This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Discovering Spatio-Temporal Rationales for Video Question Answering

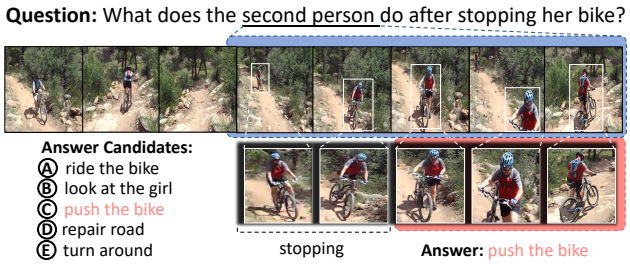
Yicong Li[1], Junbin Xiao[1]*, Chun Feng[2], Xiang Wang[2]*, Tat-Seng Chua[1]
[1]National University of Singapore, [2]University of Science and Technology of China,
liyicong@u.nus.edu, fengchun3364@mail.ustc.edu.cn, xiangwang1223@gmail.com
junbin@comp.nus.edu.sg, dcscts@nus.edu.sg

Abstract

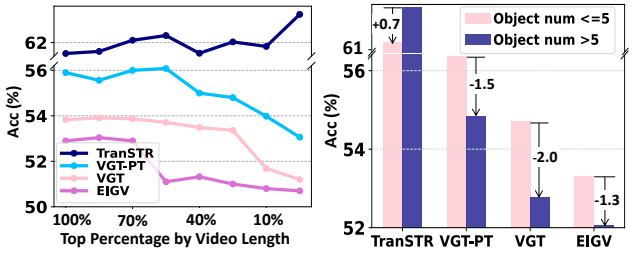
*This paper strives to solve complex video question answering (VideoQA) which features long video containing multiple objects and events at different time. To tackle the challenge, we highlight the importance of identifying question-critical temporal moments and spatial objects from the vast amount of video content. Towards this, we propose a **S**patio-**T**emporal **R**ationalization (STR), a differentiable selection module that adaptively collects question-critical moments and objects using cross-modal interaction. The discovered video moments and objects are then served as grounded rationales to support answer reasoning. Based on STR, we further propose TranSTR, a Transformer-style neural network architecture that takes STR as the core and additionally underscores a novel answer interaction mechanism to coordinate STR for answer decoding. Experiments on four datasets show that TranSTR achieves new state-of-the-art (SoTA). Especially, on NExT-QA and Causal-VidQA which feature complex VideoQA, it significantly surpasses the previous SoTA by 5.8% and 6.8%, respectively. We then conduct extensive studies to verify the importance of STR as well as the proposed answer interaction mechanism. With the success of TranSTR and our comprehensive analysis, we hope this work can spark more future efforts in complex VideoQA. Code will be released at https://github.com/yl3800/TranSTR.*

1. Introduction

The great success of self-supervised pretraining with powerful transformer-style architectures [5, 12, 17, 9, 39, 42] has significantly boosted the performance of answering simple questions (*e.g.*, "what is the man doing") on short videos (*e.g.*, 3∼15s) [13, 37]. The advances thus point towards complex video question answering (VideoQA), that features long video containing multiple objects and events [34, 20, 45]. Compared with simple VideoQA, complex VideoQA poses several unique challenges:



(a) A example of long video (52s) with multiple objects, the question-related frames and objects are located to support the reasoning.



(b) Accuracy by video length. (c) Accuracy by object number.

Figure 1: (a) Illustration of complex VideoQA, in which the videos are longer and the questions involve multiple objects and events at different time. (b) Prediction accuracy grouped by video length. We first sort all samples by video length, then select top x% to calculate accuracy. (c) Accuracy grouped by whether the video has more than 5 objects. All results are reported on NExT-QA test set [34]. Figure (b) and (c) show that our method TranSTR performs much better than the previous SoTAs for question answering of long videos with multiple objects.

1) **Longer videos with multiple objects interacting differently at different times.** The long video and rich visual content indispensably bring more background scenes that include massive question-irrelevant video moments and objects. For example, to answer the question in Fig. 1a, only the interaction between object "person" and "bike" on the last three frames encloses the answer information, leaving the massive rest as background. These backgrounds, if not

*Corresponding author.

filtered properly, will overwhelm the critical scene and interfere with answer prediction. 2) **Harder negative answer as distractors.** Negative answers in complex VideoQA are typically tailored for each video instance. Due to the massive video content, the vast question-irrelevant scene provides an ideal foundation to build a hard negative candidate as distractor. The hard negatives are very similar to the correct answer but correspond to a different video moment or object. For example, the answer candidate "A.ride the bike" of Fig. 1a, though irrelevant to the question, corresponds to a large part of the video. As a result, these distractors can seriously derail the prediction if not properly modeled.

In light of the challenges, current methods (both pretrained and task-specific architectures) hardly perform well on complex VideoQA. In Figs. 1b and 1c, we use the video length and number of objects[1] to indicate the complexity of the video questions. We can see that current methods suffer a drastic performance drop when video length increases or more objects are involved. The reason can be provided from two aspects: **First**, confronting long video and multiple objects, pretrained methods suffer from a domain gap. Because they are typically pretrained with short videos and simple questions, [17, 19] where the answer can be easily captured via a static frame, without fine-grained reasoning over multiple objects in a long video. While recent task-specific methods exploit object-level representation for fine-grained reasoning [4, 29, 35, 36], they exhibit limited generalization ability, as they handle different videos with only a fixed number of frames and objects, and cannot adapt to lengthy and varied visual content, which rigidity undermines their adaptability to a wide range of video content. **Second**, to model the answer candidate, prevailing designs [13, 10, 8, 16, 35] append the candiate to the question and treat the formed question-answer sequence as a whole for cross-modal learning. However, this makes the answer candidate directly interact with the whole video content, which gives rise to a strong spurious correlation between the hard negative candidates (*e.g.* "A. ride the bike" in Fig. 1a) and the question-irrelevant scenes (*e.g.* " riding" scene in first three frames), leading to a false positive prediction.

