8**

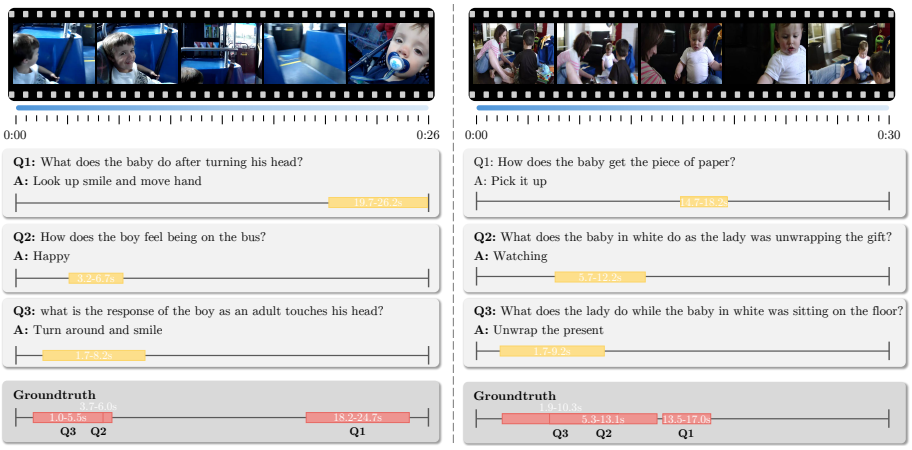

Fig. 4: Qualitative comparison on NExT-GQA. For each video, we show three questions (Q1–Q3) with their answers and the predicted grounding intervals, along with the ground-truth segments (bottom). GroundFormer produces distinct, question-conditioned temporal segments that follow the ground-truth ordering, whereas prior methods collapse different questions onto similar intervals.

5.2 Comparison With State-of-the-Art Methods

We demonstrate the robustness of GroundFormer on two Grounded VideoQA benchmarks [27, 29]. Our framework achieves new state-of-the-art Acc@GQA and temporal grounding across both datasets regardless of network capacity.

NExT-GQA. Tab. 2 compares GroundFormer with prior methods on the NExT-GQA test set. GroundFormer achieves the best Acc@GQA of 21.5, setting a new state of the art with consistent gains in localization quality across all grounding metrics and thresholds. While FrozenBiLM variants attain higher Acc@VQA (up to 74.7) due to their larger capacity, GroundFormer yields substantially stronger grounded performance: it improves Acc@GQA (21.5 vs. 18.8), mIoP (34.0 vs. 26.5), and mIoU (17.5 vs. 13.5) over the best FrozenBiLM baseline, despite using only 17.6% of its parameters.

Table 3: Comparison with state-of-the-art methods on STAR [27].

Arch.	Method	**Acc@GQA**	Acc@VQA	mIoP@0.5	mIoU@0.5
TempCLIP [29]	NG+ [29]	24.4	57.3	41.4	4.7
	CRA-GQA [3]	26.8	58.6	44.5	5.5
FrozenBiLM [32]	NG+ [29]	25.8	60.1	40.9	7.8
	CRA-GQA [3]	27.8	60.5	43.1	5.1
GroundFormer		**30.7**	**61.1**	**49.3**	**15.8**

Table 4: Ablation on NExT-GQA. We factorize design choices into structural components (two-pass training, communication tokens) and auxiliary components.

Structure		**Auxiliary Components**				**QA**		**Localization (IoP)**			**Localization (IoU)**		
Two-pass train	Comm. tokens	$\mathcal{L}_{type}$	$\mathcal{L}_{mcl}$	$\mathcal{G}_{\sigma}$	MIL	**Acc@GQA**	Acc@VQA	mIoP	TIoP@0.3	TIoP@0.5	mIoU	TIoU@0.3	TIoU@0.5
✓						14.8	58.5	26.5	33.3	24.8	14.7	21.9	10.6
✓	✓					18.6	59.8	29.6	37.0	28.6	17.0	25.8	13.0
✓	✓	✓				19.0	60.4	30.5	37.9	30.2	16.9	25.8	13.2
✓	✓	✓	✓			20.1	61.2	32.0	39.5	30.3	17.2	25.8	12.5
✓	✓	✓	✓	✓		20.8	61.2	32.4	39.3	32.0	16.5	25.0	12.0
	✓			✓	✓	15.7	58.2	26.7	32.8	25.9	14.6	21.8	10.8
✓	✓	✓	✓	✓	✓	**21.5**	**61.7**	**34.0**	**41.5**	**33.7**	**17.5**	**26.5**	**12.8**

Fig. 4 further illustrates the benefit of GroundFormer's question-conditioned grounding capability. In the left example, three questions target temporally distinct events: Q3 asks about a reaction near the beginning, Q2 about a feeling in the middle, and Q1 about an action toward the end. GroundFormer localizes each question to a different temporal region that aligns with the corresponding ground-truth segment. In the right example, three questions again refer to separate moments, and GroundFormer produces three non-overlapping predictions that match the ground-truth ordering. In both cases, NG+ and CRA-GQA collapse all predictions onto a single interval regardless of the question. Thus, our result confirms that GroundFormer ensures correct answers are derived from accurately localized temporal evidence compared to baseline methods.

