Wavelet-based Intra-video Counterfactual Reasoning for Video Question Grounding

JiaSheng Yuan[1,2] and Wei Wei[1,2⋆]

[1] School of Computer Science and Technology, Huazhong University of Science and Technology, P.R.China
[2] Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL), P.R.China
{yuanjs314,weiw}@hust.edu.cn

Abstract. Video Question Grounding (VideoQG) requires models to answer questions and localize supporting temporal evidence. Despite progress in video-language models, they remain susceptible to various biases. Prior debiasing methods primarily focus on dataset-level correlations, while easily overlooking intra-video temporal bias, where visually similar segments differ substantially in their causal relevance to answering the question. To address this challenge, we present Wavelet-based Intra-video Causal Intervention (WICI), a framework designed to disentangle causal temporal evidence from redundant visual contexts. Our WICI comprises two key components: (1) an Intra-video Causal Intervention module, which utilizes counterfactual reasoning to suppress irrelevant intra-video biases, and (2) Wavelet-based Dynamic Modeling, which decomposes contextual variations between visually similar frames to capture fine-grained temporal cues. Extensive experiments on two VideoQG benchmarks demonstrate that the proposed WICI significantly improves question grounding accuracy and yields more robust, faithful question reasoning.

Keywords: Video Question Grounding · Video Understanding · Weakly-supervised Learning

1 Introduction

Video Question Answering (VideoQA) commonly requires models to reason over temporal visual content to answer natural language questions grounded in video content. Recent advances in pre-trained vision-language models [19, 20, 35, 37] have improved multiple VideoQA benchmarks [13, 43, 46, 47, 49]. However, these gains may be unreliable, as they often exploit dataset biases or spurious correlations for high performance [9, 43, 45]. To better assess genuine reasoning ability, the Video Question Grounding (VideoQG) task has been introduced [18] to jointly predict the correct answer and localize supporting temporal evidence, while further exploring weakly-supervised settings [45].

⋆ Corresponding author.

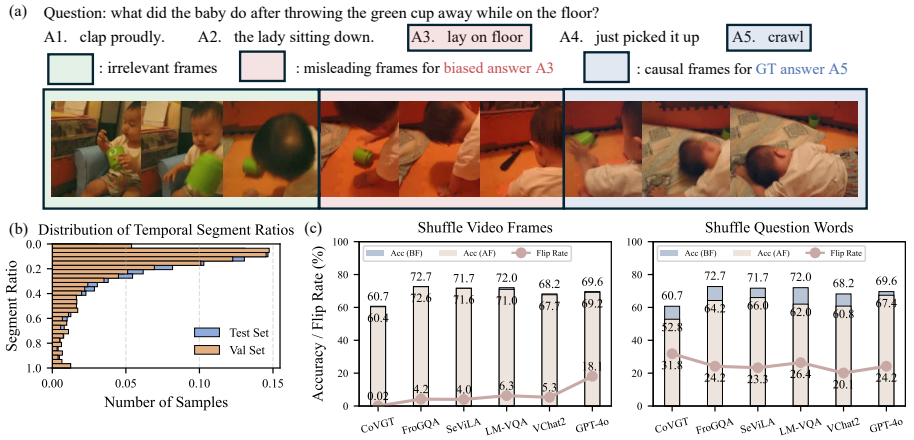

Fig. 1: (a) An example video–question pair, where frames are categorized as evidence supporting (red), irrelevant (green), and misleading (blue). Despite their similar visual appearances, these frames contribute differently to answer prediction. (b) Distribution of temporal segment ratios on the NeXT-GQA dataset. (c) Robustness analysis under temporal perturbations, including shuffled video frames and shuffled question words.

However, accurately localizing evidential segments is particularly susceptible to spurious correlations, as models may exploit biased visual or temporal cues rather than perform genuine reasoning. Existing de-biasing methods primarily target dataset-level bias [3, 5, 32, 50], but VideoQA also suffers from intra-video distractors arising from temporal bias: Unlike images, videos inherently exhibit repeated or similar visual content over time, but the semantic contributions of these frames to question-answering can vary substantially [13, 51]. As illustrated in Fig. 1(a), the frames highlighted in red contain the evidence required to correctly answer the question, whereas the green frames are irrelevant and the blue frames even provide misleading visual cues. Notably, in Fig. 1(b), such irrelevant or misleading segments often dominate the temporal span of a video in NeXT-GQA dataset [45], making the model disproportionately exposed to non-causal content. Although these segments serve different semantic roles, they remain visually similar and share the same entities (*e.g.* the baby, the green cup, the background). When processed by a pre-trained vision–language encoder [14, 19, 35], whose representations are predominantly appearance-oriented, video segments from different temporal stages but sharing similar visual content become less distinguishable in the embedding space, thereby amplifying temporal bias [31, 48]. This observation also helps explain a counterintuitive phenomenon reported in [43] and illustrated in Fig. 1(c): answer accuracy appears insensitive to temporal perturbations, as models struggle to distinguish between visually similar frames in the first place. From the perspective of Structural Causal Models (SCM) [34], the video representation is jointly influenced by causally relevant evidence and non-causal distractors. While only the causal

subset should dictate the final answer, existing models implicitly entangle causal and non-causal factors within the global video representation. This entanglement introduces spurious non-causal dependencies that severely compromise grounding accuracy and reasoning fidelity. Disentangling these factors requires moving beyond static appearance features to capture the subtle temporal dynamics, such as motion patterns and state transitions, that distinguish causal actions from visually identical, yet non-causal states. Furthermore, it necessitates an explicit learning mechanism to isolate the causal impact of specific temporal segments on the model's decision-making process.