In this regard, we propose TranSTR, a Transformer-style VideoQA architecture that coordinates a Spatio-Temporal Rationalization (STR) with a more reasonable video-text interaction pipeline for candidate answer modeling. STR first temporally collects the critical frames from a long video, followed by a spatial selection of objects on the identified frames. By further fusing the selected visual objects and frames via light-weight reasoning module, we derive spatial and temporal rationales that exclusively support answering. In addition to STR, we circumvent the spurious correlation in the current modeling of answer candidates by formulating a more reasonable video-text interaction pipeline, where the question and answer candidates are separately (instead of being appended as a whole) fed to the model at different stages. Specifically, before rationale selection, only the question is interacted with the video to filter out the massive background content while keeping question-critical frames and objects. After that, a transformer-style answer decoder introduces the answer candidates to these critical elements to determine the correct answer. Such a strategy prevents the interaction between the hard negative answers and the massive background scenes, thus enabling our STR to perform better on complex VideoQA (see TranSTR in Figs. 1b and 1c). It is worth noting that STR and the answering modeling are reciprocal. Without STR's selection, all visual content will still be exposed to answer candidates. Without our answer modeling, STR could identify the question-irrelevant frame and object as critical. Thus, the success of TranSTR is attributed to the integration of both.

Our contributions are summarized as follows:

- We analyze the necessity and challenge of complex VideoQA. To solve the task, we identify the importance of discovering spatio-temporal rationales and preventing spurious correlation in modeling candidate answers.
- We propose TranSTR that features a spatio-temporal rationalization (STR) module together with a more reasonable candidate answer modeling strategy. The answer modeling strategy is independently verified to be effective in boosting other existing VideoQA models.
- We perform extensive experiments and achieve SoTA performance on four popular benchmarks, especially for ones that features complex VideoQA (NExT-QA [34] +5.8%, CausalVid-QA [20] +6.8%)

2. Related Works

Video Question Answering (VideoQA). Substantiated as a fundamental extension of ImageQA, VideoQA has enlarged its definition by adding a temporal extension. According to the pre-extracted feature granularity, existing methods either use the frame-level features or incorporate object features for fine-grain reasoning. In task-specific designs, focusing on simple questions and short videos, earlier efforts tend to model the video sequence as a visual graph using purely frame features. As the pioneer of graph-based structure, [16] and [28] build their typologies based on the heterogeneity of input modality, while [11] enables progressive relational reasoning between multi-scale graphs. Recently, the emergence of complex VideoQA benchmarks [34, 20, 44] has prompted studies on long video with multiple visual entities. In this regard, Another line of research has prevailed by processing video as multi-level hierarchy.

[1] We acquire the object number using annotation of video relation detection dataset [32], which shares same source video as NExT-QA.

[18] first build a bottom-up pathway by assembling information first from frame-level, then merging to clip-level. The following works [4, 35] extend the hierarchy into the object-level, where a modular network is designed to connect objects on the same frame. Most recently, [36, 25] establish its improvement by enabling relation reasoning in a sense of object dynamics via temporal graph transformer. Despite effectiveness, the current designs unanimously rely on a fixed number of frames and objects, which severely compromises their transferability across diverse video instances. In sharp contrast, our method works in a fully adaptive manner to explicitly select frames and objects for reasoning over different circumstances, which demonstrates superior generalization ability. Aside from these architectural advancements, a purely data-driven approach, cross-modal pretraining, has gained increasing popularity. Related approaches take advantage of the abundant vision-text data available on the web to train transformer-style models in a self-supervised manner [36, 39, 38]. By leveraging large-scale pretraining, these models can learn to generalize better and potentially improve performance on various video-language tasks. However, they are typically pretrained with short videos and simple questions, which seriously hinder their applicability to long videos

Rationalization. In pursuit of explainability, the recent development of DNN is encouraged to reveal the intuitive evidence of their prediction, *i.e*. the rationales. As one of the prevailing practices, the rationalization has been extended from the NLP community[31] to the Graph [33] and Vision field [43]. Recently, this development also stems from the multi-modal community [24, 40, 41, 15, 14]. [27] and [7] proposes ImageQA-based tasks that inquire about additional textual evidence, [20] brings this idea to the videoQA. Despite the progress, the recent solution focus on the rationale only at frame level, and they either require a rationale finder with heavy computation overhead [22, 23] or needs to be trained in a data-hungry contrastive manner [21]. TranSTR, however, identifies both critical frames and objects from an efficient cross-modal view. Also, distinct from the token reduction method in transformer literature, which trades accuracy for efficiency. Rationalization intends to improve performance [3]. The intuition behind is that, if a model can find the causal part, they have the potential to ignore the noise.

3. Preliminaries

Modeling. Given the video V and the question Q, the VideoQA model $\phi(V, Q)$ aims to encapsulate the visual content and linguistic semantics and choose the predictive answer $\hat{A}$ from the answer candidates. Typically, an entropy-based risk function $\mathcal{L}(\phi(V, Q), A)$ is applied to approach the ground-truth answer A.

Data representation. We uniformly sample T clips and keep the middle frame of each clip to represent a video. Then, for each frame, we extract a frame feature $\mathbf{f}_t$ via a pretrained image recognition backbone and S object features $\mathbf{o}_{t,s}$ using pretrained object detector, where t, s denotes the s-th object on the t-th frame. To represent the text, we encode the question as a sequence of L tokens using a pretrained language model and obtain a textual representation $\mathbf{q}_l$ for each of them. The visual backbones are frozen during training while the language backbone is fine-tuned end-to-end as in [36]. To project the representations into a common d-dimensional space, we apply a three linear mappings on $\mathbf{f}_t$, $\mathbf{o}_{t,s}$, and $\mathbf{q}_l$, respectively, and thus acquire $\mathbf{F} = \{\mathbf{f}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$, $\mathbf{O} = \{\mathbf{o}_{t,s}\}_{t=1,s=1}^{T,S} \in \mathbb{R}^{T\times S\times d}$, and $\mathbf{Q} = \{\mathbf{q}_l\}_{l=1}^{L} \in \mathbb{R}^{L\times d}$ to denote the frame, object, and question features, respectively.

