SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2024, Vol. 67, Iss. 10, 202104:1–202104:16
https://doi.org/10.1007/s11432-023-4084-6

Robust video question answering via contrastive cross-modality representation learning

Xun YANG[1*], Jianming ZENG[1,3], Dan GUO[2], Shanshan WANG[4], Jianfeng DONG[5] & Meng WANG[2,3]

[1]*School of Information Science and Technology, University of Science and Technology of China, Hefei* 230026*, China;*
[2]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei* 230601*, China;*
[3]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei* 230088*, China;*
[4]*Institutes of Physical Science and Information Technology, Anhui University, Hefei* 230601*, China;*
[5]*School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou* 310018*, China*

Received 29 July 2023/Revised 14 November 2023/Accepted 30 January 2024/Published online 24 September 2024

Abstract Video question answering (VideoQA) is a challenging yet important task that requires a joint understanding of low-level video content and high-level textual semantics. Despite the promising progress of existing efforts, recent studies revealed that current VideoQA models mostly tend to over-rely on the superficial correlations rooted in the dataset bias while overlooking the key video content, thus leading to unreliable results. Effectively understanding and modeling the temporal and semantic characteristics of a given video for robust VideoQA is crucial but, to our knowledge, has not been well investigated. To fill the research gap, we propose a robust VideoQA framework that can effectively model the cross-modality fusion and enforce the model to focus on the temporal and global content of videos when making a QA decision instead of exploiting the shortcuts in datasets. Specifically, we design a self-supervised contrastive learning objective to contrast the positive and negative pairs of multimodal input, where the fused representation of the original multimodal input is enforced to be closer to that of the intervened input based on video perturbation. We expect the fused representation to focus more on the global context of videos rather than some static keyframes. Moreover, we introduce an effective temporal order regularization to enforce the inherent sequential structure of videos for video representation. We also design a Kullback-Leibler divergence-based perturbation invariance regularization of the predicted answer distribution to improve the robustness of the model against temporal content perturbation of videos. Our method is model-agnostic and can be easily compatible with various VideoQA backbones. Extensive experimental results and analyses on several public datasets show the advantage of our method over the state-of-the-art methods in terms of both accuracy and robustness.

Keywords video question answering, cross-modality fusion, contrastive learning, cross-media reasoning

1 Introduction

Video question answering (VideoQA) is a fundamental yet important multimedia understanding task [1] that requires a joint understanding of low-level video content and high-level textual semantics. As shown in Figure 1(a), given a natural language question and a video, the VideoQA model aims to derive the correct answer from a set of candidate answers by cross-model reasoning [2]. VideoQA has attracted increasing attention [3,4] in recent years from both multimedia and natural language processing communities because of its multimodal nature and great potential in real-world downstream applications, e.g., in-home robots and personal assistants.

To date, a large number of VideoQA models have been designed [4–7] for this challenging task. Most existing efforts mainly focus on designing complex cross-modality interaction networks to gain a better understanding of videos under the guidance of questions. Despite the promising progress of existing efforts, recent studies [8,9] indicated that visual question answering (VQA) systems mostly tend to over-rely on the superficial correlations between the questions and the answers while ignoring the visual content

* Corresponding author (email: xyang21@ustc.edu.cn)

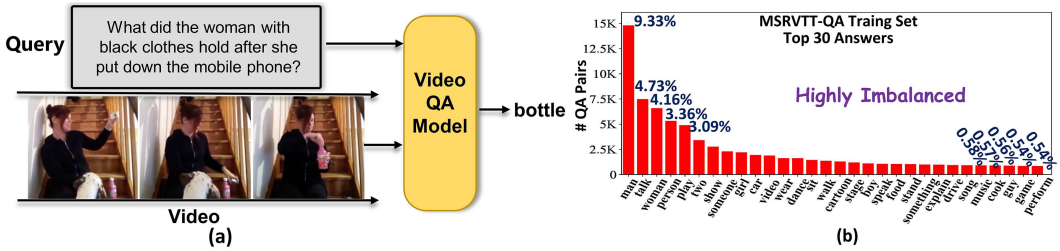

Figure 1 (Color online) Example of VideoQA (a) and illustration of the long-tailed distribution of highly imbalanced answer annotations in the MSRVTT-QA [10] dataset (b).

(aka, language biases), thus leading to unreliable results. Such a serious issue is also found in VideoQA by the latest study [3]. As shown in Figure 1(b), the number of QA pairs in each answer category is highly imbalanced. As a result, the models prefer to concentrate on the co-occurrence between the question type and the answer during training and tend to yield frequent answers for the specific type of question during inference. Notably, such spurious correlation works well on the head answer categories but fails on the tail answer categories, revealing a fundamental concern that existing models do not truly understand the temporal structure of videos and the true effect of the video content may be largely overlooked in the final prediction of VideoQA models. Such biased models have poor generalizability in unseen QA scenarios and are ineffective for real-world applications. How to effectively understand and model the temporal and semantic characteristics of videos for cross-modality reasoning is crucial but, to our knowledge, has not been well investigated for VideoQA.

In this work, to fill the research gap, we develop a robust VideoQA framework that forces the model to effectively model and leverage the temporal video semantics for cross-modality reasoning instead of exploiting the superficial statistical correlation, thereby improving the robustness of VideoQA. We expect our model to focus on the video content when making a QA decision. Therefore, we mainly concentrate on the video part for VideoQA rather than on the textural query because many efforts have already been devoted to debiasing VQA from the language part.

Specifically, we first design a self-supervised contrastive learning objective (CL) to contrast the positive and negative pairs of multimodal input (query, video), where the fused representation of original input is enforced to be closer to that of the intervened input based on clip-level video perturbation in a latent embedding space. In this way, the model is forced to focus more on the global context of videos rather than some static keyframes in cross-modality fusion, which can prevent the model from overlooking the video content in cross-modality reasoning. An effective temporal order regularization (TO) is also introduced in our framework that enforces the inherent sequential structure of videos for video representation. In particular, we first shuffle the sampled nonoverlapping video clips to a random order and then ask the model to predict the correct order. Moreover, we design a Kullback-Leibler divergence-based perturbation invariance regularization (Inv) of the predicted answer distribution to improve the robustness of the model against temporal content perturbation of videos. We force the model to capture the stable and perturbation-invariant video features for answering reasoning.

Our finally derived video representation can well preserve the valuable temporal context of video activities, be robust to local temporal content perturbation, and help the model learn an unbiased cross-modality representation under the proposed contrastive learning framework. Notably, our proposed approach is model-agnostic and can be compatible with various VideoQA backbones. We validate the effectiveness of our approach on several benchmarks, including MSVD-QA [10], MSRVTT-QA [10], and SUTD-TrafficQA [11], and demonstrate the advantages of our method over state-of-the-art (SOTA) VideoQA methods. Ablation studies also reveal that our method performs better over the less-frequent answer categories and the testing video set with longer temporal duration, supporting the improved robustness of the proposed strategy.

Overall, our technical contributions are summarized as follows:

- We develop a model-agnostic VideoQA method that aims to force the model to concentrate on the informative video content for cross-modality answer reasoning instead of over-replying on the spurious correlations rooted in dataset bias.
- We devise three simple yet effective learning objectives that are easily compatible with existing VideoQA backbones and significantly improve the robustness of VideoQA models.
- We conduct sufficient experiments on four benchmarks using several baselines that reflect the effec-

tiveness of our method.

2 Related work

Image-based QA. VQA [8, 12, 13] aims to generate natural language answers to a question posed for a specified image. Early research mainly focused on visual understanding or spatial [14] and semantic [15] relational inference in static images, and even some large-scale pretraining efforts [16] have shown a near-human capability in VQA accuracy. However, many researchers pointed out that the superior performance of the model may not be attributed to its robust reasoning capability but rather the result of language bias between the question and the answer. In light of this, several specific VQA datasets and evaluation protocols [17] have been proposed to address the language bias. The debias methods can be roughly divided into two categories: (1) Ensemble-based methods [18, 19] strive to enhance the visual attention to image content [20, 21] or use an independent question branch to explicitly model the language prior in the training data [18, 22] to ease the influence of language prior. (2) Data-balancing methods [23, 24] often adopt a data augmentation strategy to generate counterfactual samples [24] or randomly replace images and questions to conduct unbiased training by solving the unbalanced distribution problem of training data. These debia methods can alleviate the language bias and improve the robustness of VQA models.

Video-based QA. The VideoQA task involves not only the understanding of visual content but also the inference of their semantic, spatial, and temporal relationships. Previous methods [4,25–29] mainly focused on solving the correlation learning between the video and the question, including (1) attention-based models [25, 26], (2) memory-based models [27, 30], (3) graph-based models [28, 31], and (4) transformer-based models [29]. Most previous VideoQA models have tended to rely on superficial relevance and ignore the video content. Recently, the issue of language bias in VideoQA has been revealed. Specifically, Li et al. [32] proposed a new framework, i.e., IGV, which introduces invariant grounding learning to restrain spurious correlations between the question and the irrelevant scenes. In practice, IGV performs causal intervention on the video to make the model insensitive to changes in the video scene unrelated to the question, thus improving the robustness of the model. However, the superficial correlation problem in VideoQA has not been well investigated. In this study, we proposed a robust VideoQA framework that builds a self-supervised contrastive learning objective to improve the robustness of cross-modality fusion representation against the local context-based video perturbation. We also devised two simple and effective regularizers, which can effectively improve the robustness of answer prediction.

