Collaborative Aware Bidirectional Semantic Reasoning for Video Question Answering

Xize Wu, Jiasong Wu, Lei Zhu, Lotfi Senhadji, Huazhong Shu

▶ **To cite this version:**

Xize Wu, Jiasong Wu, Lei Zhu, Lotfi Senhadji, Huazhong Shu. Collaborative Aware Bidirectional Semantic Reasoning for Video Question Answering. IEEE Transactions on Circuits and Systems for Video Technology, 2024, pp.1-1. 10.1109/tcsvt.2024.3490665 . hal-04780082

HAL Id: hal-04780082

https://hal.science/hal-04780082v1

Submitted on 27 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

Collaborative Aware Bidirectional Semantic Reasoning for Video Question Answering

Xize Wu, Jiasong Wu, *Member, IEEE*, Lei Zhu, *Senior Member, IEEE*, Lotfi Senhadji, *Senior Member, IEEE*, Huazhong Shu, *Senior Member, IEEE*

***Abstract*—Video question answering (VideoQA) is the challenging task of accurately responding to natural language questions based on a given video. Most previous methods focus on designing complex cross-modal interactions to perform question-oriented video scene mining and semantic reasoning, and utilize straightforward classification and matching strategies with different decoders to forcibly associate the predicted representation with ground-truth answer. However, the limitations of question-oriented reasoning and the overlapping semantic co-occurrences between questions and candidates may cause them to fall into spurious correlation reasoning. In this paper, we propose a Collaborative aware Bidirectional Semantic Reasoning (CBSR) model to alleviate this challenging problem. Specifically, we first propose a collaborative aware adaptive correlation reasoning module to collaboratively mine multi-granularity text-aware critical video scenes and reason about the complex intrinsic correlations between them via bottom-up cross-granularity adaptive aggregation. By progressively performing video reasoning from object-level to frame-level, we can obtain a set of semantically rich critical video representations. Then, we collaboratively decode it together with question and knowledge semantics into an implicit representation through the proposed unified answer semantic collaborated decoding module. Finally, a novel bidirectional semantic reasoning learning strategy is proposed to bridge and strengthen the unique positive semantic correlation between the learned implicit representation and the ground-truth answer, and explicitly alleviate the challenge of overlapping semantic co-occurrence. Benefiting from the same model structure and learning strategy, our method can achieve seamless transfer between Open-Ended and Multi-Choice tasks. Extensive experimental results on seven commonly tested datasets (i.e. MSVD-QA, MSRVTT-QA, NExT-QA, Causal-VidQA, NExT-OOD, ActivityNet-QA and EgoSchema) verify the superior performance of our method and the effectiveness of each reasoning module. We provide our source codes and experimental datasets at https://github.com/XizeWu/CBSR.**

***Index Terms*—Video question answering, Collaborative aware, Bidirectional semantic reasoning, Multi-granularity aggregation**

This work was supported in part by the National Key Research and Development Program of China (No. 2021ZD0113202), and in part by the National Natural Science Foundation of China under Grants 62476055, 62171125, and in part by the innovation project of Jiangsu Province under grants BZ2023042, BY2022564.

Xize Wu, Jiasong Wu and Huazhong Shu are with LIST, Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing 210096, China; Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China; Centre de Recherche en Information Biomedicale Sino-français (CRIBs), Rennes F-3502, France. (e-mail: xizewu96@gmail.com, jswu@seu.edu.cn, shu.list@seu.edu.cn). Jiasong Wu is the corresponding author.

Lei Zhu is with the School of Electronic and Information Engineering, Tongji University, Shanghai 200092, China. (e-mail: leizhu0608@gmail.com).

Lotfi Senhadji is with Univ-Rennes, INSERM, LTSI-UMR 1099, Rennes F-3502, France; and with CRIBs, Rennes F-3502, France. (e-mail: lotfi.senhadji@univ-rennes1.fr).