4. Method

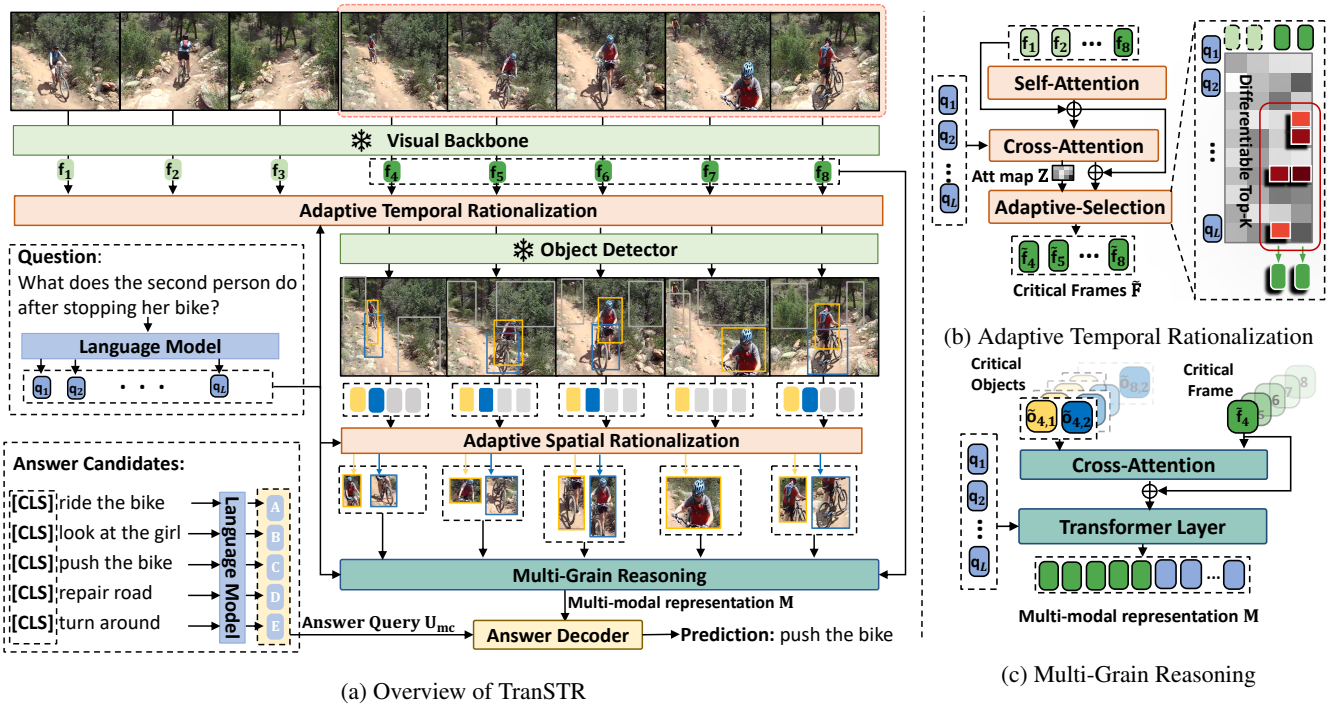
As shown in Fig. 2a TranSTR consists of three main components: Spatio-Temporal Rationalization (STR), Multi-Grain Reasoning (MGR), and Answer Decoder. First, STR follows a two-step selection process, where Adaptive Temporal Rationalization (TR) is first performed for frame selection, followed by object selection via Adaptive Spatial Rationalization (SR). Next, the selected frames and objects are then combined through MGR, which generates multi-modal representations by fusing with the question. Finally, based on the multi-modal representations, the answer decoder takes the combination of answer candidates as queries and predicts an answer. In this section, we provide a detailed illustration of each module.

4.1. Spatio-Temporal Rationalization (STR)

STR aims to find the question-critical frames and objects in a fully adaptive and differentiable manner, which comprises two components: an Adaptive Temporal-Rationalization (TR) that identifies the critical frames and an Adaptive Spatial-Rationalization (SR) that pinpoints the critical objects on the identified frames.

4.1.1 Adaptive Temporal Rationalization (TR)

To identify the critical frames from the video, TR takes the encoded video frames $\mathbf{F} \in R^{T\times d}$ as input and adaptively selects question-critical frames from a cross-modal view. As shown in Fig. 2b, a self-attention layer is first adopted to contextualize $\mathbf{F}$. Then, a cross-attention is applied by taking contextualized frame feature $\mathbf{F}'$ as query and question embedding $\mathbf{Q}$ as key and value, which yields the frame tokens $\mathbf{F}'' \in R^{T\times d}$ and the cross-attention map $\mathbf{Z} \in R^{T\times L}$.



(a) Overview of TranSTR

(b) Adaptive Temporal Rationalization

(c) Multi-Grain Reasoning

Figure 2: TranSTR (a) contains three components: the Spatio-Temporal Rationalization which adaptively selects critical frames and objects, the Multi-Grain Reasoning (c) which forms multi-modal representation by integrating the critical frames and objects together with the question semantics, and the Answer Decoder which bring in answer candidates and make a prediction based on the multi-modal representation. Notably, STR follows a "Temporal then Spatial" rationalization process, where the two steps leverage a similar deisgn. In (b), we illustrate this design using Adaptive Temporal Rationalization.

$$\mathbf{F}' = \text{Self-Attention}(\mathbf{F}) + \mathbf{F}, \tag{1}$$

$$\mathbf{F}'', \mathbf{Z} = \text{Cross-Attention}(\mathbf{F}', \mathbf{Q}) + \mathbf{F}'. \tag{2}$$

For brevity, we omit the superscript in $\mathbf{F}''$ and use $\mathbf{F}$ to denote the resulting frame tokens.

