Exploring Question Guidance and Answer Calibration for Visually Grounded Video Question Answering

Yuanxing Xu[*], **Yuting Wei**[*], **Shuai Zhong, Xinming Chen, Jinsheng Qi, Bin Wu**[†]
Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications
{xyx,yuting_wei,wnbhhdsd,cxm,qijs,wubin}@bupt.edu.cn

Abstract

Video Question Answering (VideoQA) tasks require not only correct answers but also visual evidence. The "localize-then-answer" strategy, while enhancing accuracy and interpretability, faces challenges due to the lack of temporal localization labels in VideoQA datasets. Existing methods often train the models' localization capabilities indirectly using QA labels, leading to inaccurate localization. Moreover, our experiments show that despite high accuracy, current models depend too heavily on language shortcuts or spurious correlations with irrelevant visual context. To address these issues, we propose a **Q**uestion-**G**uided and **A**nswer-**C**alibrated **TR**ansformer (QGAC-TR), which guides and calibrates localization using question and option texts without localization labels. Furthermore, we design two self-supervised learning tasks to further enhance the model's refined localization capabilities. Extensive experiments on three public datasets focused on temporal and causal reasoning show that our model not only achieves accuracy comparable to large-scale pretrained models but also leads in localization aspects. Code will be released at this link.

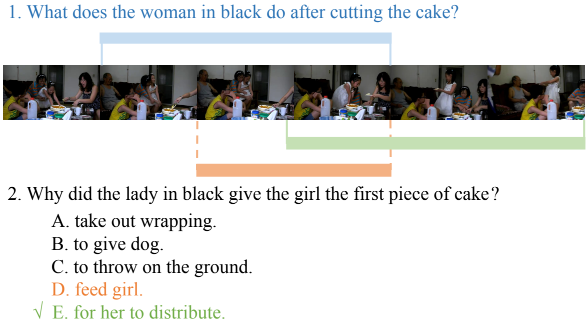

1. What does the woman in black do after cutting the cake?
2. Why did the lady in black give the girl the first piece of cake?
A. take out wrapping.
B. to give dog.
C. to throw on the ground.
D. feed girl.
√ E. for her to distribute.

Figure 1: Examples of the impact of questions and options on localization. (1) Different questions often point to different segments. (2) For the second question, without the calibration of options for localization, current models (Li et al., 2023b; Xiao et al., 2024) easily focus on the orange area and predict 'feed girl' as the answer.

1 Introduction

Videos, with their rich information, engaging content, and user connectivity, have become highly sought-after multimedia data (Wu et al., 2017; Apostolidis et al., 2021; Zhong et al., 2022). However, sifting through extensive video content is time-consuming, making it hard for users to fully understand and find specific content. Therefore, there is a growing demand for technology that can quickly locate and answer questions to aid user understanding of video content.

Recent studies have devoted significant effort to VideoQA tasks, including Transformer-based models (Xiao et al., 2022b; Gao et al., 2023; Li et al., 2023b; Xiao et al., 2023, 2024; Cherian et al., 2022; Kim et al., 2023) and vision-language pretrained models (Buch et al., 2022; Yang et al., 2022; Ye et al., 2023). Recent works (Qian et al., 2023; Gao et al., 2023; Li et al., 2023b; Xiao et al., 2024) have introduced a localization process to enhance accuracy and interpretability by using Transformers for cross-modal interaction and temporal localization of videos. On the other hand, with the advancement of pretrained language models, recent methods yield encouraging results by incorporating them into the training process for fine-tuning. However, we pose the following questions: (1) **Are the existing methods for model localization capabilities optimal, and is the text's potential for aiding localization fully exploited?** We believe that due to the absence of localization labels, models can only train their localization capabilities indirectly through QA tasks, i.e., assessing the accuracy of localization based on the correctness of the provided answers. However, higher QA accuracy does not equate to better localization performance (Xiao et al., 2024), and the indirect training approach of-

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 3121–3133
November 12-16, 2024 ©2024 Association for Computational Linguistics

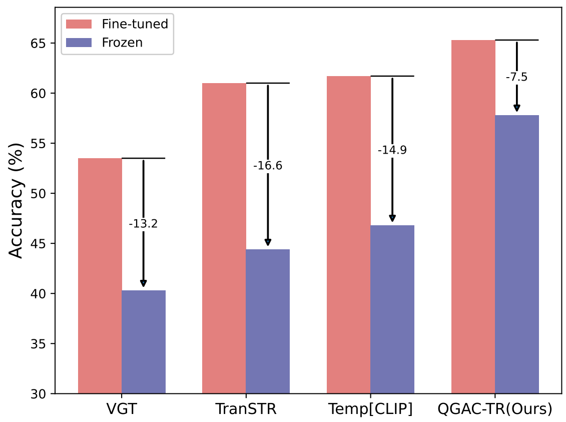


Figure 2: Accuracy grouped by whether the language model is frozen. All results are reported on NExT-QA (Xiao et al., 2021a) test set.

ten leads to inaccuracies. We emphasize that the potential of question and answer texts to guide and calibrate localization remains underexploited. As shown in Figure 1, different questions typically point to different segments of the same video, and most questions are difficult to accurately localize without answer calibration. Although there are various methods in Video Temporal Grounding tasks for reference, VideoQA are significantly different: the former's texts are declarative sentences containing all visual clues, while the latter's are interrogative sentences, often containing only part of the key visual clues, which may lead to omissions in the localization process. These challenges increase the risk of "one wrong step leading to many" in models that adopt a "localize-then-answer" strategy. (2) **To what extent do these models rely on language shortcuts?** Although early studies like IGV (Li et al., 2022) have explored this issue, as shown in Table 3, its ability to localize visual evidence remains low, with significant issues of linguistic shortcuts still present. Furthermore, we selected several representative models (Xiao et al., 2022b; Li et al., 2023b; Xiao et al., 2024) for experimentation, applying both fine-tuning and freezing of the language model, as shown in Figure 2. The data indicates that when the language model is frozen during training, the performance of current methods significantly declines. This suggests that their responses might rely more on language shortcuts or spurious correlations with irrelevant visual context, rather than truly understanding the video content.