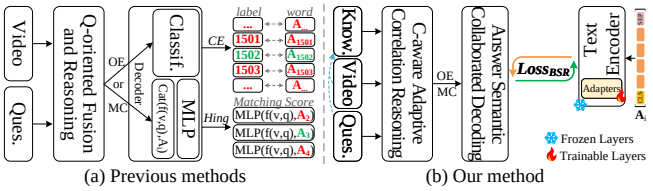


Fig. 1. Comparison between previous methods and our method. (1) Previous methods mostly utilize the given question to individually guide the model's reasoning. In contrast, the proposed method first generates a coarse knowledge description, and then performs collaborative aware reasoning together with the question semantics. (2) Most previous methods adopt a simple discrete classification strategy and matching strategy with different decoders on OE task and MC task. Differently, we propose a novel Bidirectional Semantic Reasoning (BSR) learning strategy with a unified answer decoder to uniformly perform answer decoding on both Open-Ended (OE) and Multi-Choice (MC) tasks, which facilitates seamless transfer between these two reasoning tasks.

I. Introduction

VIDEO question answering (VideoQA) [1]–[4] is one of the most representative research hotspots in interactive artificial intelligence [5]–[11], which has recently attracted increasing interests from many researchers. It requires the VideoQA model to understand the complex semantics in a given question, and then mine the crucial semantic information in the video to predict the positive answer.

To solve this challenging task, numerous methods [12]–[15] have been proposed with impressive reasoning performance. Although promising progresses have been made so far, most existing methods still suffer from the following two bottlenecks: (1) As shown in Fig. 1(a), previous methods [16]–[18] tend to design a variety of complex interactions for question-oriented video clue mining and semantic reasoning. However, the question-oriented design may cause the VideoQA model to overly focus on video scenes semantically related to the question, and eventually fall into spurious causal reasoning. As shown in Fig. 2(a), the question-oriented strategy tends to reason based on the question-focused 1st, 3rd and 4th frame ("people") while overlooking the critical information in the 2nd frame ("TV"). In fact, the ideal reasoning is that the VideoQA model predicts the correct answer based on all positive video scenes instead of only question-oriented positive video scenes. (2) In addition, most previous methods [19]–[21] regard the Open-Ended (OE) task as a simple classification task, using a simple classification layer to forcibly amplify the connection from the predicted feature to the ground-truth discrete label; or regard the Multi-Choice (MC) task as

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

a matching task, using a Multi-Layer Perceptron (MLP) to concatenate the predicted feature and each candidate feature, then widening the score gap between the positive matching pair and all negative matching pairs using a fixed margin. On the one hand, the discriminative identity semantics implicit in each answer are ignored on OE task, potentially weakening the model in distinguishing the positive answer from hard negative candidates. On the other hand, on MC task, existing matching strategy not only introduces additional computational overhead caused by the concatenation operation, but also fails to effectively alleviate the detrimental impact caused by diverse overlapping semantic co-occurrences (underlined in blue and orange in Fig. 2(b)).

Motivated by the above analysis, in this paper, we propose a Collaborative aware Bidirectional Semantic Reasoning (CBSR) method to solve the limitations of existing methods, illustrated in Fig. 3. It first generates a coarse knowledge description by feeding the given video-question pair into an off-the-shelf Visual-Language Model. We regard individual question semantics or knowledge semantics as local text semantics, and their combined augmented semantics as global text semantics. Then, we propose a Collaborative aware Adaptive Correlation Reasoning (CACR) module to collaboratively mine multi-granularity text-aware critical video scenes, and progressively reason the complex correlations among them. Specifically, it first mines critical video scenes semantically related to global and local text, and views the remaining ones as noisy video scenes and filters them out. After that, we adaptively aggregate local-aware video semantics into global-aware video semantics in a bottom-up manner to reason the complex intrinsic correlations among critical scenes. Through progressive video semantic reasoning from the object-level to the frame-level, it finally derives semantically rich critical video representations. Subsequently, a unified Answer Semantic Collaborated Decoding (ASCD) module is proposed to collaboratively decode an implicitly predicted feature as an output. Finally, we propose a Bidirectional Semantic Reasoning (BSR) learning strategy to encourage the model to learn the unique intrinsic semantic correlation between the implicitly predicted feature and the ground-truth answer on both OE and MC tasks. It motivates the model to flexibly build the positive correlation from the predicted feature to the ground-truth answer while suppressing spurious correlations to all negative candidates without a prespecified margin via forward semantic reasoning. And it widens the negative semantic gap between the positive answer and all negative candidates to reversely strengthen the unique positive semantic correlation from the positive answer to the predicted feature via reverse semantic reasoning. Therefore, the proposed BSR not only avoids the additional computational overhead introduced by the concatenation operation, but also explicitly alleviates the detrimental impact caused by overlapping semantic co-occurrences that commonly exist between question and candidates.