To address these challenges, we propose a unified framework that tackles intra-video temporal bias at both the representation and reasoning levels. Specifically, we introduce a Wavelet-based Intra-video Causal Intervention (WICI) approach. First, to enhance fine-grained temporal discriminability, we propose a Wavelet-based Dynamic Modeling (WDM) module inspired by signal processing [8, 30, 41]. WDM decomposes frame-level features to isolate high-frequency inter-frame dynamics from low-frequency appearance contexts, enriching the representations with the subtle motion cues essential for distinguishing causal segments. Second, we design an Intra-video Counterfactual Mining (ICM) module to explicitly sever the spurious link between non-causal content and the predicted answer. ICM constructs curriculum-guided [2,36], intra-video counterfactual segments that progressively increase in difficulty. By applying a margin-based objective [1] to compare the network's predictive distributions across causal, counterfactual, and reference segments, the model learns to isolate genuine evidence based on its causal contribution rather than superficial visual similarity. We summarize our main contribution as follows:

1. We identify and formulate the problem of intra-video temporal bias in Video Question Grounding, highlighting the vulnerability of current models to visually similar but non-causal distractors.

2. We introduce WICI to address intra-video temporal bias from both representation and decision level. A WDM module decomposes video features into multi-scale dynamic components to enhance inter-frame discriminability, and an ICM module introduces quantifiable supervision between causal and counterfactual predictions, encouraging the model to distinguish causally relevant frames.

3. Extensive experiments on two VideoQG benchmarks demonstrate that the proposed WICI framework significantly outperforms existing baselines, yielding substantial improvements in both visual grounding accuracy and causal reasoning robustness.

2 Related Work

2.1 Grounded Video Question Answering

Recent advancements in Video Question Answering (VideoQA) with temporal grounding aim to jointly predict correct answers and localize relevant video segments that support the reasoning process. Conventional approaches typically employ Cross-Modal Attention(CMA) between video frames and question tokens to

classify grounding intervals [4,6,15,17]. Methods like [7,16,21,28] further improve grounding performance by leveraging pretrained vision-language models and fine-grained cross-modal reasoning. However, temporal boundary annotations are costly to obtain, motivating recent efforts toward weakly-supervised settings where such annotations are not available during training. In weakly-supervised sentence grounding, prior studies commonly resort to multiple instance learning (MIL) [12, 29, 54] or reconstruction-based [24, 53] objectives to introduce indirect temporal supervisory signals. Similarly, recent efforts in weakly-supervised VideoQG [45] attempt to compensate for the lack of temporal supervision by mining contrastive video-language pair [5, 45], generating pseudo labels [10, 49] or constructing bidirectional question-answer pair [25], thereby providing auxiliary supervision. Nevertheless, existing methods rely heavily on the priors of LLMs or pretrained vision–language encoders that emphasize appearance-level semantics, while overlooking both the discriminative signals embedded in model response distributions and the critical role of subtle temporal dynamics in distinguishing causal from non-causal video segments.

2.2 Causal Learning for Multimodal Reasoning

Most existing multi-modal methods rely on learning cross-modal correlations; however, such correlation-driven reasoning is insufficient for tasks(e.g. VideoQA) involving causal dependencies, leading to unfaithful predictions due to model bias. Recent studies have increasingly attributed model bias to spurious correlations induced by dataset-specific priors, and accordingly introduce causal intervention mechanisms [33] to mitigate such effects, including backdoor adjustment [38, 39], front-door intervention [26], and counterfactual reasoning [32, 52]. [39] introduced Intervention Effect(IE) to estimates causal contributions by comparing answer distributions before and after causal interventions. While effective, IE primarily enforces relative ordering among predictions through inequality-based objectives, without explicitly regulating the magnitude of separation between different temporal segments. In VideoQG, prior works have further explored causal modeling to address spurious temporal correlations. IGV [23], EIGV [22], and VCSR [40] partition video content into causal and non-causal components conditioned on the question, and employ invariant or contrastive learning objectives to enforce their separation. MCR [50] further localizes causal visual evidence at the object level. CRA [5] combines back-door intervention on the textual modality with front-door intervention on the visual modality. Despite their effectiveness, most existing de-biasing methods in VideoQG focus on representation-level separation or prediction invariance between causal and non-causal segments. However, within a single video, temporally adjacent segments often exhibit highly similar content, and enforcing such rigid constraints can introduce ambiguous supervision and unstable optimization. In contrast, we address temporal bias from a decision-centric perspective, modeling causal relevance as a relative preference rather than requiring strict invariance or explicit disentanglement.

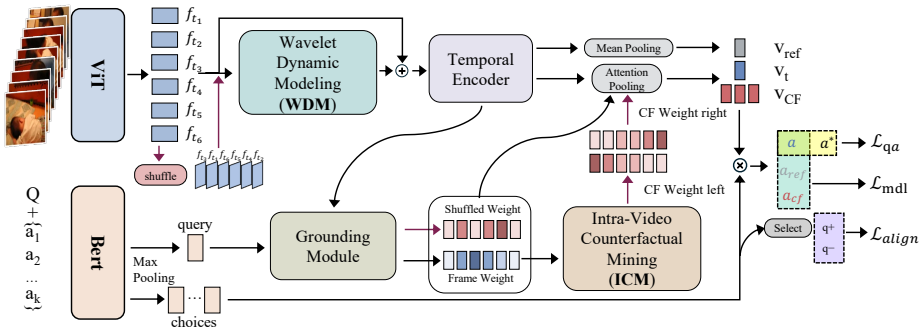

Fig. 2: An overview of WICI framework. Visual and textual features are extracted independently. The WDM module captures frame-wise temporal dynamics and, together with a temporal encoder, models global temporal dependencies. A shuffled-video branch is introduced as counterfactual input and processed in parallel. Query-conditioned frame attention identifies causal segments, which are used by the ICM module to construct curriculum-guided counterfactual samples. The model is trained with cross-entropy loss on causal predictions and margin-based discrimination against full-video and counterfactual predictions, along with contrastive learning on selected linguistic samples.

3 Method

To enhance faithful VideoQG reasoning under weak supervision, we propose WICI, a unified framework consisting of two components: (1) Intra-video Counterfactual sample Mining (ICM) module and (2) Wavelet-based Dynamic Modeling (WDM) module. The WDM module first decomposes video representations into high-frequency dynamic details and low-frequency approximations, thereby injecting complementary temporal dynamics into video features. The ICM module then generates curriculum-guided intra-video counterfactual segments, enabling the model to better discriminate causal evidence among video frames.