To address the aforementioned challenges, we propose the QGAC-TR, which aims to fully utilize the guidance and calibration of questions and options to achieve precise localization of video segments without localization labels, focusing on providing correct answers based on visual content. Specifically, after employing a cross-attention-based Transformer encoder to enforce the involvement of question text in extracting video representations, we introduce a question relevance token representing the entire video to serve as a standard for adaptive localization, designing an input-adaptive localizer. Subsequently, we designed two self-supervised learning tasks. The first involves constructing negative pairs by mixing original video-question pairs, training the model to suppress the question relevance scores of these negative pairs. Through this process, we expect the model to develop the ability to discern relevant video segments. The second task uses the relevance scores generated from video-question pairs as an anchor, treating scores related to correct and incorrect options as positive and negative samples, respectively, for self-supervised contrastive learning. This process aims to more comprehensively and accurately locate video segments, compensating for the shortcomings of the visual clues provided solely by question text.

In summary, our contributions are as follows:

(1) We reveal the insufficient use of text in existing methods and their heavy reliance on language shortcuts or spurious visual correlations.

(2) We introduce QGAC-TR, which delves into the guidance and calibration roles of question and option texts, prompting the model to rely more on visual content for accurate localization and QA.

(3) We propose an input-adaptive question relevance score predictor and have designed two types of self-supervised learning tasks aimed at enhancing the model's ability to understand and locate video content based on questions, especially in the absence of localization labels.

(4) Our extensive experimental results demonstrate that QGAC-TR surpasses state-of-the-art standard models in terms of QA and localization, and in some cases, even exceeds some large pre-trained models, showcasing its robust localization and question-answering performance.

2 Related Works

Video Question Answering. Early methods primarily included cross-modal attention (Jang et al., 2017; Li et al., 2019; Jiang et al., 2020), motion-appearance memory (Gao et al., 2018; Fan et al., 2019; Liu et al., 2021b), and graph neural networks (GNNs) (Jiang and Han, 2020; Park et al., 2021; Xiao et al., 2022a; Li et al., 2022). In recent years, Transformer-based models have demonstrated their excellence in the VideoQA task, particularly with (Xiao et al., 2022b), which combined Transformers with GNNs for significant performance improvements through fine-grained spatiotemporal modeling. (Gao et al., 2023; Li et al., 2023b; Yu et al., 2024; Xu et al., 2023; Xiao et al., 2024) introduced the "localize-then-answer" strategy, not only enhancing accuracy but also interpretability. Specifically, (Gao et al., 2023; Li et al., 2023b) located video segments/frames relevant to the questions, (Yu et al., 2024; Xu et al., 2023) selected key frames generatively on datasets with GT labels (Lei et al., 2021a) using the pretrained multimodal model, and (Xiao et al., 2024) refined attention weights with Gaussian masks for nuanced localization. However, these approaches have not fully exploited the potential of question and option texts.

Video Temporal Grounding. This task closely relates to the localization phase in VideoQA. Traditional methods are categorized into proposal-based and proposal-free approaches (See Appendix A.1 for details). Recent advancements include DETR-based and regression-based approaches, with methods like UMT (Liu et al., 2022) incorporating additional audio modalities, and (Lei et al., 2021a; Moon et al., 2023; Jang et al., 2023) developing DETR architectures. However, their direct application to VideoQA is challenging without localization labels and when questions contain only partial visual cues.

Language Shortcuts in Vision-Language Tasks. In terms of reducing language shortcuts and spurious visual correlations, our research aligns with other domains. For instance, (Goyal et al., 2017) addressed the issue of language shortcuts in VQA by creating VQAv2, which pairs questions with additional images that have similar content but different answers. (Niu et al., 2021) and (Guo et al., 2021) mitigate this issue by adjusting prediction scores. (Zeng et al., 2023) developed X2-VLM through multi-level vision-language pretraining to enhance spatial localization. These studies attempt to avoid shortcut learning in image processing by collecting new datasets, employing specialized learning strategies, or focusing on spatial grounding. Methods suitable for VideoQA remain largely unexplored.

3 Methods

Problem Formulation. Given a video V, a question Q related to it and several candidate answers (options) o_i, the goal of VideoQA is to select the correct option, formulated as follows:

$$\hat{a} = \arg\max_{a \in O} \mathcal{F}_\theta(a|V, Q, O), \tag{1}$$

where $O = \{o_n\}_{n=1}^{|O|}$ and $\mathcal{F}$ is a VideoQA model with trainable parameter θ.

Model Architecture Overview. Our overall model architecture is shown in Figure 3. (i) The model uses pretrained visual and text encoders to extract features. (ii) Tokens along with a learnable relevance token are fed into Input-adaptive Localizer. Two self-supervised learning tasks are also designed, namely mixing different samples of videos and questions for negative pair learning ($\mathcal{L}_{neg}$), and question-option pairs with correct and incorrect options serving as positive and negative samples for contrastive learning ($\mathcal{L}_{cl}$). (iii) The questions and localized frames are processed through another cross-attention Transformer Encoder for QA reasoning and prediction.