Contrastive learning. Contrastive learning [33] is a new technique to automatically learn generalized data representations by distinguishing similar and dissimilar data in an embedded space, which has been well developed in a variety of vision and language tasks, such as visual recognition [34, 35], video-text retrieval [36], and VQA [37, 38]. In VQA, to improve the reasoning power and robustness of the model, Liang et al. [37] introduced a contrastive learning mechanism to learn the relationship between original, factual, and counterfactual samples for learning a cross-modality joint embedding. Si et al. [38] proposed a new contrastive learning method, which eliminates the information related to superficial correlations in the original training samples to construct positive samples. This method uses biased samples to obtain unbiased samples, which does not diminish the importance of biased samples in training. In this work, we explore the potential of contrastive learning in training robust VideoQA models.

3 Methodology

3.1 Preliminaries

Before diving into our proposed approach, we first briefly revisit the common paradigm of VideoQA models. Given a pair of queries and its corresponding video (Q, V) sampled from a VideoQA dataset $\mathcal{D} = \{\mathcal{Q}, \mathcal{V}, \mathcal{A}\}$ consisting of N QA pairs, the goal of VideoQA is to optimize a cross-modality mapping $f : \mathcal{Q} \times \mathcal{V} \to [0, 1]^{|\mathcal{A}|}$, which can effectively identify the correct answer $\hat{A}$ from the set of candidates and is briefly formulated as follows:

$$\hat{A} = \arg\max_{A \in \mathcal{A}} f_{\mathcal{A}}(Q, V), \tag{1}$$

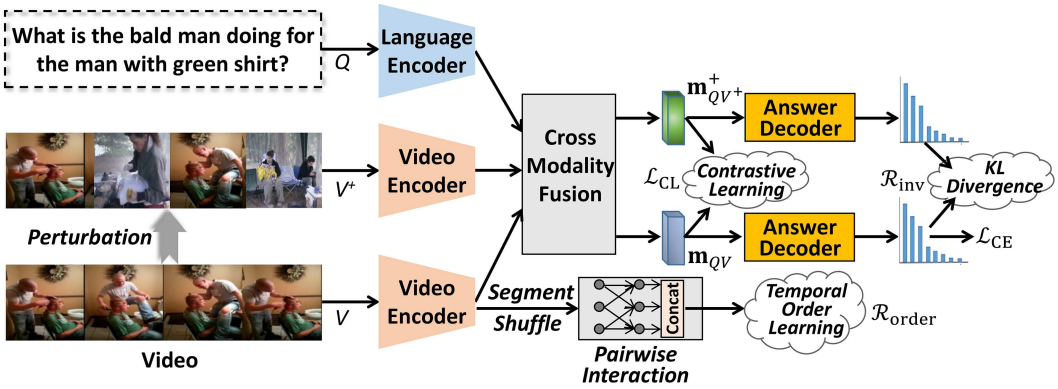

Figure 2 (Color online) Pipeline of our proposed model-agnostic robust VideoQA method that consists of a contrastive learning objective, a temporal order regularization, and a perturbation invariance regularization.

where $\mathcal{A}$ denotes the predefined answer set either in multiple-choice or open-ended QA. The objective in (1) is usually optimized by minimizing the standard cross-entropy (CE) loss in the empirical risk minimization (ERM) setting as follows:

$$\mathcal{L}_{\mathrm{CE}} = \mathbb{E}_{(Q,V,A)\in\mathcal{D}}\,\mathrm{CE}\big(\hat{A}, A\big). \tag{2}$$

Generally speaking, a VideoQA system is composed of the following:

- Video encoder, which usually encodes the original video as a sequence of visual feature vectors, represented by frame appearance features, motion features, object-level appearance features, or a combination of all of the three types of features [4] based on pretrained CNNs, such as ResNet-101 [39] and 3D ResNeXt-101 [40], and pretrained object detectors [41];
- Language encoder, which encodes the user intention in the textual query into linguistic representation in the form of global sentence representation, contextual word representation, or structural representation of grammatical dependencies [42] based on RNN or pretrained BERT models [43];
- Cross-modality fusion module, which is mainly designed for vision-language alignment and fusion based on cross-modality interaction networks, such as graph neural networks [44], hierarchical architectures [4, 45], and transformers [4], and finally yields a fused representation for answer prediction;
- Answer decoder, which maps the cross-modality fusion representation into the latent answer space to estimate the matching score between the given (Q, V) pair and the candidate answer $A \in \mathcal{A}$.

Existing efforts mainly focus on designing cross-modality fusion modules, which is a popular strategy in the vision-language field. Despite their performance improvement, existing methods may be prone to overexploiting the superficial statistical correlations between the questions and the answers or those between the noncausal part of the video and the answers [32] because of the issue of ERM criterion [46,47]; i.e., VideoQA models may easily predict the correct answer even though they have not well understood the video content. Such type of methods may be sensitive to the distribution shift of the testing set, which has poor generalizability in unseen domains, thereby motivating us to design a robust VideoQA model that can effectively leverage the informative video content for making a reliable QA decision.

3.2 Our proposed method

To address the aforementioned issue, this section presents a robust VideoQA approach that can be effectively compatible with existing methods. Our basic idea is forcing the model to stably leverage the temporal context and semantics of videos instead of overlooking the informative video content for biased prediction. Inspired by the recent progress in robust VQA [47], as shown in Figure 2, we devise a self-supervised contrastive cross-modality representation learning solution for robust VideoQA, which mainly consists of a video-perturbation-based contrastive cross-modality learning objective in Subsection 3.2.1 and two video-related regularization terms in Subsections 3.2.2 and 3.2.3, and will be deliberated as follows.

3.2.1 *Contrastive cross-modality learning*

Learning cross-modality fusion is a crucial step of VideoQA. How to ensure that the fused representation well preserves the valuable temporal characteristic of videos is the key research question of robust VideoQA. In this work, to prevent the model from leveraging the spurious correlation between the pairwise input (Q, V) and the answer A, we introduce the self-supervised contrastive learning as an auxiliary objective to boost the learning of query-video fusion representation. First, the original input (Q, V) is formulated as an anchor sample, represented as $g(Q, V)$, where $g : \mathcal{Q} \times \mathcal{V} \to \mathbb{R}^d$ denotes the cross-modality fusion module. Then, we devise a video perturbation strategy to formulate the positive sample as $g(Q, V^+)$, where V^+ denotes the perturbed video of V. Moreover, we sample K negative samples from another query-video pair (Q', V'), where (Q', V') can be either the matched or unmatched pair in the training set but should have a different answer from (Q, V). Here, we recall that the key idea of contrastive learning is to contrast positive and negative pairs of data points, aiming to push the representations of positive pairs to be closer in a latent embedding space and those of negative pairs to be more orthogonal [33]. Hence, our auxiliary contrastive learning objective is formulated as follows:

$$\mathcal{L}_{\mathrm{CL}} = \mathbb{E}_{\boldsymbol{h}_a, \boldsymbol{h}_p, \{\boldsymbol{h}_{n_k}\}_{k=1}^K} \left[-\log \frac{\mathrm{e}^{\mathrm{sim}(\boldsymbol{h}_a, \boldsymbol{h}_p)}}{\mathrm{e}^{\mathrm{sim}(\boldsymbol{h}_a, \boldsymbol{h}_p)} + \sum_{k=1}^K \mathrm{e}^{\mathrm{sim}(\boldsymbol{h}_a, \boldsymbol{h}_{n_k})}} \right], \tag{3}$$

where $\boldsymbol{h}_a = g(Q, V), \boldsymbol{h}_p = g(Q, V^+)$, and $\boldsymbol{h}_{n_k} = g(Q', V')$ denote the embeddings of the anchor, positive, and negative samples, respectively. The term $\mathrm{sim}(\cdot, \cdot)$ denotes the similarity metric that measures the similarity between two input samples. The goal of (3) is to encourage the model to be robust to the perturbation of videos and the fused representation $g(Q, V)$ to capture the key temporal cues of videos[1)].

The key to achieving our goal is how to design an effective video perturbation strategy for positive sample construction. We introduce a simple strategy as follows:

• **Sample and replace.** First, we preprocess the given video as a sequence of sampled clips $V = \{v1, \ldots, v_M\}$ and divide it into s segments $\{S_1, \ldots, S_s\}$. Then, we randomly sample one clip from each segment and replace the clip with the corresponding one in a randomly selected video. Such a strategy can ensure that the global context of the videos is intact.

The aforementioned strategy is simple and reasonable, which is different from the popular mixup augmentation [48]. mixup averages not only the RGB values of two images but also the ground truth labels to create new samples. In our work, the ground truth answer of the augmentation-based positive sample should be the same as that of the anchor. Furthermore, Eq. (3) is an auxiliary learning objective that does not include the ground truth answer and is used to boost the optimization of the CE loss in (1). We expect that our model will focus more on the informative multimodal fusion information instead of the hidden shortcuts in the dataset.