Naturally, each value in the cross-attention map indicates the cross-modal interaction activeness between a frame and a question token. To enable an adaptive frame selection that caters to different videos, we collect the K_f interactions of the highest attention score from the cross-attention map $\mathbf{Z}$, then gather their corresponding frame tokens $\tilde{\mathbf{F}} \in R^{C \times d}$ as a subset of $\mathbf{F}$, where C is the number of critical frames. This process is formally given by:

$$\tilde{\mathbf{F}} = \text{Adaptive-Selection}_K(\mathbf{F}, \mathbf{Z}) \quad \text{s.t. } K = K_f. \tag{3}$$

Notably, by gathering 1-D tokens from the 2-D interaction view, we enable an adaptive collection of $\tilde{\mathbf{F}}$ with much fewer tokens being selected as critical frames (*i.e.* $C \ll T$ and $C \ll K_f$). It is worth noting that the interaction selection via vanilla hard Top-K produces a discrete selection, making it inapplicable for end-to-end training. We address this issue by adopting a differentiable Top-K using the perturbed maximum method [1], which has empirically shown advantages over other differentiable technique (*c.f.* Tab. 4)

4.1.2 Adaptive Spatial Rationalization (SR)

Given the embedding of the selected frames $\tilde{\mathbf{F}}$, SR aims to pinpoint the question-critical objects in each frame. To achieve that, SR enable an adaptive object selection similar to Eqs. (1) to (3). Specifically, for the t-th critical frame $\tilde{\mathbf{f}}_t$, we first feed SR with fixed S object features detected on that frame, then collect top K_o interactions from its cross-modal attention map with question embedding. Finally, by gathering their corresponding object tokens, we obtain the critical object feature $\tilde{\mathbf{o}}_t \in R^{C_t \times d}$, where C_t denotes the number of critical objects on t-th critical frame. It is worth noting that, SR is applied independently to each frame, thus, different frames can adapt to different numbers of the critical objects C_t, even if we keep K_o constant for all frames.

4.2. Multi-Granularity Reasoning (MGR)

MGR aims to enhance the frame-level representation with fine-grained object embedding, while modeling the video dynamic together with question semantics. As shown in Fig. 2c, MGR first applies intra-frame aggregation via a cross-attention, which takes the frame feature of the t-th critical frame $\tilde{\mathbf{f}}_t \in R^{1 \times d}$ as query, and all critical objects in t-th frame $\tilde{\mathbf{o}}_t \in R^{C_t \times d}$ as key and value to generate an

object-enhanced representation $\mathring{\mathbf{f}}_t \in R^{1\times d}$ for the t-th frame:

$$\mathring{\mathbf{f}}_t = \text{Cross-Attention}(\tilde{\mathbf{f}}_t, \tilde{\mathbf{o}}_t) + \tilde{\mathbf{f}}_t. \quad (4)$$

By doing so to all C critical frames, we acquire $\mathring{\mathbf{F}} \in R^{C\times d}$ as the object-enhanced frame representation for all critical frames. Next, a transformer layer is adopted to establish cross-frame dynamics, which takes in the concatenation of $\mathring{\mathbf{F}}$ and question tokens $\mathbf{Q}$, and yields multi-modal representations $\mathbf{M} \in R^{(C+L)\times d}$ for answer decoding:

$$\mathbf{M} = \text{Transformer-Layer}([\mathring{\mathbf{F}}; \mathbf{Q}]), \quad (5)$$

where $[;]$ denotes concatenation operation.

4.3. Answer Decoding

Existing methods [36, 22, 21] concatenate a question with answer candidates. As analyzed in Sec. 1, these methods suffer from a spurious correlation between negative candidates and question-irrelevant video scenes. To circumvent this issue in spatio-temporal rationalization, we employ a transformer-style decoder that takes as input the question-critical multi-modal representations $\mathbf{M}$ and the representations of the candidate answers to determine the correct answer. We detail our implementations for multi-choice QA and open-ended QA in the next sub-sections.

4.3.1 Multi-Choice QA

In Multi-Choice QA, answer candidates are given as $|A_{mc}|$ sentences or short phrases that are tailored for each video-question pair. Therefore, reasoning on Multi-Choice QA typically requires fine-grain inspection of the video content as well as the interaction between the video and the candidate answers. To this end, we first prepend a [CLS] token to each answer candidate and feed the sequences to the same language model used for question encoding. Then, we gather the output of the '[CLS]' tokens for all encoded answer candidates, and form the answer query $\mathbf{U}_{mc} \in \mathbb{R}^{|A_{mc}|\times d}$. During decoding, we feed a transformer decoder with $\mathbf{U}_{mc}$ as query to interact with the multi-modal representation $\mathbf{M}$, which yields the decoded representation $\mathbf{H}_{mc} \in \mathbb{R}^{|A_{mc}|\times d}$ as:

$$\mathbf{H}_{mc} = \text{Transformer-Decoder}(\mathbf{U}_{mc}, \mathbf{M}). \quad (6)$$