3.1 Feature Extraction

As in previous work (Xiao et al., 2022a,b, 2023, 2024), we first uniformly divide the video V into T clips and collect the intermediate frames of each clip to represent the video. Then, frame features are extracted for each frame using a frozen pretrained image encoder (i.e. ViT of CLIP (Radford et al., 2021)), denoted as f_t. For the text part, we use a pretrained language model (i.e. RoBERTa (Liu et al., 2019)) as a text encoder to encode the question q as a sequence of L_q tokens and extract their features, notated as q_l; the question q is concatenated to each option o_i, where [SEP] token is inserted, also after the text encoder to extract features, notated as o_l, where L_O denotes the sequence length of the longest question-option pair. In order to project all features into a common d-dimensional space, we apply two linear mappings for the frame features and the two text features, respectively, to obtain $F = \{f_t\}_{t=1}^{T} \in \mathbb{R}^{T \times d}$, $q = \{q_l\}_{l=1}^{L_q} \in \mathbb{R}^{L_q \times d}$ and $o_n = \{o_n^l\}_{l=1}^{n \cdot L_O} \in \mathbb{R}^{L_O \times d}$. The frame feature F passes through a temporal Transformer layer to capture the temporal dynamic information between frames, and its process is omitted here for brevity.

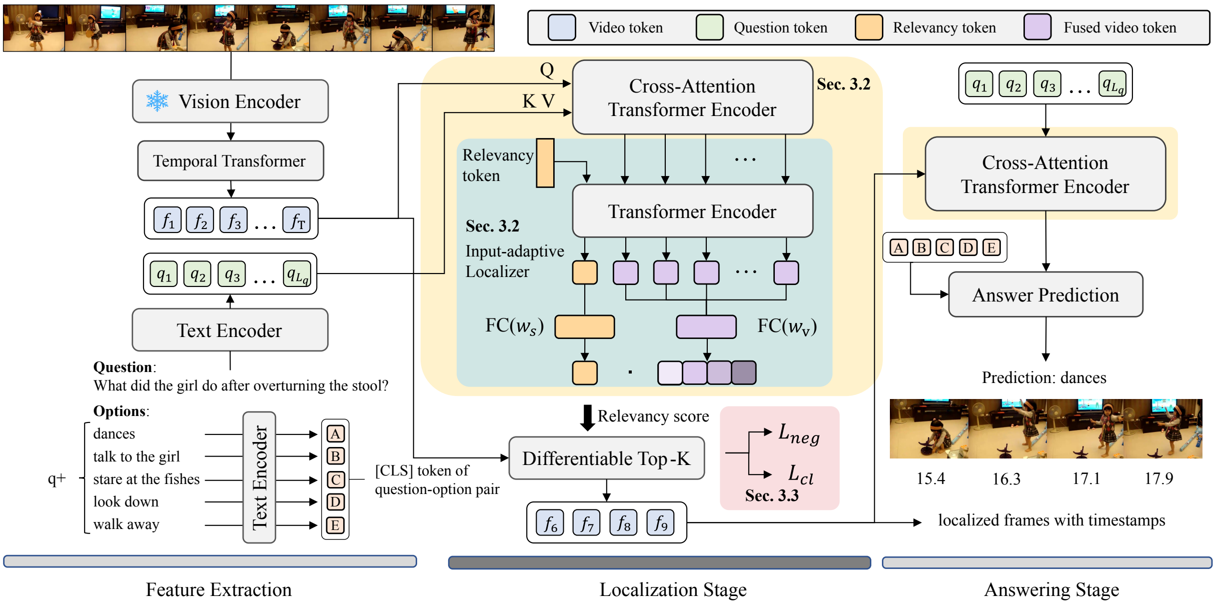


Figure 3: Overview of the proposed QGAC-TR architecture. From left to right, the model consists of three stages: (i) feature extraction, (ii) localization, and (iii) question answering. See Chapter 3 for more details.

3.2 Cross-Attention Transformer Encoder with Adaptive Localization

Deep interaction between video and question. Existing methods either simply concatenate the features of each modality or perform fusion at a later stage. However, we believe that the relationship between modalities should be more carefully scrutinized and mined, which is necessary for a deep understanding of the relationship between questions and videos. The goal of the cross-attention encoder is to produce frame-level representations with question relevance information. In cross-attention, we use the frame representation F of the contextual culture as the query and the problem feature q as the key and value, so the cross-attention layer operates as follows:

$$Q = F, K = V = q, X = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d}}}\right)\mathrm{V}. \quad (2)$$

Note that we choose to use frames rather than clips because we believe that a pre-divided segmentation approach similar to (Gao et al., 2023) is likely to cut the same key scene apart and is less flexible. However, the video frames relevant to the question tend to be concentrated, so we take an alternative approach to encourage the model to select consecutive frames, as detailed in Sec. 3.5.

Input-adaptive Localization. We introduce a relevance token x_r as an input-adaptive localizer. x_r is a learnable vector, initially randomized. When it is added to the encoded video token sequence and processed through a transformer encoder, it becomes an input-adaptive predictor, adjusting adaptively based on the specific context of the input. As shown in Figure 3, we concatenate x_r with the video token X, feeding them into the Transformer encoder. This allows x_r to reorganize itself based on relevant contextual information. Each video token and the relevance token are projected by respective fully connected layers with weights w_v and w_r, and their scaled dot product defines the relevance score. Therefore, the relevance score is computed as follows:

$$S(x_v^i) = \frac{w_s^T x_s \cdot w_v^T x_v^i}{\sqrt{d}}. \quad (3)$$

3.3 Refine Localization with Questions and Options

In this section, we aim to refine the localization process by leveraging the semantic relation between questions and multiple choice options. The refinement process is twofold: learning from negative pairs to enhance the discriminative ability of the model and employing contrastive learning to refine localization based on the labeled options. The positive and negative samples of these two self-supervised learning tasks are shown in Fig. 4.