Briefly, we summarize the main contributions as follows:

- We propose a novel Collaborative aware Adaptive Correlation Reasoning (CACR) module, which utilizes the given question and generated coarse knowledge to collaboratively mine critical video scenes and guide complex intrinsic correlation reasoning among them.
- We propose a unified Answer Semantic Collaborative Decoding (ASCD) module with a novel Bidirectional Semantic Reasoning (BSR) learning strategy, which decodes an implicitly predicted representation semantically related to the positive answer on both OE and MC tasks. By sharing the same model structure and learning strategy, our method can achieve seamless transfer between these two reasoning tasks.
- Extensive experimental results on seven widely tested datasets (including two descriptive datasets MSVD-QA and MSRVTT-QA, two causal datasets NExT-QA and Causal-VidQA, one Out-Of-Distribution dataset NExT-OOD, two long-form datasets ActivityNet-QA and EgoSchema) verify the superior performance of the proposed method.

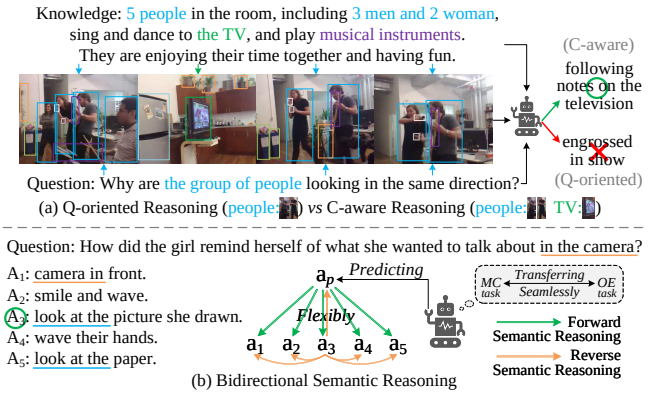


Fig. 2. (a) Previous question-oriented methods may fall into insufficient mining of critical video scenes and ultimately derive an unsatisfactory reasoning result (×). The proposed method uses the question and knowledge to collaboratively mine critical video scenes and reason the real intrinsic correlation among them to predict the correct answer (◯). (b) The proposed bidirectional semantic reasoning includes forward and reverse semantic reasoning. It encourages the model to flexibly build the positive correlation from the predicted feature to the ground-truth answer while suppressing spurious correlations to all negative candidates without a prespecified margin via forward semantic reasoning. And it widens the semantic gap between the positive answer and all negative candidates to reversely strengthen the unique positive semantic correlation from the positive answer to the predicted feature via reverse semantic reasoning. The proposed model can achieve seamless transfer between the MC and OE tasks because it shares the same model structure and learning strategy.

II. RELATED WORK

Video Question Answering (VideoQA) requires models to effectively understand the complex semantics hidden in the sequential video and deduce the correct answer based on the given question, presenting greater challenges compared to Image Question Answering (ImageQA) [22]–[26]. Existing methods can be broadly categorized as follows.