3.2.2 *Temporal order regularization*

The video-perturbation-based contrastive learning expressed in (3) helps the model capture the global context of the video, thus improving the quality of cross-modality fusion; however, it does not consider the inherent sequential structure of videos. This section describes our effort to capture the temporal semantics of videos for robust VideoQA. We use the temporal coherence characteristic as a self-supervision signal to boost the learning of robust video representation. We believe that a good video representation should well preserve the temporal order [49], which helps the VideoQA model determine the action-related answers accurately.

Given a video that is divided into s segments $V = \{S_1, \ldots, S_s\}$, we randomly shuffle the segments as $V^{\mathrm{shuffle}} = \{S_i^1, \ldots, S_j^s\}$, which will lead to A_s^s (e.g., $A_s^s = 6$ when $s = 3$) combinations of temporal orders of video segments. We aim to predict the actual order of shuffled segments in V^{shuffle} by training a A_s^s-class classifier $f_{\mathrm{order}}(\cdot)$. Specifically, we first averagely pool the embeddings of clips as the segment embedding $\boldsymbol{h}_S$. Then, the segment embeddings in V^{shuffle} are pairwise concatenated [50], resulting in C_s^2 concatenated features (e.g., $\{\langle \boldsymbol{h}_{S_3}, \boldsymbol{h}_{S_1} \rangle, \langle \boldsymbol{h}_{S_3}, \boldsymbol{h}_{S_2} \rangle, \langle \boldsymbol{h}_{S_1}, \boldsymbol{h}_{S_2} \rangle\}$ when $s = 3$, where $\langle \cdot, \cdot \rangle$ denotes the concatenation operation). Furthermore, they will be transformed to form a tuple of C_s^2 vectors and concatenated again as the input of $f_{\mathrm{order}}(V^{\mathrm{shuffle}})$, which outputs a probability distribution over different

1) The query perturbation [38] can also be adopted for constructing high-quality positive samples; however, it is not the main focus of this VideoQA work.

orders. The temporal order regularizer is formulated as the minimization of the following CE loss:

$$\mathcal{R}_{\mathrm{TO}} = \mathbb{E}_{V\in\mathcal{V}}\,\mathrm{CE}\big(f_{\mathrm{order}}(V^{\mathrm{shuffle}}), y_{\mathrm{order}}\big), \tag{4}$$

where y_{order} denotes the one-shot encoding of the actual temporal order. Notably, both the original and perturbed videos can be used as the input of (4). Temporal order prediction is a proxy task to ensure that the entire model captures the temporal semantics of the video. Other efforts in temporal order regularizer [50] may also be compatible with our objective expressed in (4).

3.2.3 *Perturbation invariance regularization (Inv)*

Eq. (3) mainly focuses on improving the robustness of cross-modality fusion to the slight perturbation of the temporal content of videos. The combination of the contrastive learning objective in (3) and the temporal order regularizer in (4) can prevent the VideoQA model from overlooking the meaningful video content features, especially the action-aware content. However, both of them did not consider the robustness of answer prediction directly. Inspired by the recent exploration of causal invariance [46] in VideoQA [32], we enforce the predicted answer distribution to be robust to the video perturbation described in Subsection 3.2.1, which aims to further boost the robustness of the proposed VideoQA model. Specifically, to achieve our goal, we maximize the consistency between the original prediction $f_{\mathcal{A}}(Q,V)$ and the intervened prediction $f_{\mathcal{A}}(Q,V^{+})$ by minimizing the following regularization term:

$$\mathcal{R}_{\mathrm{Inv}} = \mathbb{E}_{V\in\mathcal{V}}\,\mathrm{KL}\big(f_{\mathcal{A}}(Q,V), f_{\mathcal{A}}(Q,V^{+})\big), \tag{5}$$

where $\mathrm{KL}(\cdot,\cdot)$ denotes the Kullback-Leibler divergence that measures the distance between two probability distributions.

Causality and robustness. Compared with existing causality-based VQA studies [19,32], we do not formulate a causal graph that plots the hidden confounder and the causal part; thus, we cannot accurately infer the underlying true causal effect of the input (Q,V) on the output A. Our solution can be considered a type of causal intervention: we first intervene V by utilizing the perturbation strategy to create a new data distribution, which has never been observed in the original dataset; then, we enforce the prediction $P(A|(Q,V))$ to be invariant across different distributions. In this way, the model can be forced to focus more on the causal and stable features of videos, and the risk of overreliance on the spurious correlations in the dataset is considerably lessened, thereby improving the robustness of the VideoQA model.

3.2.4 *Learning*

To this end, we introduce the three key components of our work. Notably, both the CL expressed in (3) and the two regularizers, i.e., $\mathcal{R}_{\mathrm{TO}}$ and $\mathcal{R}_{\mathrm{Inv}}$, are self-supervised, which can alleviate the issue of dataset bias in our framework. We hope the three components can mutually support each other during training. The overall learning objective of our robust VideoQA method is formulated as follows:

$$\mathcal{L}_{\mathrm{final}} = \mathcal{L}_{\mathrm{CE}} + \lambda_1\mathcal{L}_{\mathrm{CL}} + \lambda_2\mathcal{R}_{\mathrm{TO}} + \lambda_3\mathcal{R}_{\mathrm{Inv}}, \tag{6}$$

where λ_1, λ_2, and λ_3 are three hyperparameters used to balance the strengths of different components during training. In the inference stage, only the main branch $f_{\mathcal{A}}(Q,V)$ is used to answer the question. Our proposed method can be easily integrated with other VideoQA backbones to improve the robustness of training.

4 Experiment

In this section, we quantitatively and qualitatively conduct extensive experiments to answer the following research questions: R1: Can our method effectively improve the performance of existing VideoQA methods? R2: Do our three key components complement each other? R3: How robust is our method against the dataset bias from the long-tailed distribution? How is the generalizability of our method when tested in unseen domains? R4: How does our method perform in different settings of hyperparameters or over different test groups with different video durations or query types?

Table 1 Statistics of four VideoQA benchmark datasets[a)]

Dataset	Topic	#Videos	#QA pairs	$\overline{L}_{\text{vid}}$ (s)	QA-T	AN-T
MSVD-QA	Various scene	1970	50505	10	OE	Auto
MSRVTT-QA	Various scene	10000	243690	15	OE	Auto
Traffic-QA	Traffic events	10080	62535	6.9	MC	Manual
NExT-QA	Various scene	5440	52044	44	MC&OE	Manual

a) QA-T denotes the types of VideoQA (i.e., open-ended [OE] or multiple-choice [MC]); AN-T denotes the types of video annotation (i.e., Auto or Manual); $\overline{L}_{\text{vid}}$ denotes the average temporal duration of all videos.

4.1 Datasets and settings

Datasets. We conduct experiments on four public VideoQA datasets, namely, MSVD-QA [10], MSRVTT-QA [10], Traffic-QA [11], and NExT-QA [3]. (1) MSVD-QA is recreated from the Microsoft Research Video Description Corpus dataset, which is a widely used video captioning benchmark and consists of 1970 short videos and 50505 QA pairs. We follow the settings of Xu et al. [10], i.e., 61% for training, 13% for validation, and 26% for testing. (2) MSRVTT-QA is recreated from the MSRVTT dataset, which is larger and more complex than MSVD-QA and consists of 10000 short videos and 243690 QA pairs. We follow the settings of Xu et al. [10], i.e., 65% for training, 5% for validation, and 30% for testing. (3) Traffic-QA [11] is collected via the combination of online harvesting and offline capturing, covers various real-world traffic scenarios, and consists of approximately 10000 in-the-wild videos and 62535 QA pairs for benchmarking the cognitive capability of causal inference and event understanding models in complex traffic scenarios. We follow the official data splits [11] of Traffic-QA for the experiments. (4) NExT-QA [3] is a manually annotated VideoQA dataset that features causal and temporal object interactions in space-time and consists of approximately 47700 manually annotated questions for multiple-choice QA collected from 5400 videos with an average length of 44 s. Table 1 shows the statistics of the four public VideoQA benchmarks.

Metrics. Current methods mainly adopt the micro-averaged accuracy as the default metric for evaluation. This work also adopts the macro-averaged accuracy as the auxiliary metric to deal with the imbalanced distribution of candidate classes for open-ended VideoQA. Specifically, macro-averaging first computes the metric for each answer class independently and then takes the average, which treats all classes equally. The difference between macro-averaging and micro-averaging is that macro-averaging weighs each class equally, whereas micro-averaging weighs each sample equally. For the multiple-choice Traffic-QA [11] and NExT-QA [3] datasets, we only report the micro-averaged accuracy because most of their annotated answers are a sequence of natural language, not a simple word.