Learning from Negative Pairs. Negative relation learning is a crucial aspect of our approach to VideoQA. By creating negative pairs, which consist of videos and questions from different samples in a batch, we train the model to discern irrelevant

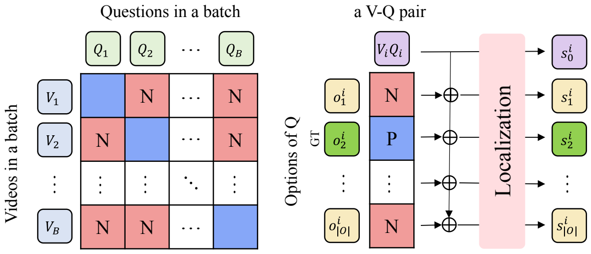


Figure 4: Labels for video-question negative pair learning (left) and contrastive learning of correct and incorrect options (right) to guide and calibrate localization. The green element (o_2^i) on the right represents the ground-truth answer, and the rest represent incorrect options. $\oplus$ means concatenate the option with the question.

content and minimize false positive localizations. The loss function for negative pair learning is as follows:

$$\mathcal{L}_{neg} = -\log(1 - S(x_v^{neg})). \tag{4}$$

This negative relation learning enhances the model's ability to distinguish between relevant and irrelevant video segments, ensuring that the localization is tightly coupled with the context of the question.

Note that the ultimate goal of the localization stage is to select the frames most relevant to the question, requiring the model to differentiate between video segments. Increasing the scores for every segment in the positive samples contradicts this. Therefore, although it is easy to collect positive samples (as shown in the blue part on the left of Figure 4), we use only negative pairs for self-supervised learning.

Contrastive Learning via Labeled Options. While negative V-q pairs enhance the global discriminative understanding of the model, the local nuances associated with each specific question-option pair still present a challenge. To address this, we employ contrastive learning, using the correlation score S^a obtained from the V-q pair as an anchor. The model is then trained to recognize the correct option by comparing the anchor score with the scores obtained from the correct and incorrect options, respectively denoted as S^+ and S^-.

$$\mathcal{L}_{cl} = -\log \frac{\exp\left(S^a \odot S^+/\tau\right)}{\sum_{*\in\{+,-\}} \exp\left(S^a \odot S^*/\tau\right)}. \tag{5}$$

The contrastive loss function above encourages the model to align the localization more closely with the video segment corresponding to the correct option while distancing it from segments that correspond to incorrect options. This way, the model not only learns to localize relevant content but also fine-tunes its understanding based on the context provided by the correct option.

3.4 Answer Reasoning and Prediction

We first utilize the TopK operation to select the K frames with the highest relevance scores. The equation is:

$$F_{loc} = \mathrm{PerturbedTopK_k(F', S(X))} \in \mathbb{R}^{\mathrm{k\times d}}. \tag{6}$$

During the inference phase, we replace the Perturbed TopK with the original Hard TopK method to achieve higher efficiency.

Answer Reasoning. We use another Transformer encoder with cross-attention, featuring distinct parameters from the localization stage (see 3.2). This setup addresses the different requirements of each stage: global relevance and key information capture in localization, and detailed understanding and utilization of located video segments in answering. The encoder specifically focuses on analyzing video details and combining textual information for precise matching and reasoning.

Answer Prediction. Classical approaches consider VideoQA as a classification task. However, recent studies like (Xiao et al., 2022b; Gao et al., 2023; Xiao et al., 2024) have demonstrated the superiority of similarity-based methods, and we follow this approach. Let F_{loc} be obtained through Equation 3 as X_{loc}, then the formula for computing the final predicted answer is:

$$\hat{a} = \underset{a\in O}{\arg\max}(X_{loc}O^{\mathrm{T}}). \tag{7}$$

3.5 Training Objectives

We follow previous work and adopt the cross-entropy loss as the main loss, denoted as $\mathcal{L}_{ce}$. The loss functions $\mathcal{L}_{neg}$ and $\mathcal{L}_{cl}$ for the two designed self-supervised objectives have been detailed in Sec. 3.3.

Additionally, as continuous scenes better maintain the coherence of the story and the integrity of the context, we have designed the following loss to encourage the model to select continuous frames. Initially, the relevance scores are normalized using a sigmoid function. Then, we incorporate a smoothing mechanism where the continuity score for each frame is updated based on its neighboring frames:

$$C_i = \alpha \cdot C_i + (1 - \alpha) \cdot \frac{C_{i-1} + C_{i+1}}{2} \tag{8}$$

where α is a smoothing coefficient. The continuity loss is defined as:

$$\mathcal{L}_{\text{cont}} = \sum_{i=1}^{T-1} (C_i - C_{i+1})^2 \tag{9}$$

This approach ensures that if one frame is chosen, its neighboring frames are also more likely to be selected, promoting a more coherent selection of key frames.

To sum up, our overall objective can be formulated as:

$$\mathcal{L}_{obj} = \mathcal{L}_{ce} + \lambda_{neg}\mathcal{L}_{neg} + \lambda_{cl}\mathcal{L}_{cl} + \lambda_{cont}\mathcal{L}_{cont}. \tag{10}$$

4 Experiments

4.1 Experimental Settings

Datasets. We primarily evaluate QGAC-TR on the NExT-GQA (Xiao et al., 2024) dataset, as it is currently the only dataset that comprehensively considers QA accuracy, GQA accuracy and detailed localization performance. Additionally, we further validate on the NExT-QA (Xiao et al., 2021a) and STAR (Wu et al., 2021) datasets how the improvements in our model's localization performance positively impact QA performance. See details of the datasets in Appendix B.1.