Notably, since the correctness of a answer candidate is invariant to its position, answer query is free of position encoding. Finally, we apply a linear projection on $\mathbf{H}_{mc}$ to get the answer prediction $\hat{A}_{mc} \in \mathbb{R}^{|A_{mc}|}$,

$$\hat{A}_{mc} = \text{Linear}(\mathbf{H}_{mc}). \quad (7)$$

4.3.2 Open-Ended QA

Open-Ended setting provides $|A_{oe}|$ simple-form answer candidates (typically a single word) that are shared among all question instances, which makes the whole candidates set it too large to be processed as Multi-Choice setting. (*i.e.* $|A_{oe}| \gg |A_{mc}|$). Instead, we take inspiration from DETR [2], and initialize a single learnable embedding $\mathbf{U}_{oe} \in \mathbb{R}^d$ as answer query. Analogous to Multi-Choice setting, we feed $\mathbf{U}_{oe}$ to the transformer decoder together with $\mathbf{M}$ and acquire the decoded representation $\mathbf{H}_{oe} \in \mathbb{R}^d$ similar to Eq. (6). As a result, we obtain the prediction $\hat{A}_{oe} \in \mathbb{R}^{|A_{oe}|}$ by projecting $\mathbf{H}_{oe}$ to the answer space $\mathbb{R}^{|A_{oe}|}$ via a linear layer:

$$\hat{A}_{oe} = \text{Linear}(\mathbf{H}_{oe}). \quad (8)$$

During training, we establish our objective on a cross-entropy loss. For inference, the differentiable Top-K is replaced with vanilla hard Top-K for better efficiency.

Table 1: Dataset Statistics. MC and OE denote Multi-Choice and Open-Ended QA respectively.

Dataset	Challenge	#QA pair	V-Len	Q-Len	QA
NExT-QA	Causal & Temporal	48K	44s	11.6	MC
Causal-VidQA	Evidence & Commonsense	161K	9s	9.5	MC
MSVD-QA	Description	50K	10s	6.6	OE
MSRVTT-QA	Description	244K	15s	7.4	OE

5. Experiments

Datasets: Experiments were conducted on four benchmarks to evaluate TranSTR's performance from different aspects. Specifically, the recent NExT-QA [34] and Causal-VidQA [20] datasets challenge models with complex VideoQA tasks using a Multi-Choice setting, which aims to test their temporal reasoning ability with complex causal and commonsense relations. In addition, MSVD-QA [37] and MSRVTT-QA [37] employ an Open-Ended setting and emphasize the description of video objects, activities, and their attributes. Their statistics are presented in Table 1.

Implementation Details: Following the convention established in [36], we sample each video as a sequence of T=16 frames, where each frame is encoded by a ViT-L [6] model that pre-trained on ImageNet-21k. To extract object-level features, we employ a Faster-RCNN [30] model that pre-trained on the Visual Genome and detect S=20 objects in each frame. For the textual encoding, we adopt a pretrained Deberta-base model [12] to encode the question and answer. During training, we optimize the model using an Adam optimizer with a learning rate of 1e-5, and set the hidden dimension d to 768. For the hyper-parameters, we set K_f=5 and K_o=12 for all datasets.

Next, we show our experimental results to answer the following questions:

- **Q1:** How is TranSTR compared with the SoTA?
- **Q2:** How effective are the proposed components?
- **Q3:** What learning pattern does the rationlizer capture?

5.1. Main Result (Q1)

In Tabs. 2 and 3, we show that TranSTR outperforms SoTAs on all question types. Our observations are as follows:

QA setting. Comparing the performance of TranSTR with state-of-the-art methods on four datasets, we observe that TranSTR achieves a greater improvement on Multi-Choice (NExT +5.8% and Causal-Vid +6.8%) compared to Open-End QA (MSVD +3.5% and MSRVTT +3.4%). This can be explained from two aspects: (1) Unlike Open-Ended datasets that contain simple questions and short videos, Multi-Choice datasets (NExT and Causal-Vid) focus on complex VideoQA, in which composite question sentences with long videos and multiple objects (see Tab. 1) makes the identifying and inspecting of critical scenes necessary. This aligns with TranSTR's design philosophy of removing redundancy and explicitly exposing critical elements. Therefore, TranSTR achieves a larger gain on complex VideoQA. (2) In multi-choice QA, SoTA methods often append candidate answers to the question during encoding, which can create a spurious correlation between the negative answer and the question-irrelevant scene, leading to a false prediction. However, such an issue is less significant in open-ended QA, where each answer candidate is treated as a one-hot category without semantic meaning. Thus, the decoder of TranSTR brings extra benefits to the multi-choice setting, and result in larger gains compared to open-ended QA.

Question-type. Based on the analysis of Multi-Choice datasets, we observe that the improvement in overall performance of TranSTR is largely due to the enhancement in answering composite questions (including Acc@C and Acc@T in NExT-QA, Acc@E and Acc@P and Acc@C in Causal-VidQA) that require deeper understanding such as causal relations and counterfactual thinking, compared to the descriptive question type (Acc@D:+1.8∼2.7%). This demonstrates TranSTR's outstanding reasoning ability for complex VideoQA tasks. In particular, for Causal-VidQA, TranSTR shows a significant improvement in answering reason-based questions (Acc@P:AR +10.5%, Acc@C:AR +8.3%). Because questions of this type require the model to justify its prediction by selecting the correct evidence, which aligns with the concept of rationalization in TranSTR's design philosophy. Therefore, TranSTR's rationalization mechanism enables it to perform optimally in answering reason-based questions.

5.2. In-Depth Study (Q2)

5.2.1 Ablative Results

We validate the key components of TranSTR by performing model ablation and discussing other implementation alternatives. As shown in Tab. 4, we first study the effectiveness of TranSTR by removing both STR and decoder ("w/o STR & decoder"), which induces a severe performance decline on every question type. However, this baseline model still outperforms existing SoTA (*i.e.*, VGT) due to a more advanced image encoder and language model. Then, we conduct experiments to study STR. As a detailed breakdown test, we notice that reasoning with all frames without temporal rationalization (w/o "TR") will cause a performance drop. A similar declination is also observed when spatial rationalization is erased (w/o "SR"), that is, all objects on the selected frame are used for reasoning. Such performance drops are expected, because a large proportion of frames only contain a question-irrelevant scene, and the pretrained object detector will inevitably introduce noisy objects. These question-irrelevant contents, if not properly ruled out, will make the background overwhelm the causal information, due to its spurious correlation with the answer

Table 2: Accuracy (%) comparison on NExT-QA, MSVD-QA, and MSRVTT-QA. Acc@C, T, D, denote questions type of Causal, Temporal, and Descriptive in NExT-QA, respectively. The **best** and 2nd best results are highlighted.