Baselines. In this work, we leverage and re-implement four different VideoQA methods as the baselines to validate the effectiveness of our robust VideoQA methodology. These methods are described as follows:

- HGA [51] contributes a heterogeneous graph alignment network over the video shots and question words, where the cross-modality information can be aligned and interacted effectively;
- IGV-B [32] is an extended version of HGA [51] and is implemented as the VideoQA backbone of IGV [32], which usually exhibits a more stable performance than HGA;
- CoMem [52] effectively fuses cues from both motion and appearance based on the dynamic memory network [53];
- MSPAN [54] is a multiscale progressive attention network to achieve relational reasoning, where multiscale clip features and linguistic semantics are effectively fused.

Notably, all of the four baselines are mainly used to implement the cross-modality fusion module $g(Q, V)$ of our robust VideoQA framework under the learning objective expressed in (6). The four baselines share the same input of video features (i.e., appearance and motion) and query features and the same answer decoder for a fair comparison.

Implementation details. (1) For the experiments on MSVD-QA [10] and MSRVTT-QA [10], we use the following feature extraction strategies for video and language modeling. We extract the video motion and appearance features using the pretrained 3D ResNeXt-101 and ResNet-152, respectively. (2) For the experiments on Traffic-QA [11], we extract the frame-level video motion and appearance features using the pretrained 3D ResNeXt-101 and ResNet-101, respectively. Each video is uniformly sampled into eight clips. The word-level query features are first extracted from Glove and then fed into a BiLSTM network for query modeling, following the method used by Xu et al. [11]. (3) For the experiments on NExT-QA [3], we extract the frame-level video motion and appearance features using the pretrained 3D

Table 2 Performance (accuracy, %) comparison with SOTAs and baselines on two datasets, i.e., MSVD-QA and MSRVTT-QA[a)]

	Method	MSVD-QA [10]		MSRVTT-QA [10]		Ref.
		Micro accuracy	Macro accuracy	Micro accuracy	Macro accuracy	
Memory network-based	AMU [10]	32.0	–	32.5	–	MM 2017
	HME [27]	33.68	4.5	33.0	–	CVPR 2019
Hierarchy structure-based	PGAT [55]	39.0	–	38.1	–	MM 2021
	HCRN [45]	36.1	–	35.6	–	CVPR 2020
	HOSTR [56]	39.4	–	35.9	–	IJCAI 2021
	HQGA [4]	41.15	9.03	38.62	6.16	AAAI 2022
Graph neural network-based	L-GCN [57]	34.3	–	33.7	–	AAAI 2020
	DVGR [58]	39.0	–	35.5	–	TMM 2021
	GMIN [28]	35.4	–	36.1	–	TIP 2021
	B2A [42]	37.2	–	36.9	–	CVPR 2021
	IGV [32]	40.48	9.92	38.21	6.40	CVPR 2022
Baseline+Ours	HGA [51]	38.89	10.48	38.01	5.41	AAAI 2020
	+Ours	$41.01^{+\mathbf{5.5\%}}$	$11.76^{+\mathbf{12.2\%}}$	$38.98^{+\mathbf{2.6\%}}$	$6.77^{+\mathbf{25.1\%}}$	Ours
	IGV-B [32]	40.09	9.81	37.88	5.09	CVPR 2022
	+Ours	$42.28^{+\mathbf{5.5\%}}$	$10.51^{+\mathbf{7.1\%}}$	$39.18^{+\mathbf{3.4\%}}$	$6.61^{+\mathbf{29.9\%}}$	Ours
	CoMem [52]	39.71	8.99	37.89	4.95	CVPR 2018
	+Ours	$40.50^{+\mathbf{2.0\%}}$	$9.56^{+\mathbf{6.3\%}}$	$38.54^{+\mathbf{1.7\%}}$	$5.60^{+\mathbf{13.1\%}}$	Ours
	MSPAN [54]	41.68	9.95	38.99	5.27	ACL 2021
	+Ours	$42.73^{+\mathbf{2.5\%}}$	$10.80^{+\mathbf{8.5\%}}$	$39.38^{+\mathbf{1.0\%}}$	$6.61^{+\mathbf{25.4\%}}$	Ours

a) Our relative improvements over the baselines are in bold.

ResNeXt-101 and ResNet-101, respectively. Each video is uniformly sampled into 16 clips. The word-level query features are first extracted from the fine-tuned BERT model [3] and then fed into a BiLSTM network to capture the sequential dependencies for query modeling. We also include the experiments on NExT-QA with the Glove-based language features. (4) Answer decoder: In this work, we use a customized answer decoder to compute the probabilistic prediction score of each candidate answer. The basic idea is to compute the normalized relevance score between the multimodal fused representation $\boldsymbol{m}_{qv}$ of the pairwise input (Q, V) and the learnable semantics-enriched answer embedding $\boldsymbol{a}$, which is formulated as follows:

$$\boldsymbol{s}_a = \text{softmax}\left(\frac{\langle \boldsymbol{m}_{qv}, \boldsymbol{a}\rangle}{\|\boldsymbol{m}_{qv}\| \cdot \|\boldsymbol{a}\|}/\tau\right), \tag{7}$$

where $\boldsymbol{s}_a$ is the normalized relevance score of all candidate answers and $\boldsymbol{m}_{qv} \in \mathbb{R}^d$ is the output of the cross-modality fusion module $g_\theta(Q, V)$. For the experiments on the open-ended QA datasets, i.e., MSVD-QA or MSRVTT-QA, where each answer is only one word, the answer embedding $\boldsymbol{a}$ is extracted from the pretrained BERT model, followed by a linear layer that maps it to the common space shared with $\boldsymbol{m}_{qv}$. For the experiments on Traffic-QA, where each answer is a phrase or sentence, $\boldsymbol{a} = \text{BiLSTM}(\{\boldsymbol{w}_i^a\})$, which is the concatenation of the last hidden states in two directions of BiLSTM, and $\boldsymbol{w}_i^a$ denotes the pretrained Glove word embedding. τ is a positive temperature constant.

Settings. We train our proposed approach for 60 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 128 by default. We select the model checkpoint with the best performance on the validation set for evaluation on the testing set. The trade-off parameters λ_1, λ_2, and λ_3 in (6) are set by applying the grid search strategy on the validation set. The similarity function $\text{sim}(\cdot, \cdot)$ in (3) is implemented by the cosine function $\cos(\cdot, \cdot)/\tau'$, where τ' is also a tunable temperature constant.

4.2 Overall performance comparison (R1)

This section aims to answer the research question R1 by extensive performance comparison between SOTA results and our results based on the four fairly re-implemented baselines (i.e., HGA, IGV-B, CoMem, and MSPAN) across the four public benchmarks.

4.2.1 *Comparison with the baselines*

We have the following observations from the experimental results presented in Tables 2 [4, 10, 27, 28, 32, 42, 45, 51, 52, 54–58] and 3 [11, 32, 45, 51, 52, 54, 59, 60].

Table 3 Performance comparison with SOTAs and baselines on the Traffic-QA dataset w.r.t. the micro accuracy (%)[a)]

	Method	Traffic-QA [11]			Ref.
		Basic-Un.	Attribution	Overall	
Results in [11]	CNN+LSTM [11]	–	–	30.78	CVPR 2022
	I3D+LSTM [11]	–	–	33.21	CVPR 2022
	BERT-VQA [59]	–	–	33.68	WACV 2020
	TVQA [60]	–	–	35.16	EMNLP 2018
	HCRN [45]	–	–	36.49	CVPR 2020
	Eclipse [11]	–	–	37.05	CVPR 2019
Baseline+Ours	HGA [51]	40.86	45.35	42.04	AAAI 2020
	+Ours	41.59$^{+\mathbf{1.8\%}}$	45.54$^{+\mathbf{0.4\%}}$	42.90$^{+\mathbf{2.0\%}}$	Ours
	IGV-B [32]	41.53	46.77	43.26	CVPR 2022
	+Ours	41.58$^{+\mathbf{0.1\%}}$	48.62$^{+\mathbf{4.0\%}}$	43.57$^{+\mathbf{0.7\%}}$	Ours
	CoMem [52]	41.61	47.20	43.93	CVPR 2018
	+Ours	42.56$^{+\mathbf{2.3\%}}$	49.66$^{+\mathbf{5.2\%}}$	45.05$^{+\mathbf{2.5\%}}$	Ours
	MSPAN [54]	40.33	44.62	42.26	ACL 2021
	+Ours	41.03$^{+\mathbf{1.7\%}}$	45.91$^{+\mathbf{2.9\%}}$	42.85$^{+\mathbf{1.4\%}}$	Ours

a) Our relative improvements over the baselines are in bold.