Evaluation Metrics. We primarily report on the most common metric in VideoQA tasks—accuracy (Acc). For the NExT-GQA benchmark, we also follow the metrics it employs, namely IoP, IoU, and Acc@GQA. See the definitions of these metrics in Appendix B.2.

Baseline Methods. We selected several VideoQA models that use "localize-then-answer" strategy or Transformer architecture as our primary baselines, including VGT (Xiao et al., 2022b), MIST (Gao et al., 2023), TranSTR (Li et al., 2023b), and Temp[CLIP] (Xiao et al., 2024), etc. Detailed descriptions of these models are provided in Appendix B.3. Recent studies (Yu et al., 2024; Xu et al., 2023; Wang et al., 2023b; Ko et al., 2023) also explore using multimodal pretrained models or large language models (MLLMs) for VideoQA, improving accuracy due to large parameter counts and extensive pretraining. However, these models still face challenges with language shortcuts and insufficient use of QA texts, as seen with FrozenBiLM's inferior localization performance in Table 3. Despite the unfairness to our model, we include some pretrained models in our comparisons.

Methods	Acc@C	Acc@T	Acc@D	Acc@All
Co-Mem (Gao et al., 2018)	45.9	50.0	54.4	48.5
HCRN (Le et al., 2021)	47.1	49.3	54.0	48.9
HGA (Jiang and Han, 2020)	48.1	49.1	57.8	50.0
IGV (Li et al., 2022)	48.6	51.7	59.6	51.3
HQGA (Xiao et al., 2022a)	49.0	52.3	59.4	51.8
VGT (Xiao et al., 2022b)	51.6	51.9	63.7	53.7
CoVGT (Xiao et al., 2023)	58.5	57.0	66.8	59.4
TranSTR (Li et al., 2023b)	59.7	60.2	70.0	61.5
Temp[CLIP]* (Xiao et al., 2024)	60.2	59.8	70.0	61.7
VGT(PT) (Xiao et al., 2022b)	52.8	54.5	67.3	55.7
CoVGT(PT) (Xiao et al., 2023)	58.0	58.0	68.4	59.7
HiTeA (Ye et al., 2023)	62.4	58.3	75.6	63.1
InternVideo (Wang et al., 2022)	62.5	58.5	75.8	63.2
SeViLA (Yu et al., 2024)	74.2	69.4	81.3	73.8
QGAC-TR (Ours)	**63.6**	**63.7**	**73.8**	**65.3**

Table 1: QA accuracies of SOTA methods on NExT-QA test set. Acc@C, T, D, denote accuracy for Causal, Temporal, and Descriptive questions respectively. Gray: pretrained models or MLLMs. Same below. *: results reproduced with the official code.

Methods	Acc@I	Acc@S	Acc@P	Acc@F	Acc@All
ClipBERT (Lei et al., 2021b)	39.8	43.6	32.3	31.4	36.7
CLIP (Radford et al., 2021)	39.8	40.5	35.5	36.0	38.0
RESERVE-B (Zellers et al., 2022)	44.8	42.4	38.8	36.2	40.5
Flamingo-9B (Alayrac et al., 2022)	-	-	-	-	43.4
AIO (Wang et al., 2023a)	47.5	50.8	47.8	44.1	47.5
Temp[ATP] (Buch et al., 2022)	50.6	52.9	49.4	40.6	48.4
MIST (Gao et al., 2023)	55.6	54.2	54.2	44.5	51.1
InternVideo (Wang et al., 2022)	62.7	65.6	54.9	51.9	58.7
SeViLA (Yu et al., 2024)	63.7	70.4	63.1	62.4	64.9
QGAC-TR (Ours)	**59.5**	**58.7**	**55.7**	**46.4**	**54.3**

Table 2: QA accuracies of SOTA methods on STAR val set. I: Interaction, S: Sequence, P: Prediction, F: Feasibility. All: Mean.

4.2 Implementation Details

Following the conventions established in (Xiao et al., 2022b; Gao et al., 2023; Li et al., 2023b; Xiao et al., 2024), we sample each video into a sequence of $T = 32$ frames, each encoded by the visual branch ViT-L/14 of CLIP (Radford et al., 2021). For text encoding, we employ the pretrained RoBERTa-base model (Liu et al., 2019) to encode questions and options. During training, we set the batch size to 64 and use the AdamW optimizer with an initial learning rate of 1e-8. The learning rate initially increases linearly to 1e-5 over 500 warmup steps and subsequently decreases gradually through cosine annealing. The hidden dimension d is set to 768, and the intermediate dimension d_{ff} of the FFN in the Transformer layers is set to 3072. Consistent with (Yu et al., 2024), we set $k = 4$ for all datasets, selecting the top 4 frames as localization results. Our model is trained on a single NVIDIA A100 40G GPU and implemented in LAVIS library (Li et al., 2023a), PyTorch.