Methods	NExT-QA				MSVD	MSRVTT
	Acc@All	Acc@C	Acc@T	Acc@D		
Co-Mem† [10]	48.5	45.9	50.0	54.4	34.6	35.3
HCRN [18]	48.9	47.1	49.3	54.0	36.1	35.6
HGA [16]	50.0	48.1	49.1	57.8	34.7	35.5
MSPAN [11]	50.9	48.6	49.8	60.4	40.3	38.0
IGV [22]	51.3	48.6	51.7	59.6	40.8	38.3
HQGA [35]	51.8	49.0	52.3	59.4	41.2	38.6
EIGV [21]	52.9	51.2	51.5	61.0	42.6	39.3
VGT [36]	53.7	51.6	51.9	63.7	-	39.7
VGT-PT [36]	55.7	52.8	54.5	67.3	-	-
TranSTR	**61.5**	**59.7**	**60.2**	**70.0**	**47.1**	**43.1**
vs. SoTA	+5.8	+6.9	+5.7	+2.7	+3.5	+3.4

Table 3: Accuracy (%) comparison on Causal-VidQA. D: Description, E: Explanation, P: Prediction, C: Counterfactual. *: Reproduced result using official implementation.

Methods	Acc@D	Acc@E	Acc@P			Acc@C			Acc@All
			A	R	AR	A	R	AR	
HCRN[18]	56.4	61.6	51.7	51.3	32.6	51.6	53.4	32.7	48.1
HGA[16]	65.7	63.5	49.4	50.6	32.2	52.4	55.9	34.3	48.9
B2A[28]	66.2	62.9	49.0	50.2	31.2	53.3	56.3	35.2	49.1
VGT*[36]	70.8	70.3	55.2	56.9	38.4	61.0	59.3	42.0	55.4
TranSTR	**73.6**	**75.8**	**65.1**	**65.0**	**48.9**	**68.6**	**65.3**	**50.3**	**62.2**
vs. SoTA	+1.8	+5.5	+9.9	+8.1	+10.5	+7.6	+6.0	+8.3	+6.8

Table 4: Ablation Study

Variants	NExT-QA			
	Acc@All	Acc@C	Acc@T	Acc@D
TranSTR	**61.5**	**59.7**	**60.2**	**70.0**
w/o STR & decoder	59.6	58.2	58.0	67.3
w/o STR	60.3	59.1	58.3	67.9
w/o TR	60.8	59.4	58.6	69.5
w/o SR	60.7	59.6	58.2	69.0
w/o decoder	60.1	58.2	58.9	68.5
w/o MGR	60.1	58.9	57.6	68.6
Random K	54.6	53.6	51.6	64.0
SinkHorn Top-K	61.0	59.4	59.7	68.7

Table 5: Performance comparison, grouped by video length and object number. diff = Acc($>$ 80s) − Acc($\leq$80s)

Model	Video Length			Object Number			Total
	$\leq$ 80s	$>$ 80s	diff(↑)	$\leq$ 5	$>$ 5	diff(↑)	
EIGV [21]	53.3	51.3	-2	53.3	52.1	-1.2	52.9
VGT [36]	54.4	52.2	-2.2	54.7	52.8	-1.9	53.7
VGT-PT [36]	55.8	54.5	-1.3	56.4	54.9	-1.5	55.7
w/o STR & decoder	59.8	58.7	-1.1	60.0	59.2	-0.8	59.6
w/o STR	60.5	58.9	-1.6	60.1	60.6	+0.5	60.3
TranSTR	61.4	62.4	**+1**	61.2	61.9	**+0.7**	61.5

distractor. As a result, we witness a more significant performance drop when both temporal and spatial rationalization are removed (w/o "STR"). Next, we validate the effectiveness of our decoder design by adopting a conventional implementation that concatenates each answer candidate with the question before feeding it to the model. This variant, remarked as "w/o decoder", also caused a substantial performance drop, which highlights the importance of our video-text interaction pipeline in eliminating the background-distractor correlation. Comparing the performance of (w/o "STA") and (w/o "deocder") to ("w/o STR & decoder"), we show that removing both STR and decoder induce a more severe decline, which demonstrates that STR can coordinate well with the proposed decoder and their benefits are mutually reinforcing. We also evaluate the importance of our MGR module by replacing it with average pooling. This variant, denoted as "w/o MGR", uses an average pooling to gather all objects on each frame and adds the pooled representation to the corresponding critical frames. We observed a significant performance drop when compared to the original TranSTR, which confirms the necessity of MGR in aggregating the multi-grain evidence for reasoning. To validate that the STR indeed learns to focus on the critical elements instead of making random choices, we replace our differentiable top-K module, with a random K selection. As a result, the performance of "Random K" drops drastically, which verifies the proposed STR is fully trainable to capture the answer information. In addition, we also verify our choice of the differentiable module by replacing our perturbed maximum method with the "SinkHorn Top-K" [26], and the results validate our implementation.

5.2.2 Analysis on Complex VideoQA

In Tab. 5, we compare the results of TranSTR on the simple and complex VideoQA, where the test set of NExT-QA [34] is split by the length and object number of the source video, respectively. By calculating the accuracy within each group, we notice that all existing methods, as well as the TranSTR baseline (TranSTR without STR and proposed decoder), suffer from a performance drop when the video length exceeds 80 seconds (diff: -1.3%∼-2%) or the video contains more than 5 objects (diff: -0.5%∼-1.9%)). Such a phenomenon still exists even when the proposed decoder is adopted. In contrast, the STR module has alleviated this issue by explicitly ruling out redundant frames and noisy objects, thus resulting in even better performance on complex samples.