Memory learning methods [1], [27]–[29]: Early methods often develop diverse intra- and inter-modal memory mechanisms, which perform iterative read and write operations to understand video content. These designs are well consistent with the temporal properties of videos. Representatively, AMU [30] introduces a co-memory attention mechanism to gradually

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

refine both appearance and motion features by leveraging question guidance, and finally fuse them for prediction. HME [12] proposes a pair of heterogeneous memory mechanisms to learn the global video semantics and understand the complex question semantics, respectively. However, these methods ignore the exploration of fine-grained visual semantics within individual frames, resulting in insufficient visual perception. Different from them, our method introduces fine-grained visual scenes and further explores their complex interrelationships to understand fine-grained visual contents.

Graph learning methods [13], [16], [31]–[33]: Some methods leverage the powerful modeling capabilities of graph neural networks [34], [35] to explore the intrinsic adjacency semantics between video scenes. Such designs explicitly enhance the understanding of relational structures among complex interactive scenes. For instance, LGCN [33] models the correlation between visual objects through a vanilla GCN to capture the locally-aware video semantics, and then fuses video features with question features to predict the answer. B2A [13] proposes a question-to-visual interaction to mine question-related appearance features and motion features, and then employs a question graph as a bridge to align complementary appearance semantics and motion semantics. However, these methods regard all frames or objects as nodes in a graph to model redundant correlations between them, which may weaken the critical adjacency semantics and result in suboptimal performance. Unlike them, we first filter out noisy visual scenes and then progressively reason about complex intrinsic correlations between critical scenes via the proposed cross-granularity adaptive aggregation.

Hierarchical learning methods [14], [17], [36], [37]: Some methods argue that complete video (or question) semantics can be hierarchically decomposed into multi-level and multi-granular visual (or textual) semantics. This hierarchical design facilitates a better understanding of visual structures and textual concepts by progressively aggregating multi-granular video and question semantics. HCRN [14], a representative hierarchical learning method, proposes a general-purpose reusable conditional relational network unit to capture video content. It takes video frame features as input and uses video motion features and question features as conditional information to guide the hierarchical reasoning process. HQGA [17] introduces a conditional graph hierarchy with multiple levels of question injection to reason and aggregate low-level visual elements into high-level video elements. However, similar to graph learning methods, these approaches introduce a large amount of question-irrelevant noisy video scenes for hierarchical aggregation, which inevitably brings negative impacts to the answer prediction. Our method instead filters out noisy video scenes via the proposed collaborative pruning mechanism, and then adaptively reasons and aggregates collaborative aware appearance and motion features for prediction.

Causal learning methods [15], [18], [38]–[41]: Some recent methods advocate that only a few visual scenes are critical for reasoning about the correct answer, thus focusing on designing various causal interventions to explicitly explore the few but essential visual scenes relevant to the question semantics. These designs are advantageous in alleviating the negative impact of noisy scenes and establishing causal links from learned rationales to predictions. IGV [15] is the first model to address VideoQA from a causal perspective. It proposes a grounding indicator and scene intervener to eliminate question-irrelevant scenarios, and predict answers based on identified question-relevant visual scenarios. To improve reasoning transparency and mitigate spurious correlations caused by co-occurring concepts, VSCR [18] performs scene separation interventions at both the frame and segment levels, and then adopts a contrastive learning paradigm to supervise the model to learn positive predictions based on question-related video scenes. However, due to the incomplete semantics of the question, existing methods may overemphasize question-related superficial visual scenes, leading to spurious predictions, as illustrated in Fig. 2(a). Conversely, we propose a collaborative aware reasoning framework to mine complementary visual semantics and a bidirectional reasoning strategy to mitigate the impact of negative answers.

TABLE I
NOTATIONS AND DESCRIPTION.