Methods	Acc@QA	Acc@GQA	mIoP	IoP@0.3	IoP@0.5	mIoU	IoU@0.3	IoU@0.5
IGV (Li et al., 2022)	50.1	10.2	21.4	26.9	18.9	14.0	19.8	9.6
VGT (Xiao et al., 2022b)	50.9	12.7	24.7	26.0	24.6	3.0	4.2	1.4
VIOLETv2 (Fu et al., 2023)	52.9	12.8	23.6	25.1	23.3	3.1	4.3	1.3
Temp[Swin] (Xiao et al., 2024)	55.9	14.4	25.3	26.4	25.3	3.0	3.6	1.7
Temp[CLIP] (Xiao et al., 2024)	59.4	14.7	24.1	26.2	24.1	6.1	8.3	3.7
Temp[CLIP](NG+) (Xiao et al., 2024)	60.2	16.0	25.7	31.4	25.5	12.1	17.5	8.9
FrozenBiLM (Yang et al., 2022)	69.1	15.8	22.7	25.8	22.1	7.1	10.0	4.4
SeViLA (Yu et al., 2024)	68.1	16.6	29.5	34.7	22.9	21.7	29.2	13.8
QGAC-TR (Ours)	**63.6**	**18.3**	**28.3**	**32.8**	**27.7**	**15.7**	**18.6**	**11.7**

Table 3: Grounded QA performance on NExT-GQA test set. The grounded QA accuracy (Acc@GQA) is the percentage of questions that are correctly answered and also visually grounded (i.e., IoP ≥ 0.5). Other results are token from (Xiao et al., 2024).

	CATE	ALoc	NP	OC	CF	Acc@QA	Acc@GQA	mIoP	IoP@0.3	IoP@0.5	mIoU	IoU@0.3	IoU@0.5
(a)						59.4	14.7	24.1	26.2	24.1	6.1	8.3	3.7
(b)	✓					60.6	15.1	25.0	26.5	24.1	8.2	10.5	5.3
(c)	✓	✓				61.0	16.3	25.5	29.3	25.4	12.1	13.4	7.6
(d)	✓	✓	✓			61.7	17.7	26.7	31.1	26.7	13.6	17.4	9.8
(e)	✓	✓		✓		62.6	16.8	26.0	30.4	26.8	13.4	17.1	9.3
(f)	✓	✓	✓	✓		63.3	17.8	27.8	32.2	27.5	15.6	18.8	11.2
(g)	✓	✓	✓	✓	✓	63.6	18.3	28.3	32.8	27.7	15.7	18.6	11.7

Table 4: Ablation study on NExT-GQA test set. CATE: Cross-Attention Transformer Encoder. ALoc: Adaptive Localizer. NP: Negative Pair Learning. OC: Contrastive Learning for Option Calibration.

4.3 Comparison with State-of-the-arts

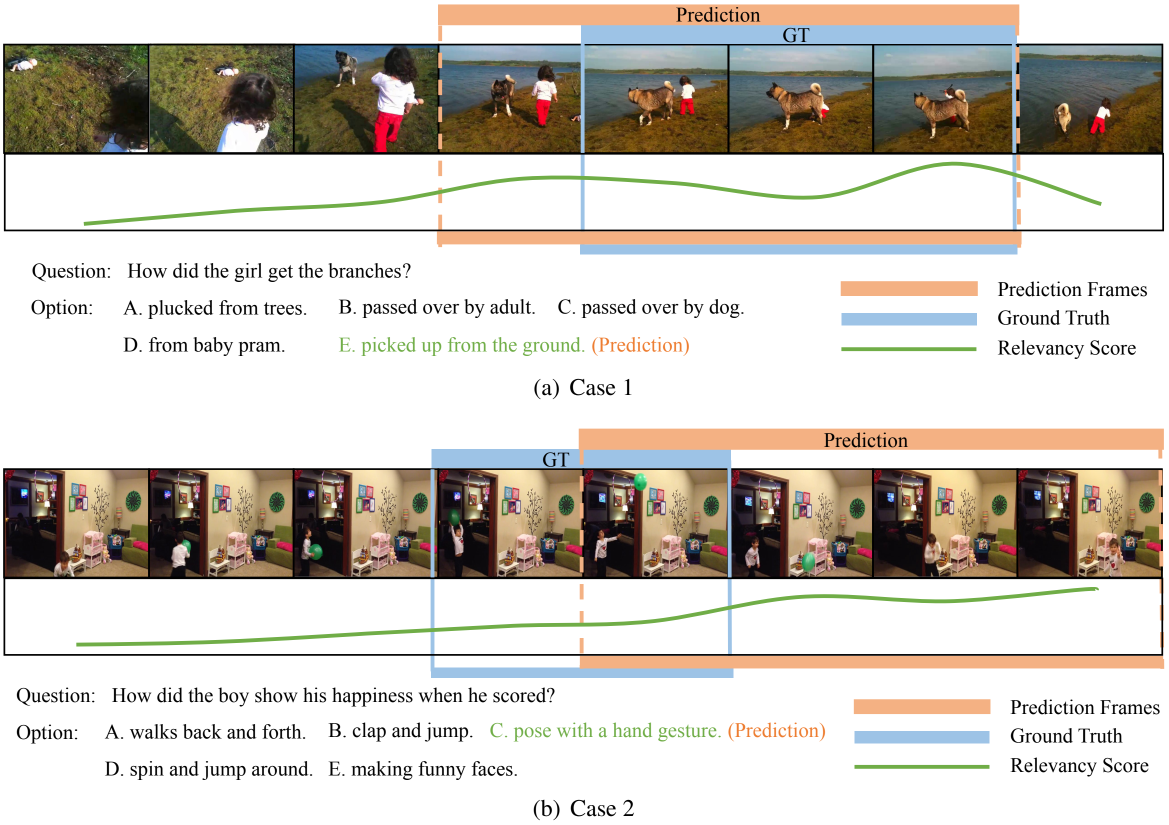
As shown in Tables 1, 2 and 3, our proposed method achieved state-of-the-art performance across all three datasets, surpassing nearly all existing methods, including large-scale pretrained models and those using additional training data (marked in gray in the tables). For instance, on NExT-QA, QGAC-TR outperformed video-language pretrained models (Ye et al., 2023; Wang et al., 2022) and (Xiao et al., 2022b), which used extra training data and fine-grained object features. The IoP and IoU metrics in Table 3 show that QGAC-TR achieves more accurate localization through its two self-supervised tasks, and the increase in Acc@GQA indicates that better localization leads to more accurate answers. Note that the notable performance of SeViLA (Yu et al., 2024) on localization metrics is due to its supervised fine-tuning on labeled datasets like (Lei et al., 2021a).