As illustrated in Fig. 2, we begin by sampling n frames evenly from each video and extracting their visual embeddings using a pre-trained CLIP encoder [35], yielding frame-level features $v \in \mathbb{R}^{n\times d}$, where d denotes the feature dimension. In the language branch, we use a shared RoBERTa [27] encoder to separately encode the question and each concatenated question–choice pair, yielding their respective textual representations. The question-only embedding $l_g \in \mathbb{R}^{1\times d}$ serves as the query for keyframe grounding, while the representation of question-choice pairs $l_a \in \mathbb{R}^{m\times d}$ are used for final answer prediction, where m denotes the number of choices.

3.1 Wavelet-based Dynamic Modeling

After extracting video embeddings, the representations mainly encode frame-level visual semantics, while fine-grained temporal variations are not sufficiently

emphasized. Such temporal cues, however, are critical for video reasoning tasks that require precise temporal discrimination. To complement the temporal modeling of the backbone, we introduce a lightweight wavelet decomposition layer inspired by discrete wavelet transform (DWT) [11] in signal processing. Given a temporal signal $x[n]$, DWT applies low-pass and high-pass filtering followed by a $2\times$ downsampling to produce the approximation and detail components, respectively:

$$a_n^{j+1} = \sum_k h[k] a_{2n-k}^{(j)}, \qquad d_n^{j+1} = \sum_k g[k] a_{2n-k}^{(j)} \tag{1}$$

where h and g denote the wavelet low-pass and high-pass filters. While classical DWT provides a principled way to separate coarse temporal trends from fine-grained variations, directly applying fixed wavelet filters to high-level video representations limits their adaptability to task-specific temporal patterns.

As shown in Fig. 3, we implement DWT using 1D group convolution initialized with wavelet kernels along the temporal dimension, where each channel is processed independently. The convolution stride is set to 2 to mimic the down-sampling behavior of standard DWT, producing an approximation signal $a^{(1)}$ and a detail signal $d^{(1)}$. Since the approximation component $a^{(l)}$ encodes the low-frequency temporal structure of the input sequence, it is naturally used as the input for the next wavelet decomposition level. By recursively applying the same wavelet filtering operation to $a^{(l)}$, the WDM module constructs a multi-level hierarchy of detail components $\{d^{(l)}\}_{l=1}^{L}$, where lower levels capture short-range temporal variations and higher levels model longer-range frame dynamics. As the temporal resolution of $d^{(l)}$ decreases with increasing decomposition depth, each detail component is up-sampled via interpolation to match the original temporal length. The reconstructed detail signals are then fused with the original video embeddings, enriching them with complementary temporal dynamics while preserving the original appearance-dominated representations.

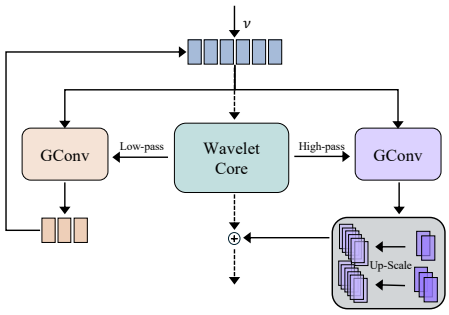

Fig. 3: Wavelet-based Dynamic Modeling (WDM) module. The low-pass approximation is recursively fed into subsequent wavelet levels to serve as the visual input for hierarchical temporal modeling.

After the WDM module, we adopt a two-layer Transformer as the temporal encoder to model global interactions among frames. To estimate the temporal interval t corresponding to the question-relevant video segment, we employ the Gaussian Smoothing Grounding module proposed in [5], which smooths the temporal attention distribution and encourages more stable temporal weights w. The final video representation is obtained via attention pooling, formulated as $v_t = v \times w$, and is subsequently used for the VideoQG task. Similar to

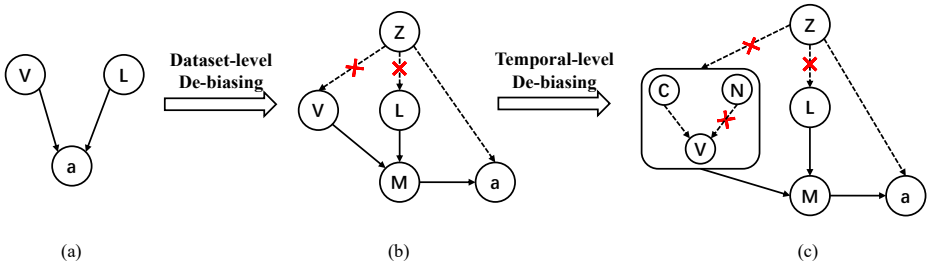

Fig. 4: Structural Causal Models for VideoQG tasks. (a) represents the standard QA setting. (b) indicates dataset-level bias and the deconfounding strategy adopted by [5]. (c) is our proposed framework, eliminates both dataset-level and temporal bias.

Temp[CLIP] [45], the temporal grounding interval t is determined by selecting the frame with highest attention weight, along with its neighboring frames within a predefined threshold.

3.2 Intra-video Counterfactual sample Mining

We model VideoQG task from a Structural Causal Model (SCM) perspective. As shown in Fig. 4(a), the video V and the question L jointly influence the predicted answer a, which can be formulated as $P(a|V, L)$.

However, video-language models are generally susceptible to various forms of bias. One issue is dataset-level bias, which introduces spurious correlation through confounding variables Z, forming back-door paths $V \leftarrow Z \rightarrow L$ and $Z \rightarrow a$. Following prior work [5], which has shown effectiveness in mitigating dataset-level bias, we adopt the same de-bias strategy in our model, as shown in Fig. 4(b). Beyond cross-sample statistical bias, VideoQG models also suffer from temporal bias within the video, which remains underexplored. In Fig. 4(c), a video typically contains both causally relevant information (C) and non-causal information (N), which jointly contribute to the entire video representation V, constructing the causal path $C \rightarrow V \leftarrow N$. To approximate these factors empirically, we define C as the frames that causally attribute to the correct answer and their underlying temporal ordering, while N denotes irrelevant or redundant frames. Therefore, the desired path is $C \rightarrow V \rightarrow a$, whereas the spurious path $N \rightarrow V \rightarrow a$ should be identified and suppressed by the model.