Notations	Description
O, F, M	Video features of object, appearance and motion
$\hat{O}^c$, $\hat{O}^q$, $\hat{O}^k$	The critical object associated with c, q or k
$\hat{O}$	The critical object after cross-granular aggregation
$\tilde{O}$, $\tilde{F}$, $\tilde{M}$	Video features of object, appearance and motion enhanced by adjacency semantics
Q, K	Local question feature and knowledge feature
C	Global text feature after concatenation of Q and K
$\bar{q}$, $\bar{k}$, $\bar{c}$	Sentence-level semantic feature of Q, K and C
$e_{i,j,s}^{c\leftarrow q}$	Cross-granularity correlation from the local q-aware s-th object to the global c-aware j-th object in the i-th frame
$e_{i,j,s}^{c\leftarrow k}$	Cross-granularity correlation from the local k-aware s-th object to the global c-aware j-th object in the i-th frame
g	Adaptive gating vector
A, $\hat{A}$	Candidate matrix for forward reasoning and pseudo candidate matrix for backward reasoning

III. THE PROPOSED METHOD

Fig. 3 shows the basic learning framework of the proposed method, which mainly consists of the following modules:

(a) Multi-modal Feature Extracting: Generating a coarse knowledge for each video-question pair via an off-the-shelf Visual-Language Model (VLM), and extracting multi-modal features via different feature extractors.

(b) Collaborative aware Adaptive Correlation Reasoning (CACR): Collaboratively mining global text-aware and local text-aware (question-aware and knowledge-aware) critical video scenes to progressively reasoning the complex intrinsic correlations among them from object-level to frame-level.

(c) Answer Semantic Collaborated Decoding (ASCD): Collaboratively decoding an implicitly predicted representation semantically related to the ground-truth answer.

(d) Candidate Learning with Adapters (CLA): Learning discriminative sentence-level candidate embeddings by optimizing the newly introduced lightweight adapters while freezing the original model parameters.