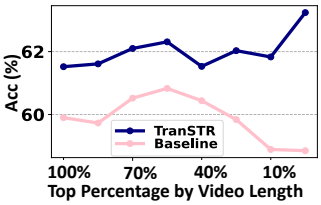


Figure 3: Acc by video length

Methods	Decoder	
	w/o	w
MSPAN [11]	50.9	52.7
EIGV [21]	51.3	53.3
TranSTR	60.1	61.5

Table 6: Apply our decoder to SoTAs.

Similar to Fig. 1b, we also investigate how the TranSTR performs on samples with different video lengths. In Fig. 3, we sort all test samples based on their video length and use a subset with the top percentage of longest videos to calculate accuracy. For example, 10% on the x-axis denotes the accuracy of samples with the top 10% longest videos, and 100% denotes all samples considered. Although the performance of TranSTR and baseline are initially comparable, the advantage of TranSTR becomes more pronounced as videos become longer. As a result, when the subset narrows down to the 10% longest videos, we observe a difference in the accuracy of over 4%.

5.2.3 Study of Decoder

Our video-text interaction pipeline is able to cater to any SoTA methods without compromising their structure. Thus, we apply the proposed decoder to three VideoQA backbones by separating the answer candidates from the question and feeding them to our answer decoder. As shown in Tab. 6, our decoder is able to consistently improve the performance of all backbones, validating the assumption that isolating the answer candidates from the video-text encoding can eliminate the spurious correlation between a negative answer and background scenes, resulting in favorable gains for the backbone models.

Moreover, our decoder design is also more efficient compared to the conventional implementation. In the traditional approach, a question needs to be fused with the video multiple times, with each time concatenating a different answer candidate. In contrast, our design only forwards the question once, as the answer candidates are introduced only after the video-question fusion, resulting in a much more efficient architecture.

5.2.4 Study of Hyper-parameter

To validate the sensitivity of TranSTR to the number of collected interactions, we conduct experiments with variations

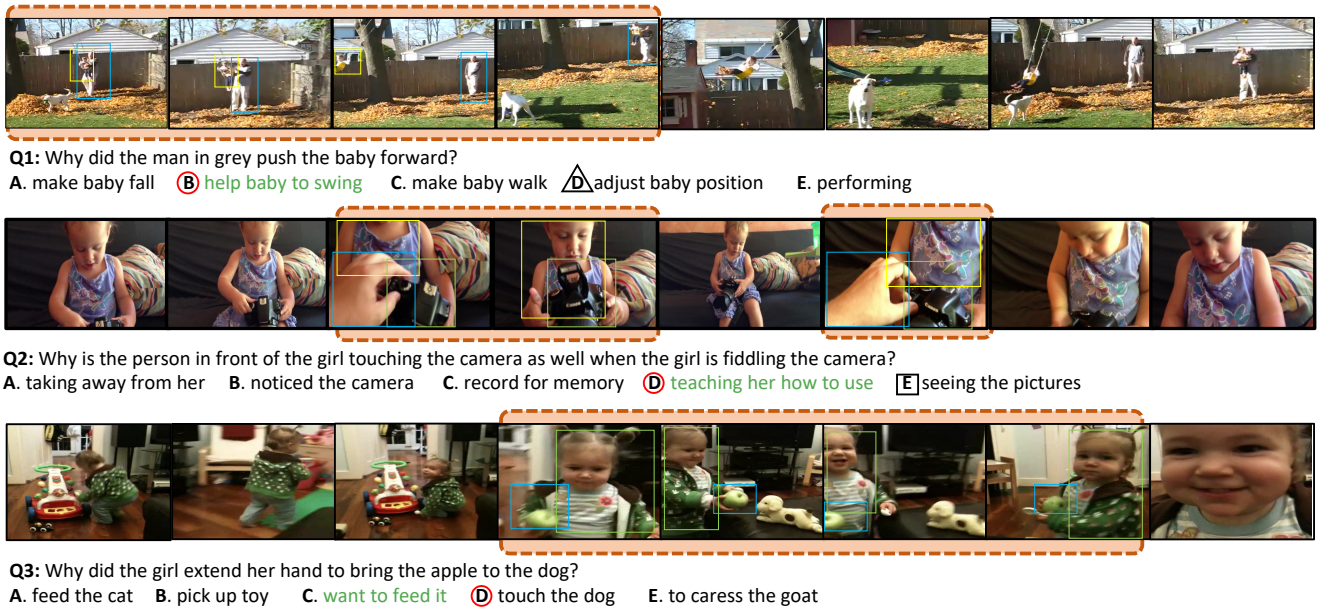


Figure 4: Case-Study on NExT-QA test set, the critical frames and objects are highlighted. Q1 and Q2 present the effect of the proposed STR and answer decoder, respectively. Q3 shows a failure case. The ground truth is colored in green. (◯:prediction of TranSTR, △: prediction of TranSTR w/o STR, □: prediction of TranSTR w/o decoder)

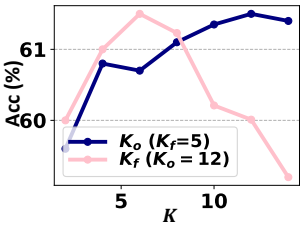


Figure 5: Study of hyper-parameters.

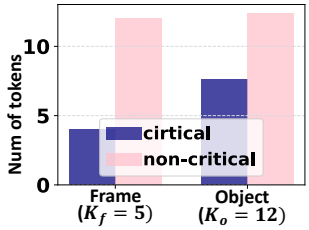


Figure 6: Study of critical frames and objects.

of K_f and K_o on NExT-QA. Without loss of generality, we tune K_f (K_o) while setting K_o = 12 (K_f = 5). According to Fig. 5, we observe that the performance of TranSTR varied mostly in the range of 61% to 61.5% under different combinations of hyperparameters, which demonstrated the effectiveness of TranSTR's adaptive design. However, we also notice a significant drop in some corner cases. When K_f (K_o) is too small, the number of critical frames (objects) is limited, which hinders the performance as some visual evidence for answering is missing. Similarly, when K_f is larger than 10, it introduces too much background, thus hurting the performance.