Table 4 Performance (accuracy, %) comparison with SOTAs and baselines on the NExT-QA dataset

Method	NExT-QA [3]									
	Text-Rep.	Causal	Temporal	Description	Accuracy	Text-Rep.	Causal	Temporal	Description	Accuracy
EVQA [8]	Glove	28.69	31.27	41.44	31.51	BERT-FT	42.64	46.34	45.82	44.24
STVQA [51]	Glove	36.25	36.29	55.21	39.21	BERT-FT	44.76	49.26	55.86	47.94
HME [27]	Glove	37.97	36.91	51.87	39.79	BERT-FT	46.18	48.20	58.3	48.72
HCRN [45]	Glove	39.09	40.01	49.16	40.95	BERT-FT	45.91	49.26	53.67	48.20
IGV [32]	Glove	40.29	40.87	60.21	43.74	BERT-FT	48.56	51.67	59.64	51.34
HGA [51]	Glove	39.72	40.08	57.94	42.82	BERT-FT	47.67	48.96	60.71	50.21
+Ours	Glove	41.71	42.38	60.43	44.99	BERT-FT	49.27	50.43	60.00	51.39
MSPAN [54]	Glove	42.51	41.96	60.50	45.29	BERT-FT	48.13	49.91	58.86	50.44
+Ours	Glove	42.91	43.66	61.71	46.23	BERT-FT	49.47	51.30	60.21	51.80

• The micro accuracies of all of the baselines are consistently improved by our method. Specifically, for the experiments on MSVD-QA, compared with the widely used HGA [51], the relative improvements of the micro accuracies over the four baselines are 5.5%, 5.5%, 2%, and 2.5%. For the experiments on the two larger datasets, the relative improvements are slightly smaller, i.e., 2.6%, 3.4%, 1.7%, and 1% on MSRVTT-QA and 2%, 0.7%, 2.5%, and 1.4% on Traffic-QA. The main reason might be that the public datasets are highly imbalanced, and the baseline methods can perform well on the head answer categories, which dominate the testing set.

• The macro accuracies of all of the baselines are significantly improved by a larger margin. Specifically, the relative improvements are 12.2%, 7.1%, 6.3%, and 8.5% on MSVD-QA and 25.1%, 29.9%, 13.1%, and 25.4% on MSRVTT-QA. The performance improvements indicate the effectiveness of our method in improving both accuracy and robustness. Benefiting from our method, the enhanced models possibly perform better in predicting the less-frequent answers.

• On the Attribution set of Traffic-QA, our results are significantly higher than the baselines, i.e., IGV-B, CoMem, and MSPAN. In particular, our result outperforms that of CoMem by an absolute improvement of 2.46%, indicating a strong capability of perceiving the causes of video activity, e.g., answering the question—What could cause this accident? or Which factors might have contributed to the accident?

• As shown in Table 4, we investigate the effectiveness of our method on the NExT-QA dataset using two types of textual representation, namely, Glove-based representation and fine-tuned BERT-based representation. We observe from the results shown in Table 4 that our model-agnostic method can improve the overall performance of two strong baselines. In particular, our method significantly outperforms its counterparts on the Causal and Temporal testing sets, which demonstrates the effectiveness of our method in improving the robustness and reliability of VideoQA models.

Table 5 Ablation studies of the effect of three components, namely, temporal order regularization (TO), contrastive learning objective (CL), and perturbation invariance regularization (Inv), on two datasets, i.e., MSVD-QA and MSRVTT-QA[a)]

Method	MSVD-QA [10]		MSRVTT-QA [10]	
	Micro accuracy (%)	Macro accuracy (%)	Micro accuracy (%)	Macro accuracy (%)
HGA [51]	38.89	10.48	38.01	5.41
HGA+mixup [48]	40.38	10.95	38.12	6.36
+TO	$39.35^{+\mathbf{1.2\%}}$	$10.44^{-0.4\%}$	$38.26^{+\mathbf{0.7\%}}$	$5.72^{+\mathbf{5.7\%}}$
+CL+TO	$39.40^{+\mathbf{1.3\%}}$	$11.02^{+\mathbf{5.2\%}}$	$38.12^{+\mathbf{0.3\%}}$	$6.48^{+\mathbf{19.8\%}}$
+CL+TO+Inv (ours)	$41.01^{+\mathbf{5.5\%}}$	$11.76^{+\mathbf{12.2\%}}$	$38.98^{+\mathbf{2.6\%}}$	$6.77^{+\mathbf{25.1\%}}$
MSPAN [54]	41.68	9.95	38.99	5.27
MSPAN+mixup [48]	41.28	10.56	38.61	5.70
+TO	$41.73^{+\mathbf{0.1\%}}$	$10.21^{+\mathbf{2.6\%}}$	$38.91^{-0.2\%}$	$5.52^{+\mathbf{4.7\%}}$
+CL+TO	$42.05^{+\mathbf{0.9\%}}$	$10.31^{+\mathbf{3.6\%}}$	$39.14^{+\mathbf{0.4\%}}$	$6.08^{+\mathbf{15.4\%}}$
+CL+TO+Inv (ours)	$42.73^{+\mathbf{2.5\%}}$	$10.80^{+\mathbf{8.5\%}}$	$39.38^{+\mathbf{1.0\%}}$	$6.61^{+\mathbf{25.4\%}}$

a) Our relative improvements over the baselines are in bold.

4.2.2 *Comparison with the SOTAs*

In Table 2, we include three groups of SOTA methods, namely, memory network-based, hierarchy structure-based, and graph neural network-based. Overall, all of our results based on the four baselines outperform those of the SOTA results, which demonstrates the strong potential of our proposed method. We also observe similar performance improvements over SOTAs on Traffic-QA in Table 3. Although we did not include the latest VideoQA results of transformer-based multimodal models pretrained on large-scale extra data [61, 62], we believe that our method can be effectively compatible with them and improve their performance by preventing the model from exploiting superficial correlations rooted in dataset bias. In Table 4, we observe that, by integrating our method with the MSPAN baseline, our method outperforms the SOTA IGV [32] method by a large absolute improvement of accuracy of 2.49% using Glove-based text representation. Specifically, our method outperforms the IGV by an absolute improvement of 2.6% on the causal group. When using the strong BERT-FT-based text representation, our results are comparable to those of IGV. Notably, our method can still outperform IGV on the challenging causal group.

4.3 In-depth study of our method

4.3.1 *Effects of different components (R2)*

As shown in Table 5, we investigate the effect of our method by gradually adding the three key components to two baselines, i.e., HGA and MSPAN. We observe that the three components, i.e., TO, CL, and Inv, can complement each other well. Specifically, (1) in general, only using the TO cannot bring consistent improvement over two baselines, as shown in Table 5, because it mainly aims to enforce the inherent sequential structure of videos in video representation by predicting the temporal orders as an auxiliary task. How much of the effect of TO can be retained in the final stage of cross-modality answer reasoning is unclear. However, TO still results in stable relative improvement w.r.t. macro accuracy on the large MSRVTT-QA dataset, i.e., 5.7% based on HGA and 4.7% based on MSPAN, which reflects the effectiveness of TO. (2) Using the combination of TO and CL, we observe small but consistent improvement of micro accuracy and more significant improvement of macro accuracy, e.g., 19.8% and 15.4% on MSRVTT-QA. This finding indicates that CL can indeed facilitate the learning of robust cross-modality fused representation, thus improving the accuracy of VideoQA. (3) By jointly using all of the three components, we observe the remarkable performance improvement w.r.t. both micro and macro accuracies, which indicates that our three components can mutually support each other. The complementation of TO, CL, and Inv has been well exploited in our method. We also compare our method with a popular self-supervised augmentation strategy, i.e., mixup [48], as shown in Table 5. We determine that, based on the same baseline framework, the performance of our method is more stable and significant because the mixup strategy is more suitable for the image-based classification task. For the challenging video sequence, the effect of the mixup strategy is insignificant.

Table 6 Performance (accuracy, %) comparison with four baselines across three testing subsets (i.e., High-Freq, Med-Freq, and Low-Freq) with different answer frequencies (i.e., high, medium, and low) in the training sets of MSVD-QA and MSRVTT-QA[a)]