STAR. Tab. 3 demonstrates the generalization capability of GroundFormer on STAR benchmark. Compared to existing methods under both TempCLIP [29] and FrozenBiLM [32] architectures, GroundFormer establishes a new state-of-the-art with an Acc@GQA of 30.7. The improvement is driven by robust temporal grounding, as shown by an mIoP@0.5 of 49.3 and mIoU@0.5 of 15.8.

5.3 Analysis

How does each module contribute? Tab. 4 summarizes how each design choice contributes to question-conditioned grounding on NExT-GQA. Starting from a baseline that uses a standard cross-attention grounding head (no communication tokens or auxiliary objectives), we obtain 14.8 Acc@GQA and 26.5 mIoP. Introducing communication tokens with *Linguistic Transfer* and *Visual Refinement* yields a large jump to 18.6 Acc@GQA and 29.6 mIoP, confirming

Table 5: Acc@GQA (%) by question type on NExT-GQA.

Method	Causal		Temporal		
	Why	How	Before & After	When	Present
NG+ [29]	16.9	17.2	12.0	17.5	10.8
GroundFormer (Ours)	**22.2**	**23.1**	**17.3**	**25.8**	**12.9**

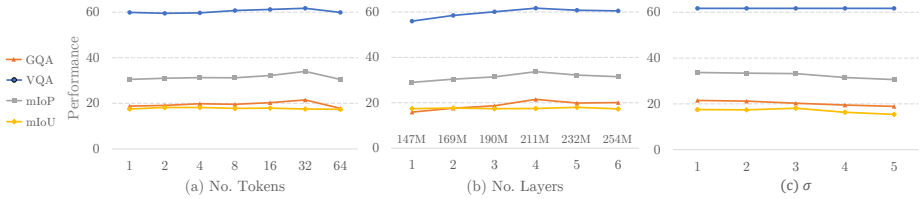

Fig. 5: Hyperparameter sensitivity on NExT-GQA. (a) Number of learnable query tokens N, (b) number of Transformer layers in the GroundFormer block, and (c) Gaussian smoothing width σ. We report Acc@VQA, Acc@GQA, mIoP, and mIoU.

that conditioning video features on question semantics *before* localization is critical. Adding our auxiliary objectives provides further gains: question-type supervision and the hierarchical contrastive loss improve Acc@GQA to 19.0 and 20.1, respectively, indicating that coarse reasoning priors and cross-modal alignment are complementary. Finally, adding the localization heads (Gaussian smoothing and MIL) strengthens localization and leads to the best overall performance.

We also ablate the *two-pass training* strategy in Tab. 4. The second-to-last row reports a *single-pass* variant trained using only the answer pass; in this setting, the question-pass objectives (*e.g.*, $\mathcal{L}_{qns}$, $\mathcal{L}_{type}$ and the V→Q term in $\mathcal{L}_{mcl}$) are not applicable. This single-pass training results in 15.7 Acc@GQA and 26.7 mIoP. In contrast, enabling two-pass training by adding the question pass activates the full hierarchical objectives and improves performance to 21.5 Acc@GQA and 34.0 mIoP (last row). Importantly, this gain comes without inference overhead, since only the answer pass is executed at test time.

How does question type shape grounding? Tab. 5 reports Acc@GQA performance separately for each question types defined in NExT-QA [28]: Causal (Why, How) and Temporal (Before & After, When, Present). GroundFormer outperforms NG+ across all five types, confirming that our question conditioning improves grounding across question types. In addition, we found that the largest gains appear on Temporal-When (+8.3) and Causal-How (+5.9), both of which demand precise temporal discrimination to locate a specific moment.

Hyperparameter sensitivity. Fig. 5 reports performance under different numbers of learnable tokens N, Transformer layers, and Gaussian kernel width σ. The Acc@GQA and mIoP peak at $N = 32$, after which additional tokens yield diminishing returns while increasing parameter count. Performance increases steadily up to 4 layers and saturates thereafter. For σ, the optimum lies at σ=2;