To address this issue, we propose an Intra-video Counterfactual Mining (ICM) strategy that approximates $do(\cdot)$ interventions to block the spurious $N \rightarrow V \rightarrow a$ pathway from the following two aspects. Frame-based counterfactuals suppress non-causal information by intervening on video segments that are irrelevant to answer the question. Shuffled-video counterfactual preserves frame-level appearance while disrupting the temporal ordering of frames. Unlike cross-instance negatives or hard negatives in video grounding [54], which mainly focus on representation-level separation, our goal is to isolate the non-causal factor N by preserving similar visual content while distinguishing segments through their

causal effects on answer prediction, thereby explicitly suppressing the spurious $N \rightarrow V \rightarrow a$ pathway.

Counterfactual Weight Construction For frame-based counterfactual, we perturb the temporal location of causal segment and construct counterfactual proposals on both left and right side of the video to mine progressively harder counterfactual samples, as shown in Fig. 5. This design is motivated by the observation: segments closer to the estimated causal segment share similar motion patterns and visual context, making them more confusing. We therefore start from the initial short, boundary-aligned counterfactuals that are less likely to cover core evidence and then gradually moved inward during training. Together with the QA loss, which penalizes localizations that harm answer prediction, this yields difficulty-ordered supervision while reducing reliance on precise early grounding. In practice, given frame-level attention weights $\{w_i\}_{i=1}^{n}$, we first normalize frame indices as $\tau_i = \frac{i-1}{n-1}$ to estimate the temporal center and variance of the causal segment as:

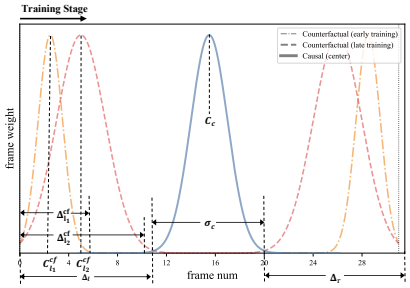

Fig. 5: Counterfactual samples are moved closer to the causal segment during training.

$$c_c = \sum_{i=1}^{n} w_i \cdot \tau_i, \qquad \sigma_c^2 = \sum_{i=1}^{n} w_i (\tau_i - c_c)^2. \tag{2}$$

We then compute the distances from the estimated causal boundaries to the video boundaries as:

$$\Delta_l = c_c - \frac{\sigma_c}{2}, \qquad \Delta_r = 1 - c_c - \frac{\sigma_c}{2}. \tag{3}$$

For each side, we place a counterfactual proposal whose boundary is aligned with the video edge. Its width is controlled by a distance ratio ρ, which determines how far the counterfactual segment resides from the causal segment. Formally, the widths of the left and right counterfactual proposals Δ_l^{cf} and Δ_r^{cf} satisfy $\frac{\Delta_l^{cf}}{\Delta_l} = \frac{\Delta_r^{cf}}{\Delta_r} = \rho$. We gradually increase ρ based on training process to enable curriculum learning, defined as:

$$\rho(e) = \left(\frac{e}{E_{\max}} \right)^{\gamma} \tag{4}$$

where γ controls the curriculum pace, e and E_{max} denote the current and total training epochs.

Their temporal centers can be obtained as $c_l^{cf} = \frac{\Delta_l^{cf}}{2}$ and $c_r^{cf} = 1 - \frac{\Delta_r^{cf}}{2}$. Each counterfactual segment induces a Gaussian-shaped temporal weight centered at its corresponding location, producing frame-level weights w_{cf}. On the other hand, shuffled-video counterfactual samples preserve frame-level appearance while disrupting the temporal ordering of frames. Concretely, given an frame-level video feature v, we generate a shuffled counterfactual $v^{shuf} = \pi(v)$, where $\pi(\cdot)$ denotes a random temporal permutation, and process v^{shuf} through the same encoding pipeline as the original video to obtain its counterfactual temporal interval w_{cf}^{shuf}. The final counterfactual video representation is obtained via weighted pooling: $v_{cf} = v \times w_{cf}$.

Margin-based Relevance Discrimination To alleviate the instability of inequality-based and contrastive supervision, we introduce a margin-based discrimination loss that quantifies the prediction gaps implied by following inequalities. Given a question–answer pair (L, a^*), the model is encouraged to satisfy the following scoring hierarchy:

$$P(a^*|L, V = v_t) > P(a^*|L, V = v) > P(a^*|L, V = v_{cf}) \tag{5}$$

where v_t and v_{cf} denote causal segment and counterfactual segment, respectively. We further treat the full video v as a reference as it contains both causally relevant and irrelevant content. We then compute the cross-entropy losses with predictions conditioned on causal, reference, and counterfactual video inputs, and impose margin-based constraints to regulate their separations:

$$\begin{aligned} \mathcal{L}_{\text{mdl}} = & \max\left(\mathcal{L}_{ce}^{c} - \mathcal{L}_{ce}^{r} + \lambda_1, 0\right) \\ & + \max\left(\mathcal{L}_{ce}^{c} - \mathcal{L}_{ce}^{cf_l} + \lambda_2, 0\right) \\ & + \max\left(\mathcal{L}_{ce}^{c} - \mathcal{L}_{ce}^{cf_r} + \lambda_2, 0\right) \end{aligned} \tag{6}$$

where $\mathcal{L}_{ce}^{(\cdot)}$ denotes the cross-entropy under different video inputs. λ_1 and λ_2 are hyperparameters satisfying $\lambda_1 < \lambda_2$, ensuring stronger separation between causal and counterfactual than causal and reference. This formulation imposes a quantitative realization of the desired ranking constraints, offering a more stable and discriminative supervisory signal for faithful temporal reasoning.