Method	MSVD-QA [10]					
	Micro accuracy			Macro accuracy		
	High-Freq	Med-Freq	Low-Freq	High-Freq	Med-Freq	Low-Freq
HGA	60.46	35.62	9.97	55.86	27.79	5.5
+Ours	$63.03^{+\mathbf{4.3\%}}$	$37.15^{+\mathbf{4.3\%}}$	$12.50^{+\mathbf{25.4\%}}$	$56.53^{+\mathbf{1.2\%}}$	$28.35^{+\mathbf{2.0\%}}$	$6.94^{+\mathbf{26.2\%}}$
IGV-B	63.32	36.22	8.86	57.25	28.23	4.53
+Ours	$66.03^{+\mathbf{4.3\%}}$	$38.60^{+\mathbf{6.6\%}}$	$10.39^{+\mathbf{17.3\%}}$	$60.77^{+\mathbf{6.1\%}}$	$29.41^{+\mathbf{4.2\%}}$	$5.04^{+\mathbf{11.3\%}}$
CoMem	64.33	34.12	7.79	56.90	26.50	3.88
+Ours	$63.56^{-1.2\%}$	$37.32^{+\mathbf{9.4\%}}$	$8.86^{+\mathbf{13.7\%}}$	$58.60^{+\mathbf{3.0\%}}$	$28.33^{+\mathbf{6.9\%}}$	$4.16^{+\mathbf{7.2\%}}$
MSPAN	66.14	37.42	8.93	60.62	28.36	4.56
+Ours	$65.74^{-0.6\%}$	$40.35^{+\mathbf{7.8\%}}$	$10.80^{+\mathbf{20.9\%}}$	$61.09^{+\mathbf{0.8\%}}$	$30.50^{+\mathbf{7.5\%}}$	$5.17^{+\mathbf{13.4\%}}$
Method	MSRVTT-QA [10]					
	Micro accuracy			Macro accuracy		
	High-Freq	Med-Freq	Low-Freq	High-Freq	Med-Freq	Low-Freq
HGA	54.67	33.89	7.38	43.65	30.56	2.94
+Ours	$55.01^{+\mathbf{0.6\%}}$	$35.62^{+\mathbf{5.1\%}}$	$9.35^{+\mathbf{26.7\%}}$	$43.07^{-\mathbf{1.3\%}}$	$31.97^{+\mathbf{4.6\%}}$	$4.33^{+\mathbf{47.3\%}}$
IGV-B	55.21	32.23	6.98	43.46	28.57	2.73
+Ours	$55.40^{+\mathbf{0.3\%}}$	$35.69^{+\mathbf{10.7\%}}$	$9.23^{+\mathbf{32.2\%}}$	$43.97^{+\mathbf{1.2\%}}$	$32.44^{+\mathbf{13.5\%}}$	$4.11^{+\mathbf{50.5\%}}$
CoMem	54.57	34.20	6.80	41.98	30.20	2.49
+Ours	$54.83^{+\mathbf{0.5\%}}$	$35.37^{+\mathbf{3.4\%}}$	$8.04^{+\mathbf{18.2\%}}$	$43.54^{+\mathbf{3.7\%}}$	$31.67^{+\mathbf{4.9\%}}$	$3.07^{+\mathbf{23.3\%}}$
MSPAN	55.74	36.07	7.24	43.45	32.42	2.67
+Ours	$55.45^{-0.5\%}$	$36.53^{+\mathbf{1.3\%}}$	$9.24^{+\mathbf{27.6\%}}$	$44.18^{+\mathbf{1.7\%}}$	$33.19^{+\mathbf{2.4\%}}$	$4.06^{+\mathbf{52.1\%}}$

a) Our relative improvements over the baselines are in bold.

4.3.2 *Robustness to imbalanced answer annotations (R3)*

We investigate the robustness of our method by partitioning the testing set into three groups, namely, High-Freq, Med-Freq, and Low-Freq, based on the answer frequencies (i.e., high, medium, and low) in the corresponding training sets on MSVD-QA and MSRVTT-QA. As shown in Table 6, the relative improvements of all baselines become more significant when tested on three subsets in the order of High-Freq, Med-Freq, and Low-Freq. Specifically, when MSPAN is enhanced by our method on MSVD-QA, the performance w.r.t. micro accuracy is slightly decreased by 0.4% on High-Freq, whereas the relative improvements on Med-Freq and Low-Freq are larger, i.e., 7.8% and 20.9%, respectively. We observe similar trends of relative performance improvements over other baselines w.r.t. both micro and macro accuracies. These results provide clear evidence that can support our argument that most existing VideoQA methods usually perform well on the head answer categories but perform poorly on tail answer categories because of the overreliance on the spurious correlations between the question and the answer mainly rooted in the long-tailed distribution of answer annotations. Instead, our method is devised to force the model to concentrate on the crucial video content during the stages of cross-modality fusion and answer reasoning, which can prevent the model from exploiting the dataset bias, thus achieving better performance on the less-frequent groups, i.e., Med-Freq and Low-Freq and validating that our model can indeed improve the robustness of the VideoQA model.

4.3.3 *Out-of-distribution generalization (R3)*

We further investigate the effectiveness of our method by generalizing the well-trained models to the out-of-distribution domain. We observe from Figure 3 that, when conducting cross-dataset generalization from MSRVTT-QA to MSVD-QA, our method can significantly outperform the baselines by a large margin, i.e., 144.4% (micro) and 72.1% (macro) relative improvements based on HGA and 1.7K% (micro) and 5.5K% (macro) relative improvements based on MSPAN. This finding indicates the strong advantage of our method in improving the generalizability of the model. Furthermore, when conducting cross-dataset generalization from MSVD-QA to MSRVTT-QA, the improvements are inconsistent because MSVD-QA is a small dataset, and the model trained on it has limited generalizability.

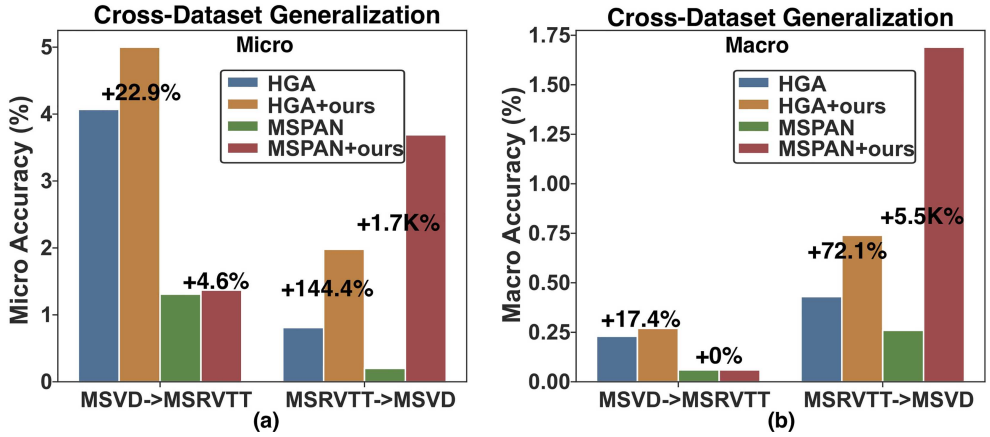

Figure 3 (Color online) Study of the cross-dataset generalizability of our approach on MSVD-QA and MSRVTT-QA. (a) Micro; (b) Macro.

Table 7 Performance (accuracy, %) comparison with baselines on six groups with different query types in Traffic-QA[a)]

Method	Traffic-QA [11]: query types					
	What	How	Which	Where	Was	Others
HGA	42.55	34.33	52.81	42.77	45.73	40.92
+Ours	$43.29^{+\mathbf{1.7\%}}$	$35.03^{+\mathbf{2\%}}$	$54.07^{+\mathbf{2.4\%}}$	$45.25^{+\mathbf{5.8\%}}$	$46.73^{+\mathbf{2.2\%}}$	$41.79^{+\mathbf{2.1\%}}$
MSPAN	41.98	34.24	51.54	43.18	46.23	45.24
+Ours	$42.35^{+\mathbf{7.4\%}}$	$35.03^{+\mathbf{2.3\%}}$	$52.67^{+\mathbf{2.2\%}}$	$43.60^{+\mathbf{1\%}}$	$48.24^{+\mathbf{4.3\%}}$	$48.13^{+\mathbf{6.4\%}}$

a) Our relative improvements over the baselines are in bold.

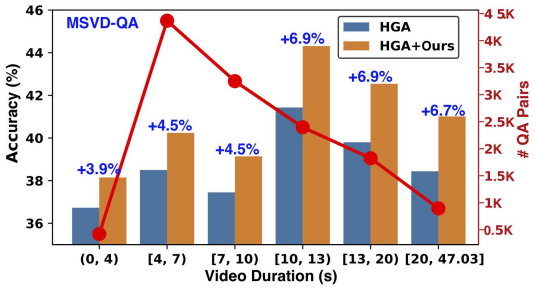

Figure 4 (Color online) Performance (micro, %) comparison across subsets with variant video lengths on the MSVD-QA testing set.

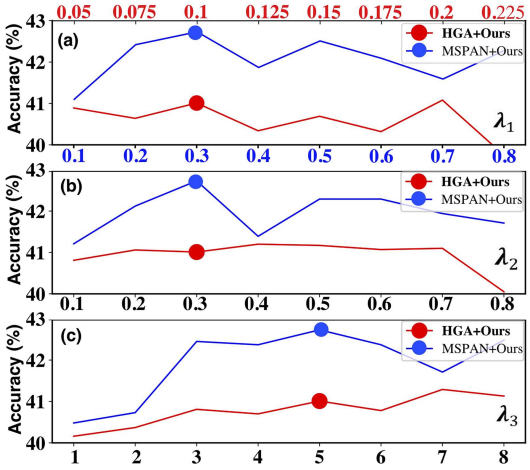

Figure 5 (Color online) Effects of the hyperparameters λ_1 (a), λ_2 (b), and λ_3 (c), investigated on MSVD-QA using HGA and MSPAN.

4.3.4 *Query types, video lengths, and hyperparameters (R4)*

To answer the research question R4, we further study the effect of our method across different testing subsets with variant query types or video durations, as shown in Table 7 and Figure 4. We also investigate the effects of the three hyperparameters expressed in (6) on MSVD-QA, as shown in Figure 5.

• **Query types.** We observe from Table 7 that, although our work mainly considers the video-based processing strategy rather than the language part, our method still consistently improves the two baselines for different query types in Traffic-QA. Specifically, the relative improvements over HGA are 1.7%, 2%, 2.4%, and 5.8% for the four popular query types, namely, What, How, Which, and Where, respectively. This finding can be mainly attributed to the CL, which can flexibly consider diverse negative pairs of multimodal input and enforce the model to carefully understand the linguistic semantics of queries.