QGAC-TR shows a larger improvement on NExT-QA compared to STAR (4.6% *vs.* 3.2%). This is likely because NExT-QA videos are longer and have more complex plots. Guided by questions and option calibration, QGAC-TR can better localize relevant parts and filter out noise in these longer videos, aligning with its design intent. Regarding question types, our model's overall performance improvement is mainly due to better handling of questions requiring causal and temporal reasoning. For example, in NExT-QA, it improved Causal (+3.4) and Temporal (+3.9) questions, and in STAR, it improved Interaction (+3.9), Sequence (+4.5), and Prediction (+1.5) questions. These question types require deeper understanding, indicating that QGAC-TR's "localize-then-answer" approach enhances reasoning capabilities.

As shown in Fig.2, QGAC-TR exhibits a smaller performance decline when the language model is frozen. This demonstrates that our model focuses more on understanding video content, effectively mitigating the issue of language shortcuts.

4.4 Ablation Study

In Table 4, we investigate the effectiveness of each component. For the baseline (a), we adopt Temp[CLIP] (Xiao et al., 2024). Rows (b) to (g) clearly validate the advantages of each component. Specifically, (b) demonstrates that the cross-attention Transformer encoder effectively extracts question-relevant video representations. (c) shows that, compared to methods selecting video frames based on attention scores from questions and video frames (Gao et al., 2023; Li et al., 2023b,c), our adaptive predictor allows dynamic adjustment of its representation according to the specific video context. This means the predictor can better adapt to different input conditions rather than merely relying on a static attention mechanism. (d) and (e)



(a) Case 1

(b) Case 2

Figure 5: Visualization of results predicted by QGAC-TR. Predicted and ground-truth moments are bounded by the lines. The cases are selected from NExT-QA/NExT-GQA (Xiao et al., 2021a, 2024).

show that each self-supervised learning task significantly improves performance, and when combined in (f), performance is further enhanced, demonstrating their effectiveness in leveraging question and option text to improve precise localization, thereby explainably increasing question-answering accuracy. The performance improvement in (g) relative to (f) supports our hypothesis that "continuous scenes better maintain contextual coherence, aiding model comprehension." Furthermore, (g), which incorporates all designed components, achieves the best results.

4.5 Qualitative Analysis

We explore how video representations, guided by questions and option calibration, respond sensitively to changes in the text query context. As shown in Figure 5, we select two typical cases and visualize their localization scores.

In Case 1, the scores of video frames closely related to the question are significantly higher than other parts. Notably, our model assigns the highest relevance score to the moment when the girl lifts the branch (the seventh frame). However, the key frame is the sixth frame, where the scene of the girl squatting down to pick up the branch is obscured by the dog, resulting in a relatively lower score for that frame. Despite this, QGAC-TR accurately selects this key frame, demonstrating its precise localization capability. In Case 2, there is a discrepancy between the ground truth and our frame localization. The ground truth focuses on the “when he scored” aspect mentioned in the question, failing to cover all frames relevant to the query. In contrast, our QGAC-TR, guided by the question, not only pays attention to “when he scored” but also captures “happiness,” achieving higher relevance scores in the sixth to eighth frames. Specifically, the sixth frame captures the boy’s smile when he scores, and the eighth frame shows him posing with a hand gesture.

This result demonstrates that our model can more comprehensively interpret the intent of the question and perform more precise video content localization, providing a richer and more detailed visual response to the query. The sensitivity to question nuances and high localization accuracy highlight the strengths of our approach, especially in handling complex video question answering tasks.

5 Conclusion

QGAC-TR introduces an innovative approach to VideoQA by using question and option texts for guided localization in the absence of labels. It employs a Transformer-based architecture with cross-attention and an input-adaptive localizer to enhance localization precision. Additionally, two self-supervised learning tasks further improve the model's understanding and localization of video content. The experimental results show that QGAC-TR not only achieves competitive QA accuracy but also excels in localization precision. This demonstrates that it deeply exploits the helpfulness of question and option texts for localization, and effectively mitigates the model's reliance on language shortcuts and spurious correlations.

Acknowledgments

This work is supported by the National Natural Science Foundations of China under Grant (62372060).

Limitations

Model Performance and Underutilization of Visual Evidence. Although current methods have achieved impressive QA accuracy, as shown in Table 3, the accuracy on visually grounded VideoQA (Acc@GQA) is significantly lower. This discrepancy suggests that most correct responses are not based on correct localization of visual evidence. While our QGAC-TR model demonstrates advanced visual localization capabilities, its performance on Acc@GQA still significantly lags behind Acc@QA, indicating a vast area for further exploration in visually grounded VideoQA.

Limitations of Self-Supervised Contrastive Learning. Our approach includes a self-supervised contrastive learning task that relies on the use of options. Current VideoQA research predominantly focuses on multiple-choice datasets, whereas generative open-ended VideoQA, which is more pragmatically relevant, cannot utilize this self-supervised task. Although theoretically, the localization component of our trained model could be adapted for open-ended VideoQA tasks, the high cost of collecting quality multiple-choice data remains prohibitive. Training models to proactively uncover key visual clues not mentioned in questions without relying on options remains an unresolved critical challenge.