To facilitate cross-modal representation learning, we incorporate a contrastive alignment loss $\mathcal{L}_{align}$ following [45]. This loss adopts a standard InfoNCE formulation to align grounded video representations with question–answer features using in-batch negatives. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{qa} + \alpha\mathcal{L}_{align} + \beta\mathcal{L}_{mdl} \tag{7}$$

where α and β are hyperparameters that balance the contributions of the alignment loss and the margin-based discrimination loss, respectively.

4 Experiment

4.1 Experiment Settings

Datasets **1) NeXT-GQA** [45] is tailored for weakly supervised video question grounding. Built upon NeXT-QA [44], it excludes descriptive questions and those answerable from the global video context, retaining only causal and temporal reasoning queries. **2) STAR** [42] is organized around Situation, which jointly models entities, events, temporal points, and their surrounding context.

Evaluation metrics Following prior work on weakly supervised VideoQG [45], we evaluate both answer accuracy and temporal grounding quality. We report **Acc@VQA** to assess question answering performance. We then adopt Intersection over Prediction **(IoP)**, which measures whether the predicted temporal segment covers the ground-truth evidence. Accordingly, we report mean IoP (mIoP), TIoP@0.3 and TIoP@0.5. We include **TIoU@0.5** as a secondary metric for reference, following prior practice of reporting IoU at a fixed threshold. **Acc@GQA** measures the percentage of questions that are answered correctly and grounded correctly, where a prediction is considered grounded if its IoP exceeds 0.5.

Baseline To provide a comprehensive performance comparison, we evaluate our method against a diverse set of strong vision–language models that span different architectural paradigms, text encoders, and video representation strategies. The baselines considered in our experiments include:

- **IGV** [23] and **SeViLA** [49]: Originally proposed for VideoQA, these models are adapted to grounding by adding keyframe-aware localization and post-hoc temporal alignment.
- **VIOLETv2** [7]: VIOLETv2 adopts a Swin Transformer to encode spatio-temporal visual features and a BERT-based text encoder, followed by cross-modal fusion for video-language reasoning.
- **VGT** [46]: VGT adopts a graph-transformer architecture to model object-level visual relations and jointly reason over cross-modal relevance.
- **Temp[Swin]**, **Temp[CLIP]**, **Temp[BLIP]** [45]. These variants share the same temporal grounding framework but differ in the underlying visual encoder. The "PH" and "NG+" settings correspond to alternative strategies for generating grounding intervals during inference.
- **TimeCraft** [25]: TimeCraft introduces bidirectional reasoning for VideoQG and leverages large language models for data augmentation and self-supervised learning of both answering and temporal grounding.
- **CRA** [5]: CRA incorporates explicit causal intervention through back-door and front-door adjustments to mitigate textual and cross-modal spurious correlations, and further employs Gaussian filtering and bidirectional contrastive learning to enhance representation alignment.

Implementation details Following prior work [45], we uniformly sample 32 frames from each video and extract visual features using a frozen CLIP-L. Textual inputs are encoded with a RoBERTa backbone, which is fine-tuned during training. The backdoor- and frontdoor- interventions for dataset bias are strictly follow [5]. For WDM module, we adopt the standard Haar wavelet as the temporal filtering kernel. The extracted inter-frame dynamic components are fused to the original video features with a fixed weight 0.3. The decomposition level is set to 2 to capture multi-scale temporal dynamics while maintaining computational efficiency. For ICM module, we set the curriculum parameter γ to 0.5. The margin hyperparameters λ_1 and λ_2 are set to 0.15 and 0.2, respectively. The model is optimized using AdamW with an initial learning rate of 1×10^{-5}, while loss weights α and β are set to 1 and 0.3, respectively. All the experiments are conducted on an NVIDIA 4090 GPU.

4.2 Main Results

We evaluate our method on NeXT-GQA and STAR datasets, and compare against state-of-the-art VideoQG and VideoQA baselines under the weakly supervised grounding setting.

Results on NeXT-GQA Table 1 reports quantitative comparisons on NeXT-GQA benchmark. Our method (WICI) achieves best overall performance across major evaluation metrics, demonstrating consistent and meaningful improvements over temporal grounding baselines and causal reasoning approaches.

Comparison with Strong Temporal Baselines Compared with the strongest temporal grounding baseline Temp[CLIP](NG+), WICI improves Acc@GQA from 16.0% to 19.0%. Notably, this improvement is achieved with only a comparable model size (146.7M vs. 130.6M parameters), indicating a favorable trade-off between performance and model capacity. Meanwhile, WICI also improves Acc@VQA (60.2% $\rightarrow$ 61.5%) and mIoP (25.7% $\rightarrow$ 28.8%), suggesting that the gains stem from more accurate causal grounding rather than merely answer 'guessing'.

Comparison with Causal Reasoning Methods CRA provides a representative benchmark in causal reasoning, as it explicitly introduces causal interventions to mitigate spurious correlations. WICI further surpasses CRA by 0.8% in Acc@GQA (18.2% $\rightarrow$ 19.0%) and by 0.9% in TIoP@0.5 (28.5% $\rightarrow$ 29.4%), under a similar backbone and marginal model size increment. These improvements suggest that explicitly modeling inter-frame temporal dynamics, together with intra-video counterfactual mining and margin-based supervision, mitigates temporal bias by discouraging reliance on spurious temporal cues. Following CRA [5], we additionally report **Bias Error** and **Unfaithful** to quantify failure modes. Both metrics are defined over mis-grounded predictions, where the predicted segment has IoP < 0.3. Bias Error refers to cases where the model predicts an incorrect answer under mis-grounded evidence, while Unfaithful denotes cases where the predicted answer is correct but grounded on irrelevant segment (Table 2).

Table 1: VideoQG performances on NeXT-GQA test set. BT: BERT. RBT: RoBERTa. FT5: FLan-T5-XL. Temp[C]: Temp[CLIP]. Temp[B]:Temp[BLIP]. Human: indicates the performance achieved by human annotators. Random: always predicts a fixed answer ID and returns the entire video as the grounding. ∗: pretrained on video-language grounding dataset. Data in gray shading in the table are only provided as reference and are not included in comparison. TimeCraft augments its training set with extra data constructed using the LLaMA-2-13B model. We compare methods that use Temp[CLIP] as the backbone to ensure a fair evaluation.