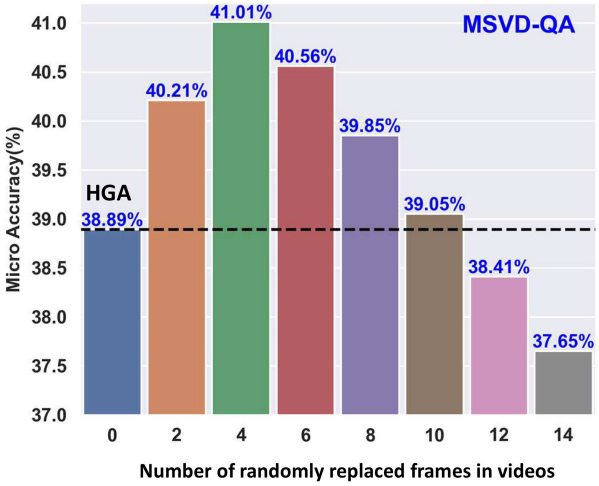

Figure 6 (Color online) Effect of our video perturbation strategy with different numbers of randomly replaced frames on MSVD-QA using HGA.

• **Video lengths.** We observe from Figure 4 that both the original baseline and our enhanced method do not perform well on short video groups because they do not contain sufficient temporal cues to answer complex natural language queries. The original baseline and our enhanced method exhibit the best performance in the medium-length video group $[10, 13)$. However, the performance of the original baseline and our enhanced method deteriorates when the video duration increases, which makes sense because long videos are more difficult to understand than short videos. Surprisingly, benefiting from our robust VideoQA method, we not only consistently surpass the HGA baseline but also achieve higher relative improvement on long video groups (approximately 7%) than short video groups (less than 5%). This finding indicates the advantage of our method in better modeling and leveraging the complex temporal content cues in videos for answer reasoning over long videos.

• **Hyperparameters.** We observe from Figure 5 that our method is insensitive to the three hyper-parameters. Take the strong baseline MSPAN as an example. Overall, we observe an increasing trend of performance when we increase the values of λ_1 and λ_2 from 0.1 to 0.3 and the value of λ_3 from 1 to 5. After peaking at $\lambda_1 = \lambda_2 = 0.3$ and $\lambda_3 = 5$, the performance gradually decreases with short-term fluctuations. We also observe a smoother change in the performance of our method based on HGA, as shown in Figure 5. As shown in Figure 6, we also test the effect of our sample and replace video perturbation strategy with different numbers of randomly replaced frames in videos. When the number of replaced frames is set to 0, the result corresponds to the performance of baseline HGA. When the number of replaced frames varies from 2 to 8, the performance of our method is relatively stable and peaks at 4. When the number of replaced frames is high, the performance of our method degrades rapidly because the key video content is damaged. Overall, the results indicate the robustness of our method against variations of different hyperparameters.

4.3.5 *Visualization*

We first visualize four representative VideoQA cases sampled from the MSVD-QA, MSRVTT-QA, and Traffic-QA datasets in Figure 7. In the Object case (a), enhanced by our strategy, both methods accurately predict the correct answer horse, instead of the false positive answer dog predicted by the baselines. Notably, the annotation frequency of dog is three times higher than that of horse in the training set. The baseline methods tend to exploit the dataset bias instead of the visual fact in a given video, thus having limited generalizability. A similar observation can also be drawn in the Number case (b). For the query type "How many, the baseline methods return the Top 1 answer two whose annotation frequency is seven times higher than that of the true answer three. Meanwhile, our method can prevent the model from exploiting spurious correlations, thereby identifying and justifying the correct answer. Our method also performs well in the Action case (c), which indicates that our method has indeed comprehensively watched the video before making a decision. We have a similar observation of the Attribution case (d), validating the effectiveness of our VideoQA approach. We also visualize two representative and challenging VideoQA cases sampled from the Temporal and Causal groups in the NExT-QA dataset in Figure 8.

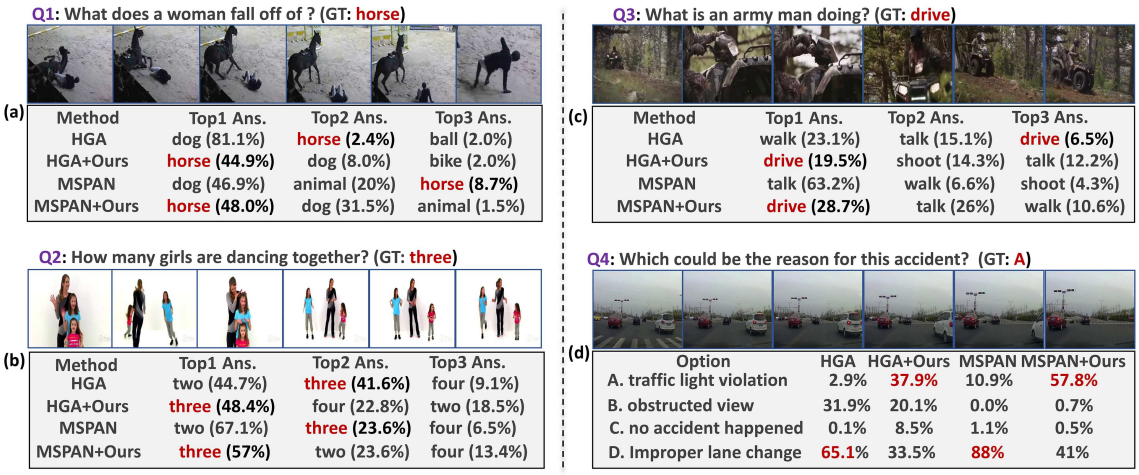

Figure 7 (Color online) Visualization of VideoQA cases from MSVD-QA (a), MSRVTT-QA (b, c), and Traffic-QA (d).

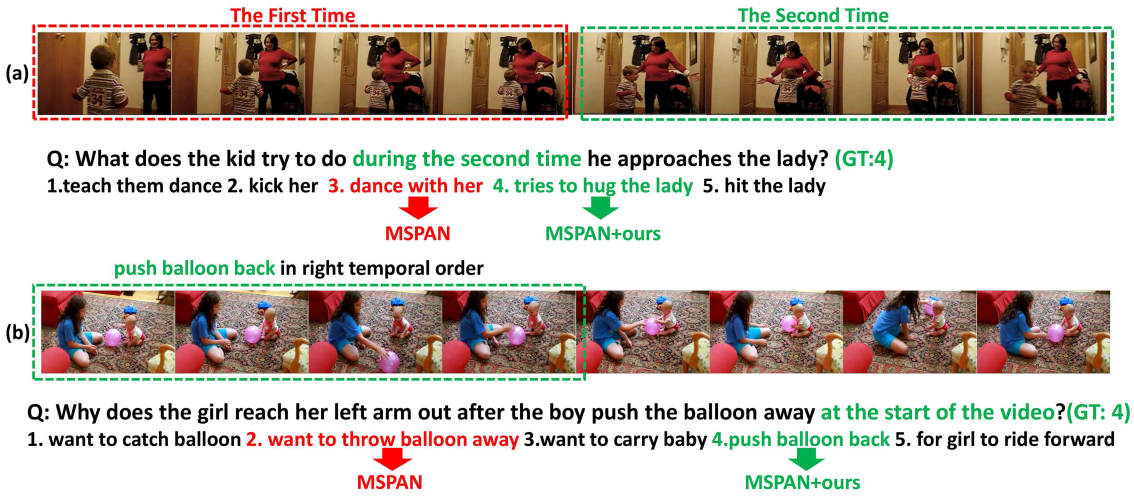

Figure 8 (Color online) Visualization of VideoQA cases from the temporal (a) and causal (b) groups in the NExT-QA dataset.

As shown in Figure 8(a), the model should correctly understand the temporal video characteristic. In particular, the model should be able to identify the differences between the actions dance and hug. The results show that our model can understand the spatiotemporal interactions in videos better than its counterparts, which indicates the effectiveness of our solution.

5 Conclusion

In this work, we have devised an effective, robust learning strategy for VideoQA. The main goal is to force the model to concentrate more on the crucial video content instead of overrelying on the superficial correlations rooted in dataset bias. Our robust learning strategy mainly consists of three parts, namely, self-supervised CL, TO, and Inv. Sufficient experimental results show that our method consistently outperforms different baselines in both accuracy and robustness. The efficacy of our approach has been well investigated and verified. In the future, we will attempt to couple our robust learning strategy with the transformer-based large multimodal model [62] for addressing other video-language tasks [63–66], such as video grounding [67, 68].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62272435, U22A2094), and Advanced Computing Resources Provided by the Supercomputing Center of the University of Science and Technology of China (USTC).