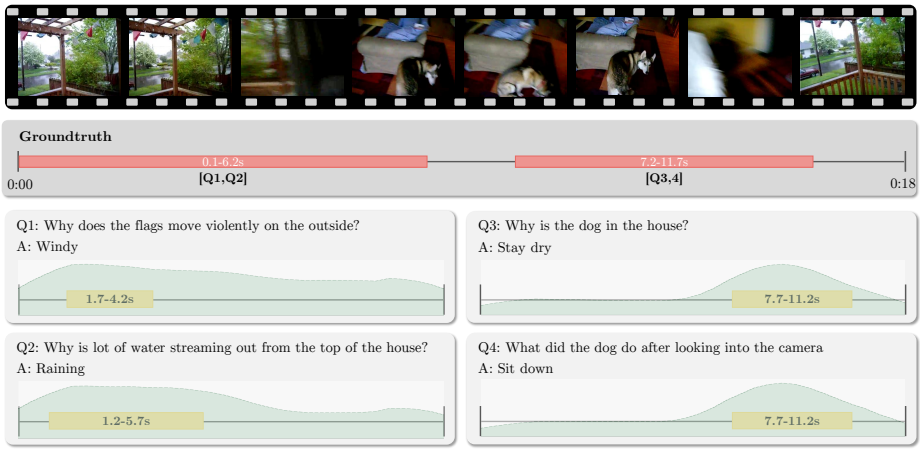

Fig. 6: Grounding for semantically similar questions. In the same video, GroundFormer groups Q1–Q2 (outdoor weather) into an early interval and Q3–Q4 (indoor dog event) into a later interval, matching the ground-truth split. Shaded spans indicate predicted grounded segments.

larger values over-smooth the temporal signal and degrade mIoU. We adopt $N = 32$, 4 layers, and $\sigma = 2$ throughout all other experiments.

How does grounding behave for similar questions? Fig. 6 shows that GroundFormer produces *consistent* temporal grounding for semantically related questions while remaining *discriminative* across different events within the same video. Q1 (windy flags) and Q2 (rain streaming from the roof) both refer to the outdoor weather, and GroundFormer grounds them in the early portion of the video (0.1–6.2 s), consistent with the ground-truth intervals (Q1: 1.7–4.2 s; Q2: 1.2–5.7 s). In contrast, Q3 (dog stays indoors to stay dry) and Q4 (dog sits after looking into the camera) target the later indoor event, and the predicted grounding shifts to 7.2–11.7 s, aligning with the shared ground-truth segment (7.7–11.2 s). These results suggest that GroundFormer's temporal grounding follows question semantics rather than collapsing onto a single salient moment.

6 Conclusion

We identified *question-invariant grounding* as a key failure mode in weakly-supervised Grounded VideoQA: existing methods produce nearly identical temporal predictions regardless of question intent, a consequence of modality isolation and weak question injection in the grounding module. To address this, we proposed GroundFormer, which conditions visual tokens on question semantics before localization via asymmetric masked attention and couples answer selection with temporal grounding through MIL Cross-Attention. To further bind grounding to question intent, we introduce a hierarchical multi-modal contrastive

objective that aligns video, question, and answer embeddings progressively along a V→Q→A. As a result, GroundFormer achieves state-of-the-art Acc@GQA on both NExT-GQA and STAR, with PIoU and GT-Pred Correlation analyses confirming substantially more question-discriminative grounding.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World) and the NRF grant (No. RS-2023-00208506). Junhyug Noh was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2026-25499022), and Global – Learning & Academic research institution for Master's · PhD students, and Postdocs (G-LAMP) Program of the NRF funded by the Ministry of Education (RS-2025-25442252).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Chen, W., Liu, Y., Chen, B., Su, J., Zheng, Y., Lin, L.: Cross-modal causal relation alignment for video question grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24087–24096 (2025)
4. Choi, S., On, K.W., Heo, Y.J., Seo, A., Jang, Y., Lee, M., Zhang, B.T.: Dramaqa: Character-centered video story understanding with hierarchical qa. In: Proceedings of the aaai conference on artificial intelligence. vol. 35, pp. 1166–1174 (2021)
5. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
7. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)
8. Gao, M., Davis, L., Socher, R., Xiong, C.: Wslln: Weakly supervised natural language localization networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 1481–1487 (2019)
9. Gupta, A., Roy, A., Chellappa, R., Bastian, N.D., Velasquez, A., Jha, S.: Toga: Temporally grounded open-ended video qa with weak supervision. arXiv preprint arXiv:2506.09445 (2025)
10. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020)
11. 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)
12. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017)
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
16. Liu, H., Ma, X., Zhong, C., Zhang, Y., Lin, W.: Timecraft: Navigate weakly-supervised temporal grounded video question answering via bi-directional reasoning. In: European Conference on Computer Vision. pp. 92–107. Springer (2024)

17. Liu, Y.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 **364** (2019)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. 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: European conference on computer vision. pp. 156–171. Springer (2020)
20. Min, J., Buch, S., Nagrani, A., Cho, M., Schmid, C.: Morevqa: Exploring modular reasoning models for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13235–13245 (2024)
21. 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)
22. 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)
23. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)
24. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
25. Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., Huang, L.: Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290 (2024)
26. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
27. 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 (NeurIPS) (2021)
28. 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)
29. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13204–13214 (2024)
30. 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)
31. Xu, Y., Wei, Y., Zhong, S., Chen, X., Qi, J., Wu, B.: Exploring question guidance and answer calibration for visually grounded video question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3121–3133 (2024)
32. 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)
33. Yin, C., Tang, J., Xu, Z., Wang, Y.: Memory augmented deep recurrent neural network for video question answering. IEEE transactions on neural networks and learning systems **31**(9), 3159–3167 (2019)

34. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)