Method	Model	Param	Vision	Text	Acc		IoP			TIoU@0.5
					@GQA	@VQA	mIoP	@0.3	@0.5	
Human	-	-	-	-	81.2	93.3	72.1	91.7	86.2	70.3
Random	-	-	-	-	1.7	20.0	21.1	20.6	8.7	8.7
SeViLA∗	BLIP-2	4.1 B	ViT-G	FT5	16.6	68.1	29.5	34.7	22.9	13.8
IGV	-	-	ResNet	BT	10.2	50.1	21.4	26.9	18.9	9.6
MIST	Temp[C]	-	ViT-L	BT	12.1	-	23.6	29.3	20.7	7.0
PH	VIOLETv2	-	VSWT	BT	12.8	52.9	23.6	25.1	23.3	1.3
PH	VGT	-	RCNN	RBT	14.4	55.7	25.3	26.4	25.3	1.7
PH	Temp[S]	-	SWT	RBT	13.5	55.9	23.1	24.7	23.0	2.3
PH	Temp[C]	-	ViT-B	RBT	14.7	57.9	24.1	26.2	24.1	3.7
PH	Temp[B]	-	ViT-B	RBT	14.9	58.5	25.0	27.8	25.3	4.5
PH	Temp[C]	130.3M	ViT-L	RBT	15.2	59.4	25.4	28.2	25.5	4.1
NG	Temp[C]	130.6M	ViT-L	RBT	15.5	59.4	25.8	28.8	25.9	4.6
NG+	Temp[C]	130.6M	ViT-L	RBT	16.0	60.2	25.7	31.4	25.5	8.9
TimeCraft	Temp[C]	130.6M	ViT-L	RBT	18.2	-	28.1	**35.1**	27.8	9.6
CRA	Temp[C]	145.1M	ViT-L	RBT	18.2	61.1	28.6	34.3	28.5	**10.6**
WICI(Ours)	Temp[C]	146.7M	ViT-L	RBT	**19.0**	**61.5**	**28.8**	34.3	**29.4**	10.2

In practice, our model consistently favors compact temporal segments with high causal confidence. This behavior arises naturally from our decision-centric causal intervention, which suppresses visually similar yet non-causal segments through the ICM module. As a result, the model is encouraged to identify minimal but sufficient temporal evidence that most strongly supports the answer, rather than maximizing temporal coverage. Such compact predictions lead to higher precision-oriented localization, which is reflected in improved IoP performance, while potentially sacrificing IoU that rewards broader span overlap.

Table 2: Debiasing analysis on the NeXT-GQA test set. BE and UF denote Bias Error and Unfaithful, respectively.

Method	Acc@GQA↑	Acc@VQA↑	BE↓	UF↓
PH	15.3	59.0	28.5	41.4
CRA	18.2	61.1	27.4	40.0
WICI	**19.0**	**61.5**	**27.1**	**39.6**

Results on STAR We further evaluate our model on the STAR dataset to verify its generalization under a different Situation-oriented VideoQG benchmark. As shown in Table 3, our method consistently outperforms prior baselines across all grounding-related metrics. While the improvement in Acc@VQA is relatively

Table 3: VideoQG performance on STAR validation set.

Method	Acc@GQA	Acc@VQA	TIoP@0.5	TIoU@0.5
SeViLA*	17.0	45.5	37.6	19.5
IGV	13.1	31.7	42.8	1.7
Temp[CLIP](NG+)	24.4	57.3	41.4	4.7
Temp[CLIP](CRA)	26.8	58.6	44.5	5.5
Temp[CLIP](WICI)	**29.2**	**58.8**	**48.5**	**5.9**

Table 4: Ablation study on NeXT-GQA.

Align	WDM	ICM	Acc@GQA	Acc@VQA	TIoP@0.5	TIoU@0.5
✓	✓	✓	19.0	61.5	29.4	10.2
✓		✓	17.8(↓ 1.2)	60.5(↓ 1.0)	28.6(↓ 0.8)	9.5(↓ 0.7)
✓	✓		17.7(↓ 1.3)	60.8(↓ 0.7)	28.4(↓ 1.0)	7.1(↓ 3.1)
✓			17.3(↓ 1.7)	60.1(↓ 1.4)	28.1(↓ 1.3)	7.5(↓ 2.7)

moderate, we observe substantial gains in both Acc@GQA and TIoP@0.5. Specifically, compared with the strongest baseline Temp[CLIP](CRA), our method improves Acc@GQA from 26.8 to 29.2, and boosts TIoP@0.5 from 44.5 to 48.5. This validates that our approach improves the model's ability to answer questions grounded in factual video evidence, rather than merely boosting answer accuracy, leading to more reliable and faithful predictions in VideoQG.

4.3 Ablation Study

Breakdown Ablation We conduct ablation experiments on NeXT-GQA to examine the contribution of each major component in Table 4. Removing the WDM module leads to Acc@GQA(↓ 1.2) and Acc@VQA(↓ 1.0) degradation, suggesting that explicitly modeling inter-frame temporal dynamics is critical to distinguish semantically similar but functionally different video segments. This result confirms that dynamic cues are as important as static appearances in video representation. When the ICM module is disabled, the Acc@GQA drops more substantially (↓ 1.3), highlighting the importance of explicitly mitigating intra-video temporal bias. Without ICM, the model tends to aggregate visually redundant yet non-causal segments, resulting in unfaithful question answering.

We further conduct a temporal sensitivity analysis in Table 5 by shuffling frames at inference. WICI shows larger degradation across all metrics, especially grounding-related metrics, under shuffled video inputs. This degradation is expected as a temporally faithful model should be sensitive to temporal order, whereas excessive invariance would indicate reliance on static appearance shortcuts rather than causal temporal reasoning. Additional ablation studies and qualitative visualizations are provided in the supplementary material.