5.3. Study of Critical Frames and Objects (Q3)

Quantitative study. To grasp the learning insight of TranSTR, we inspect the number of frames and objects that are collected as critical by the adaptive rationalization. Concretely, we draw the number of frames C and non-critical frames T-C in Fig. 5. For the visual objects, since the number of critical objects C_t varies according to frame content, we take the average of C_t overall all critical frames, while leaving the rest objects as non-critical. As a result, TranSTR can pinpoint a small group of tokens as critical tokens while leaving the rest as redundancy, which manifests the mass of question-irrelevant content in the original video, thus pointing the necessity of rationalization.

Qualitative study. To capture the learning pattern of TranSTR, we present some prediction results in Fig. 4 along with the identified frames and objects. In general, TranSTR can locate very few indicative elements as visual evidence. In question 1, we show the effectiveness of the STR. In temporal selection, it rules out the environment scene and targets the first four "swing" frames as critical. Next, in the spatial selection, it excludes non-causal objects (*i.e.* "dog") and focuses on the relation between question-relevant objects (*i.e.* "man" and "baby"). By aggregating the critical elements in frames and objects, TranSTR successfully reaches the gold answer. As a comparison, when the STR is removed, the massive background overwhelms the salient reasoning pattern and leads to a false prediction. Question 2 demonstrates the effect of our answer decoder. We can see that TranSTR targets three critical frames that encompass the question-referred "person", while selecting "camera" and "girl" as critical objects to correctly infer the person's intention. However, when the decoder is removed, the prediction falls into negative answer "E.seeing the pictures". This is because implementation without our answer decoder inevitably suffers from a spurious correlation between the negative answer "E.seeing the pictures" and frames where

the girl is actually seeing pictures, even though these frames are irrelevant to the question. Lastly, we present a failure case in question 3, where TranSTR fails to capture the subtle difference between "feed" and "touch", although the critical visual elements are located, which leads to a false prediction of the girl's intention.

6. Conclusion

For the first time, this paper addresses complex VideoQA, where long multi-object video and hard answer distractors have crippled existing methods, owing to their incapacity in handling massive visual backgrounds and modeling hard distractor answers. We then propose STR to adaptively trim off question-irrelevant scenes, and further develop a novel answer decoding scheme that coordinates with STR to overcome the spurious correlation resulted from distractor-background interaction. Instantiating this pipeline with transformer architecture, we show our method, TranSTR, achieves significant improvements over SoTAs, especially on complex VideoQA tasks. We hope our success can shed light on answering questions in the context of long videos with multiple objects.

References

[1] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *NeurIPS*, pages 9508–9519, 2020. 4

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 213–229, 2020. 5

[3] Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. Can rationalization improve robustness? In *NAACL*, pages 3792–3805, 2022. 3

[4] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In *IJCAI*, pages 636–642, 2021. 2, 3

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

[7] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. Beyond VQA: generating multi-word answers and rationales to visual questions. In *CVPR*, pages 1623–1632, 2021. 3

[8] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007, 2019. 2

[9] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1

[10] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585, 2018. 2, 6

[11] Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. Multi-scale progressive attention network for video question answering. In *ACL*, pages 973–978, 2021. 2, 6, 7

[12] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *ICLR*, 2021. 1, 5

[13] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 1, 2

[14] Wei Ji, Xi Li, Lina Wei, Fei Wu, and Yueting Zhuang. Context-aware graph label propagation network for saliency detection. *IEEE Transactions on Image Processing*, 29:8177–8186, 2020. 3

[15] Wei Ji, Yicong Li, Meng Wei, Xindi Shang, Junbin Xiao, Tongwei Ren, and Tat-Seng Chua. Vidvrd 2021: The third grand challenge on video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4779–4783, 2021. 3

[16] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116, 2020. 2, 6

[17] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, volume 48, pages 1378–1387, 2016. 1, 2

[18] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, pages 9969–9978, 2020. 3, 6

[19] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 2

[20] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR*, pages 21241–21250, 2022. 1, 2, 3, 5

[21] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. Equivariant and invariant grounding for video question answering. *CoRR*, abs/2207.12783, 2022. 3, 5, 6, 7

[22] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2928–2937, 2022. 3, 5, 6

[23] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Transformer-empowered invariant grounding for video question answering. *TPAMI*, pages 1–12, 2023. 3

[24] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. Interventional video relation detection. In *ACM MM*, pages 4091–4099, 2021. 3

[25] Yicong Li, Xun Yang, An Zhang, Chun Feng, Xiang Wang, and Tat-Seng Chua. Redundancy-aware transformer for video question answering. *arXiv preprint arXiv:2308.03267*, 2023. 3

[26] Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *ICLR*, 2018. 7

[27] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, pages 8779–8788, 2018. 3

[28] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering, 2021. 2, 6

[29] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *ACM MM*, 2021. 2

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 5

[31] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *KDD*, pages 1135–1144, 2016. 3

[32] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM MM*, pages 1300–1308, 2017. 2

[33] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022. 3

[34] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 1, 2, 5, 7

[35] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*, pages 2804–2812, 2022. 2, 3, 6

[36] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, pages 39–58. Springer, 2022. 2, 3, 5, 6, 7

[37] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, pages 1645–1653, 2017. 1, 5

[38] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pages 6787–6800, 2021. 3

[39] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 1, 3

[40] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 2021. 3

[41] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1939–1947, 2020. 3

[42] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 1

[43] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *CVPR*, pages 6261–6270, 2019. 3

[44] Yaoyao Zhong. Video question answering: Datasets, algorithms and challenges. *EMNLP*, 2022. 2

[45] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, 2022. 1