Relation with Large Language Models. The advent of large language models (LLMs) has reshaped the performance benchmarks for various tasks. Although our method does not directly leverage these LLMs, it has matched or even exceeded some LLM-based methods in certain cases. However, LLMs have a higher potential ceiling. Currently, LLM-based approaches have not addressed the two major issues we identified in Sec. 1. Exploring how to integrate and adapt our method with LLMs is an important future direction for our research.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171.

Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. 2022. (2.5+ 1) d spatio-temporal scene graphs for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 444–453.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng

Liu. 2023. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585.

Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 245–253. IEEE.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. 2021. Loss re-scaling vqa: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing*, 31:227–238.

Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116.

Sungdong Kim, Jin-Hwa Kim, Jiyoung Lee, and Minjoon Seo. 2023. Semi-parametric video-grounded text generation. *arXiv preprint arXiv:2301.11507*.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2021. Hierarchical conditional relation networks for multimodal video question answering. *International Journal of Computer Vision*, 129(11):3027–3050.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021a. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021b. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. 2023a. Lavis: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8658–8665.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023b. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878.

Yicong Li, Xun Yang, An Zhang, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023c. Redundancy-aware transformer for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3172–3180.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021a. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244.

Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078.

Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. 2021b. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1698–1707.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 15–24.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multimodal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15526–15535.

Tianwen Qian, Ran Cui, Jingjing Chen, Pai Peng, Xiaowei Guo, and Yu-Gang Jiang. 2023. Locate before answering: Answer guided question localization for video question answering. *IEEE Transactions on Multimedia*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. 2018. Find and focus: Retrieve and localize video events with natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 200–216.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023a. All in one: Exploring unified video-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming Lin, and Shan Yang. 2023b. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2312.08367*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*.

Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. 2017. Deep learning for video classification and captioning. In *Frontiers of multimedia research*, pages 3–29.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021a. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022a. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022b. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer.

Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. 2023. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021b. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Yuanxing Xu, Yuting Wei, and Bin Wu. 2023. Query-aware long video localization and relation discrimination for deep video understanding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9591–9595.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2023. X 2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019b. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1230–1238.

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019c. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Details of Related Works

A.1 Video Temoral Grounding

Proposal-based methods rely on predefined proposals like sliding windows (Anne Hendricks et al., 2017; Liu et al., 2018; Ge et al., 2019; Yuan et al., 2019; Zhang et al., 2019b) and temporal anchors (Chen et al., 2018; Yuan et al., 2019; Zhang et al., 2019a,c; Liu et al., 2020) or learn to generate proposals (Shao et al., 2018; Xu et al., 2019; Zhang et al., 2020; Liu et al., 2021a; Xiao et al., 2021b). Proposal-free methods encode multimodal knowledge and predict time spans using regression heads.

B Details of Experimental Settings

B.1 Datasets

We evaluated the effectiveness of QGAC-TR on three public and popular datasets, which focus on temporal and causal reasoning and require a deep understanding of the video content, namely NExT-QA(Xiao et al., 2021a), STAR(Wu et al., 2021), and NExT-GQA(Xiao et al., 2024).

- **NExT-QA** (Xiao et al., 2021a) is a VideoQA benchmark for causal and temporal reasoning. It contains a total of 5,440 videos with an average length of 44s and about 52k questions. NExT-QA contains 3 different question types: Temporal (Tem.), Causal (Cau.), and Description (Des.).

- **STAR** (Wu et al., 2021) is a benchmark for situated reasoning, which contains 22K video clips with an average length of 12s, along with 60K questions. STAR contains 4 different question types: Interaction (Int.), Sequence (Seq.), Prediction (Pre.), and Feasibility (Fea.).

- **NExT-GQA** (Xiao et al., 2024) is a very recent and currently the only benchmark for visually grounded VideoQA. It retains the Temporal and Causal type questions in NExT-QA and provides GT moment labels for the validation set and test set to evaluate the model's localization ability.

B.2 Evaluation Metrics

We primarily report on the most common metric in VideoQA tasks—accuracy (Acc). For the NExT-GQA benchmark, we also follow the metrics it employs, namely IoP, IoU, and Acc@GQA. IoP represents the proportion of the intersection between the model-located time interval and the actual time interval relative to the actual time interval. IoU denotes the ratio of the intersection between the model-located time interval and the actual time interval to their union. Acc@GQA refers to the percentage of questions that are answered correctly and accurately located (IoP>=0.5).

B.3 Baseline Methods

We initially selected several VideoQA models that also utilize Transformer architecture or the "localize-then-answer" strategy, including (Xiao et al., 2022b; Gao et al., 2023; Li et al., 2023b; Xiao et al., 2024) as our primary baseline models. VGT (Xiao et al., 2022b) refines granularity to the object level and aligns it, capturing temporal and spatial information through Transformers and GNNs, and aggregates video representations using a hierarchical structure (Xiao et al., 2022a). However, it lacks a dedicated localization stage. As shown in Figure 2, VGT's performance improvements mainly depend on language shortcuts, and its performance significantly drops below (Xiao et al., 2022a) when the language model is frozen. Both MIST (Gao et al., 2023) and TranSTR (Li et al., 2022) employ the "localize-then-answer" strategy but only indirectly train the model's localization ability using QA labels, which could lead to incorrect localizations. These models also localize using attention scores between video segments or frames and questions, exhibiting low flexibility and adaptability with insufficient generalization ability. Temp[CLIP] (Xiao et al., 2024), achieving high accuracy with a very simple dual architecture, performs poorly in localization. Therefore, Temp[CLIP](NG+) enhances localization accuracy through cross-modal self-supervision by using video moments marked by Gaussian masks as anchors, optimizing the proximity of question-answer pairs in feature space while ensuring they are further from unrelated pairs. However, it still does not fully utilize the guidance and calibration role of question and option texts.