Table 5: Temporal sensitivity analysis under inference-time frame-order perturbation. * denotes evaluation with shuffled input frames.

Method	Acc@GQA	Acc@VQA	TIoP@0.5	TIoU@0.5
WICI	19.0	61.5	29.4	10.2
w/o WICI	17.3	60.1	28.1	7.5
WICI*	15.0(↓ 4.0)	60.9(↓ 0.6)	23.3(↓ 6.1)	6.3(↓ 3.9)
w/o WICI*	13.9(↓ 3.4)	59.9(↓ 0.2)	22.4(↓ 5.7)	5.8(↓ 1.7)

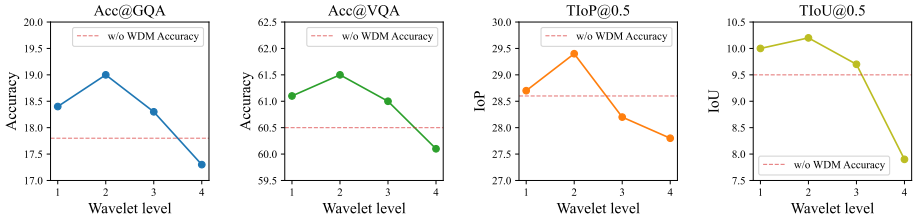

Fig. 6: Ablation study on different wavelet levels.

Further Experiment on WDM We conduct a hyperparameter ablation study to investigate the impact of the number of wavelet decomposition levels in the proposed Wavelet Decomposition Module (WDM). The quantitative results are summarized in Fig. 6. We argue that effective VideoQG requires a balanced integration of dynamic temporal cues and static semantic information. While moderate wavelet decomposition enhances discriminative inter-frame dynamics, overly deep WDM amplifies high-frequency temporal variations at the expense of low-frequency semantic consistency. Such over-emphasis on fine-grained dynamics may hinder the model's ability to aggregate temporally extended but semantically coherent evidence, which is critical for reliable question answering.

Further Experiment on ICM We further decompose ICM into its key elements to assess their contributions. Specifically, ICM consists of two orthogonal aspects: (i) how counterfactual samples are constructed and scheduled, and (ii) how counterfactual supervision is imposed. As shown in Table 6, omitting the shuffled counterfactual samples drops the Acc@GQA from 19.0 to 18.2. Further, replacing curriculum-guided counterfactual mining with a naive strategy, where counterfactual segments are sampled with a fixed distance from the causal segment, leads to degraded Acc@GQA (↓ 0.8). This highlights the benefit of progressively increasing counterfactual difficulty, enabling stable optimization by avoiding overly aggressive early perturbations. Replacing the proposed margin-based loss with a contrastive objective over intra-video counterfactuals results in inferior Acc@GQA (↓ 0.7), with a larger gap observed (↓ 1.3) when WDM module is removed. Without explicit dynamic cues, contrastive learning tends to over-separate highly similar video representations, leading to degraded perfor-

Table 6: Ablation study w.r.t. ICM module. w/o TO-CF removes the temporal-shuffle branch during training. w/o CM denotes naive ICM module, which omits the curriculum schedule and the distance ρ is fixed to 0.5. w/o ML denotes ICM module using contrastive loss instead of margin-based loss. w/o ML* further removes WDM module.

Method	Acc@GQA	Acc@VQA	TIoP@0.5	TIoU@0.5
WICI	**19.0**	**61.5**	**29.4**	**10.2**
w/o TO-CF	18.2	60.9	28.8	8.9
w/o CM	18.2	60.7	28.8	8.7
w/o ML	18.3	61.2	29.0	8.6
w/o ML*	17.7	60.4	28.4	7.0

mance, whereas margin-based ranking remains stable, indicating more reliable supervision for intra-video causal discrimination.

5 Conclusion

In this paper, we presented the Wavelet-based Intra-video Causal Intervention (WICI) framework, designed to tackle underexplored intra-video temporal bias in VideoQG. Unlike prior methods that predominantly rely on static appearance features, WICI effectively addresses the challenge of distinguishing genuine causal evidence from visually similar but non-causal distractors. By introducing wavelet transforms to capture fine-grained inter-frame temporal dynamics and incorporating a causal intervention mechanism to eliminate interference from visually similar but irrelevant segments, our method significantly enhances the model's ability to identify key evidence in complex video scenarios.

Despite these advancements, the counterfactual sample construction is mainly designed for settings where most temporal/causal questions depend on one continuous evidence segment, and may be less effective for multiple discrete spans. Future research will focus on exploring mixture-based multi-span counterfactual modeling and more robust temporal bias confounder generation strategies to extend this causal intervention framework to a broader spectrum of video-language tasks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

1. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Brit. Mach. Vis. Conf. (2016)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Int. Conf. Mach. Learn. vol. 382, pp. 41–48 (2009)
3. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. In: Adv. Neural Inform. Process. Syst. pp. 839–850 (2019)
4. Chen, Q., Di, S., Xie, W.: Grounded multi-hop videoqa in long-form egocentric videos. In: AAAI. vol. 39, pp. 2159–2167 (2025)
5. Chen, W., Liu, Y., Chen, B., Su, J., Zheng, Y., Lin, L.: Cross-modal causal relation alignment for video question grounding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 24087–24096 (2025)
6. Di, S., Xie, W.: Grounded question-answering in long egocentric videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12934–12943. IEEE (2024)
7. Fu, T., 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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22898–22909 (2023)
8. Fujieda, S., Takayama, K., Hachisuka, T.: Wavelet convolutional neural networks. CoRR (2018), `http://arxiv.org/abs/1805.08620`
9. Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R.S., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Mach. Intell. **2**, 665–673 (2020)
10. Gupta, A., Roy, A., Chellappa, R., Bastian, N.D., Velasquez, A., Jha, S.: TOGA: temporally grounded open-ended video QA with weak supervision. In: Int. Conf. Comput. Vis. pp. 23593–23603. IEEE (2025)
11. Heil, C., Walnut, D.F.: Continuous and discrete wavelet transforms. SIAM Rev. **31**, 628–666 (1989)
12. Huang, J., Liu, Y., Gong, S., Jin, H.: Cross-sentence temporal and semantic relations in video activity localisation. In: Int. Conf. Comput. Vis. pp. 7179–7188. IEEE (2021)
13. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: TGIF-QA: toward spatio-temporal reasoning in visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1359–1367 (2017)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Int. Conf. Mach. Learn. vol. 139, pp. 4904–4916 (2021)
15. Kim, H., Tang, Z., Bansal, M.: Dense-caption matching and frame-selection gating for temporal localization in videoqa. In: Annu. Meeting Assoc. Comput. Linguist. pp. 4812–4822 (2020)
16. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7331–7341 (2021)
17. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: localized, compositional video question answering. In: Conf. Empirical Methods Natural Lang. Process. pp. 1369–1379 (2018)
18. Lei, J., Yu, L., Berg, T.L., Bansal, M.: TVQA+: spatio-temporal grounding for video question answering. In: Annu. Meeting Assoc. Comput. Linguist. pp. 8211–8225 (2020)

19. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Int. Conf. Mach. Learn. vol. 202, pp. 19730–19742 (2023)
20. Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. In: Adv. Neural Inform. Process. Syst. pp. 9694–9705 (2021)
21. Li, L., Chen, Y., Cheng, Y., Gan, Z., Yu, L., Liu, J.: HERO: hierarchical encoder for video+language omni-representation pre-training. In: Conf. Empirical Methods Natural Lang. Process. pp. 2046–2065 (2020)
22. Li, Y., Wang, X., Xiao, J., Chua, T.: Equivariant and invariant grounding for video question answering. In: ACM Int. Conf. Multimedia. pp. 4714–4722 (2022)
23. Li, Y., Wang, X., Xiao, J., Ji, W., Chua, T.: Invariant grounding for video question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2918–2927 (2022)
24. Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. In: AAAI. vol. 34, pp. 11539–11546 (2020)
25. Liu, H., Ma, X., Zhong, C., Zhang, Y., Lin, W.: Timecraft: Navigate weakly-supervised temporal grounded video question answering via bi-directional reasoning. In: Eur. Conf. Comput. Vis. Lecture Notes in Computer Science, vol. 15063, pp. 92–107 (2024)
26. Liu, Y., Li, G., Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 11624–11641 (2023)
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR (2019), `http://arxiv.org/abs/1907.11692`
28. Ma, H., Pathak, V., Wang, D.Z.: Bridging vision language models and symbolic grounding for video question answering. CoRR (2025), `https://doi.org/10.48550/arXiv.2509.11862`
29. 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: Eur. Conf. Comput. Vis. Lecture Notes in Computer Science, vol. 12373, pp. 156–171 (2020)
30. Mallat, S.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
31. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: Eur. Conf. Comput. Vis. Lecture Notes in Computer Science, vol. 13664, pp. 1–18 (2022)
32. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X., Wen, J.: Counterfactual VQA: A cause-effect look at language bias. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12700–12710. Computer Vision Foundation / IEEE (2021)
33. Pearl, J.: Causal diagrams for empirical research (with discussions). Biometrika **82**(4), 669–688 (1995)
34. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. vol. 139, pp. 8748–8763 (2021)
36. Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum learning: A survey. Int. J. Comput. Vis. **130**, 1526–1565 (2022)

37. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: pre-training of generic visual-linguistic representations. In: Int. Conf. Learn. Represent. (2020)
38. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense representation learning via causal inference. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1547–1550 (2020)
39. Wang, W., Gao, J., Xu, C.: Weakly-supervised video object grounding via causal intervention. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3933–3948 (2023)
40. Wei, Y., Liu, Y., Yan, H., Li, G., Lin, L.: Visual causal scene refinement for video question answering. In: ACM Int. Conf. Multimedia. pp. 377–386 (2023)
41. Williams, T.L., Li, R.: Wavelet pooling for convolutional neural networks. In: Int. Conf. Learn. Represent. (2018)
42. Wu, B., Yu, S., Chen, Z., Tenenbaum, J., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. In: Adv. Neural Inform. Process. Syst. (2021)
43. Xiao, J., Huang, N., Qin, H., Li, D., Li, Y., Zhu, F., Tao, Z., Yu, J., Lin, L., Chua, T., Yao, A.: Videoqa in the era of llms: An empirical study. Int. J. Comput. Vis. **133**, 3970–3993 (2025)
44. Xiao, J., Shang, X., Yao, A., Chua, T.: Next-qa: Next phase of question-answering to explaining temporal actions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9777–9786. Computer Vision Foundation / IEEE (2021)
45. Xiao, J., Yao, A., Li, Y., Chua, T.: Can I trust your answer? visually grounded video question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13204–13214 (2024)
46. Xiao, J., Zhou, P., Chua, T., Yan, S.: Video graph transformer for video question answering. In: Eur. Conf. Comput. Vis. Lecture Notes in Computer Science, vol. 13696, pp. 39–58 (2022)
47. Xiao, J., Zhou, P., Yao, A., Li, Y., Hong, R., Yan, S., Chua, T.: Contrastive video question answering via video graph transformer. IEEE Trans. Pattern Anal. Mach. Intell. **45**(11), 13265–13280 (2023)
48. Xu, H., Ghosh, G., Huang, P., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: Conf. Empirical Methods Natural Lang. Process. pp. 6787–6800 (2021)
49. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 76749–76771 (2023)
50. Zang, C., Wang, H., Pei, M., Liang, W.: Discovering the real association: Multimodal causal reasoning in video question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19027–19036 (2023)
51. Zeng, R., Gan, C., Chen, P., Huang, W., Wu, Q., Tan, M.: Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. IEEE Trans. Image Process. **28**(12), 5797–5808 (2019)
52. Zhang, X., Zhang, F., Xu, C.: Reducing vision-answer biases for multiple-choice VQA. IEEE Trans. Image Process. **32**, 4621–4634 (2023)
53. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: AAAI. vol. 36, pp. 3517–3525 (2022)
54. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15555–15564 (2022)