References

1 Yang H, Chaisorn L, Zhao Y, et al. VideoQA: question answering on news video. In: Proceedings of ACM International Conference on Multimedia, 2003. 632–641
2 Li K, Li J, Guo D, et al. Transformer-based visual grounding with cross-modality interaction. ACM Trans Multimed Comput Commun Appl, 2023, 19: 1–19
3 Xiao J, Shang X, Yao A, et al. NExT-QA: next phase of question-answering to explaining temporal actions. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 9777–9786
4 Xiao J, Yao A, Liu Z, et al. Video as conditional graph hierarchy for multi-granular question answering. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 2804–2812
5 Zhu L, Xu Z, Yang Y, et al. Uncovering the temporal context for video question answering. Int J Comput Vis, 2017, 124: 409–421
6 Kim K M, Choi S H, Kim J H, et al. Multimodal dual attention memory for video story question answering. In: Proceedings of European Conference on Computer Vision, 2018. 673–688
7 Li Y, Yang X, Zhang A, et al. Redundancy-aware transformer for video question answering. In: Proceedings of ACM International Conference on Multimedia, 2023. 3172–3180
8 Antol S, Agrawal A, Lu J, et al. VQA: visual question answering. In: Proceedings of International Conference on Computer Vision, 2015. 2425–2433
9 Kafle K, Kanan C. An analysis of visual question answering algorithms. In: Proceedings of International Conference on Computer Vision, 2017. 1965–1973
10 Xu D, Zhao Z, Xiao J, et al. Video question answering via gradually refined attention over appearance and motion. In: Proceedings of ACM International Conference on Multimedia, 2017. 1645–1653
11 Xu L, Huang H, Liu J. SUTD-TrafficQA: a question answering benchmark and an efficient network for video reasoning over traffic events. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 9878–9888
12 Zhou S, Guo D, Li J, et al. Exploring sparse spatial relation in graph inference for text-based VQA. IEEE Trans Image Process, 2023, 32: 5060–5074
13 Zhou S, Guo D, Yang X, et al. Graph pooling inference network for text-based VQA. ACM Trans Multimed Comput Commun Appl, 2024, 20: 1–21
14 Li L, Gan Z, Cheng Y, et al. Relation-aware graph attention network for visual question answering. In: Proceedings of International Conference on Computer Vision, 2019. 10312–10321
15 Li G, Wang X, Zhu W. Boosting visual question answering with context-aware knowledge aggregation. In: Proceedings of ACM International Conference on Multimedia, 2020. 1227–1235
16 Zhang P, Li X, Hu X, et al. VinVL: revisiting visual representations in vision-language models. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 5575–5584
17 Agrawal A, Batra D, Parikh D, et al. Don't just assume; look and answer: overcoming priors for visual question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2017. 4971–4980
18 Cadene R, Dancette C, Ben-Younes H, et al. RUBi: reducing unimodal biases for visual question answering. In: Proceedings of Conference on Neural Information Processing Systems, 2019. 841–852
19 Niu Y, Tang K, Zhang H, et al. Counterfactual VQA: a cause-effect look at language bias. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 12700–12710
20 Selvaraju R R, Lee S, Shen Y, et al. Taking a hint: leveraging explanations to make vision and language models more grounded. In: Proceedings of International Conference on Computer Vision, 2019. 2591–2600
21 Wu J, Mooney R J. Self-critical reasoning for robust visual question answering. In: Proceedings of Conference on Neural Information Processing Systems, 2019. 8604–8614
22 Han X, Wang S, Su C, et al. Greedy gradient ensemble for robust visual question answering. In: Proceedings of International Conference on Computer Vision, 2021. 1564–1573
23 Chen L, Yan X, Xiao J, et al. Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020. 10797–10806
24 Chen L, Zheng Y, Xiao J. Rethinking data augmentation for robust visual question answering. In: Proceedings of European Conference on Computer Vision, 2022. 95–112
25 Wang H, Guo D, Hua X, et al. Pairwise vlad interaction network for video question answering. In: Proceedings of ACM International Conference on Multimedia, 2021. 5119–5127
26 Li F, Bai T, Cao C, et al. Relation-aware hierarchical attention framework for video question answering. In: Proceedings of International Conference on Multimedia Retrieval, 2021. 164–172
27 Fan C, Zhang X, Zhang S, et al. Heterogeneous memory enhanced multimodal attention model for video question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2019. 1999–2007
28 Gu M, Zhao Z, Jin W, et al. Graph-based multi-interaction network for video question answering. IEEE Trans Image Process, 2021, 30: 2758–2770
29 Zhang J, Shao J, Cao R, et al. Action-centric relation transformer network for video question answering. IEEE Trans Circ Syst Video Technol, 2022, 32: 63–74
30 Cai J, Yuan C, Shi C, et al. Feature augmented memory with global attention network for videoQA. In: Proceedings of International Joint Conferences on Artificial Intelligence, 2020. 998–1004
31 Cherian A, Hori C, Marks T K, et al. (2.5 + 1)D spatio-temporal scene graphs for video question answering. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 444–453
32 Li Y, Wang X, Xiao J, et al. Invariant grounding for video question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2022. 2928–2937
33 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning, 2020. 1597–1607
34 Zhu J, Wang Z, Chen J, et al. Balanced contrastive learning for long-tailed visual recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2022. 6898–6907
35 Yang C, An Z, Zhou H, et al. Online knowledge distillation via mutual contrastive learning for visual recognition. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 10212–10227
36 Ma Y, Xu G, Sun X, et al. X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In: Proceedings of ACM International Conference on Multimedia, 2022. 638–647
37 Liang Z, Jiang W, Hu H, et al. Learning to contrast the counterfactual samples for robust visual question answering.

In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2020. 3285–3292
38 Si Q, Liu Y, Meng F, et al. Towards robust visual question answering: making the most of biased samples via contrastive learning. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2022. 6650–6662
39 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2016. 770–778
40 Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018. 6546–6555
41 Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018. 6077–6086
42 Park J, Lee J, Sohn K. Bridge to answer: structure-aware graph interaction network for video question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2021. 15526–15535
43 Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186
44 Jiang J, Chen Z, Lin H, et al. Divide and conquer: question-guided spatio-temporal contextual attention for video question answering. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 11101–11108
45 Le T M, Le V, Venkatesh S, et al. Hierarchical conditional relation networks for video question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2020. 9972–9981
46 Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization. 2019. ArXiv:1907.02893
47 Liu J, Hu Z, Cui P, et al. Heterogeneous risk minimization. In: Proceedings of International Conference on Machine Learning, 2021. 6804–6814
48 Zhang H, Cisse M, Dauphin Y N, et al. mixup: beyond empirical risk minimization. In: Proceedings of International Conference on Learning Representations, 2018
49 Xu D, Xiao J, Zhao Z, et al. Self-supervised spatiotemporal learning via video clip order prediction. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2019. 10334–10343
50 Lee H Y, Huang J B, Singh M, et al. Unsupervised representation learning by sorting sequences. In: Proceedings of International Conference on Computer Vision, 2017. 667–676
51 Jiang P, Han Y. Reasoning with heterogeneous graph alignment for video question answering. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 11109–11116
52 Gao J, Ge R, Chen K, et al. Motion-appearance co-memory networks for video question answering. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018. 6576–6585
53 Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: dynamic memory networks for natural language processing. In: Proceedings of International Conference on Machine Learning, 2016. 1378–1387
54 Guo Z, Zhao J, Jiao L, et al. Multi-scale progressive attention network for video question answering. In: Proceedings of Association for Computational Linguistics, 2021. 973–978
55 Peng L, Yang S, Bin Y, et al. Progressive graph attention network for video question answering. In: Proceedings of ACM International Conference on Multimedia, 2021. 2871–2879
56 Dang L H, Le T M, Le V, et al. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In: Proceedings of International Joint Conferences on Artificial Intelligence, 2021. 636–642
57 Huang D, Chen P, Zeng R, et al. Location-aware graph convolutional networks for video question answering. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 11021–11028
58 Wang J, Bao B K, Xu C. DualVGR: a dual-visual graph reasoning unit for video question answering. IEEE Trans Multimedia, 2022, 24: 3369–3380
59 Yang Z, Garcia N, Chu C, et al. BERT representations for video question answering. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2020. 1556–1565
60 Lei J, Yu L, Bansal M, et al. TVQA: localized, compositional video question answering. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2018. 1369–1379
61 Tang Z, Cho J, Lei J, et al. PERCEIVER-VL: efficient vision-and-language modeling with iterative latent attention. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2023
62 Xiao J, Zhou P, Yao A, et al. Contrastive video question answering via video graph transformer. 2023. ArXiv:2302.13668
63 Song P, Guo D, Yang X, et al. Emotion-prior awareness network for emotional video captioning. In: Proceedings of ACM International Conference on Multimedia, 2023. 589–600
64 Yang X, Dong J, Cao Y, et al. Tree-augmented cross-modal encoding for complex-query video retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020. 1339–1348
65 Dong J, Li X, Xu C, et al. Dual encoding for video retrieval by text. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 4065–4080
66 Zheng Q, Dong J, Qu X, et al. Progressive localization networks for language-based moment localization. ACM Trans Multimedia Comput Commun Appl, 2023, 19: 1–21
67 Yang X, Feng F, Ji W, et al. Deconfounded video moment retrieval with causal intervention. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. 1–10
68 Yang X, Wang S, Dong J, et al. Video moment retrieval with cross-modal neural architecture search. IEEE Trans Image Process, 2022, 31: 1204–1216