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

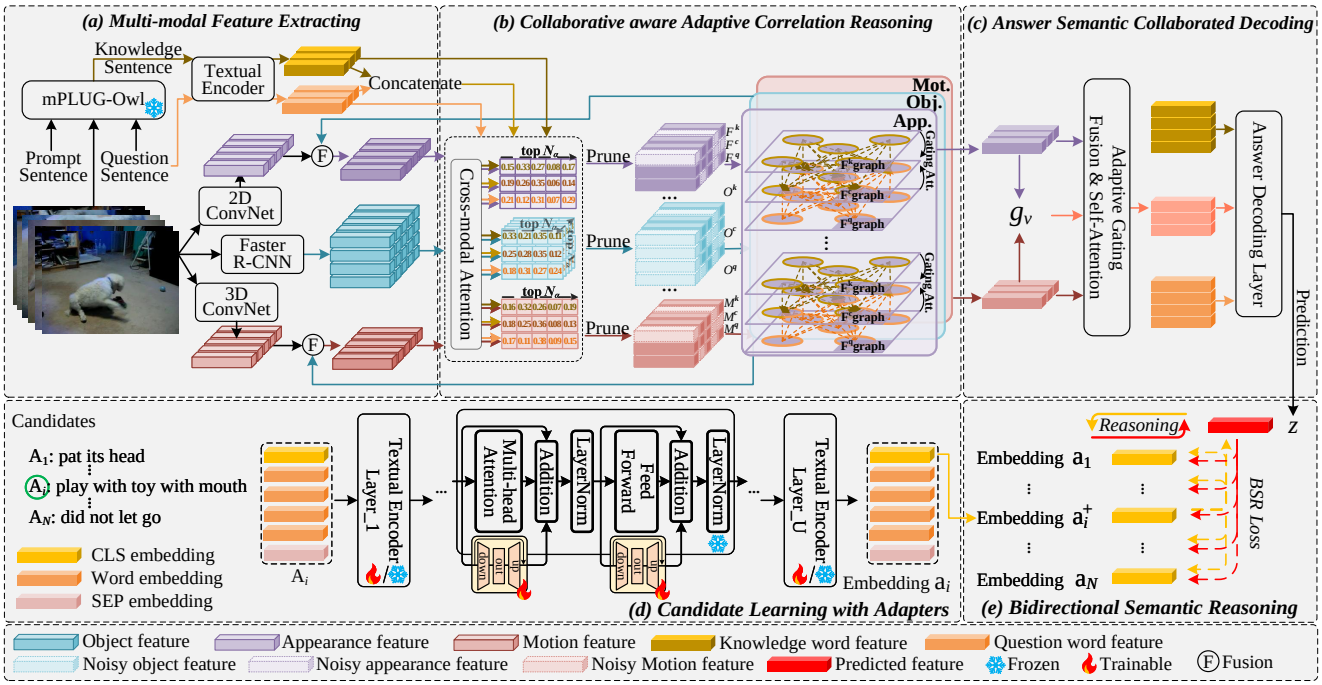


Fig. 3. The basic learning framework of the proposed CBSR. In (a), we first generate coarse knowledge based on the given video and question, and then extract video and text features through the modality-specific backbone. With progressive design (b), the Collaborative aware Adaptive Correlation Reasoning (CACR) module exploits question and knowledge semantics to collaboratively mine critical video scene information, and reasoning complex correlations between critical scenes through cross-granularity adaptive aggregation. Thereafter (c), we propose a unified Answer Semantic Collaborated Decoding (ASCD) module to adaptively fuse appearance and motion features, and then equip it with knowledge and question semantics to collaboratively decode an implicitly predicted representation. In (d), sentence-level candidate embeddings are learned by optimizing lightweight adapters. (e) The proposed Bidirectional Semantic Reasoning (BSR) learning strategy, including forward semantic reasoning and reverse semantic reasoning.

(e) Bidirectional Semantic Reasoning (BSR): Driving model learning through both forward semantic reasoning and reverse semantic reasoning.

For convenience, the relevant notations and descriptions mainly used in this paper are shown in Table I.

A. Multi-modal Feature Extracting

In this paper, we split a given video into N_T frames. Similar to the previous methods [14], we use pre-trained ResNet [42] and ResNeXt [43], [44] to extract the appearance and motion features of the video, respectively, and then use a FC layer to map them into a d-dimensional hidden space, denoted as $F = \{f_i\}_{i=1}^{N_T}$ and $M = \{m_i\}_{i=1}^{N_T}$, respectively. f_i and m_i represent the i-th appearance and motion feature, respectively. The commonly used Faster R-CNN [45] is adopted to extract N_O objects in each frame, denoted as $O = \{o_i\}_{i=1}^{N_T} = \{o_{i,j}\}_{i=1,j=1}^{N_T,N_O}$, where o_i represents all object features in the i-th frame, $o_{i,j}$ represents the feature of the j-th object in the i-th frame. Additionally, we design a prompt "***Please briefly describe the content of this video in one sentence based on the given question: {Q}***", and feed it into an off-the-shelf VLM (e.g., mPLUG-Owl [46], [47]) together with each video-question pair to obtain a coarse knowledge related to the given question, where ***{Q}*** represents a placeholder for a question. For a given question and generated knowledge, we utilize the commonly used Bert [48] model to encode each word into a d-dimensional hidden layer embedding, denoted as $Q = \{q_i\}_{i=1}^{L_Q}$ and K=$\{k_i\}_{i=1}^{L_K}$, where L_Q and L_K denote the length of question words and knowledge words, respectively. Furthermore, we concatenate question and knowledge to obtain a joint representation $C = \{c_i\}_{i=1}^{L_C}$, where $L_C = L_Q + L_K$. Since the joint feature C contains both question and knowledge semantics, we regard it as a global text feature while Q and K as local text features.

B. Collaborative aware Adaptive Correlation Reasoning

First, we introduce a Collaborative Pruning Mechanism (CPM) module to mine global text-aware and local text-aware critical video objects, and regard the remaining ones as noisy objects to be pruned out. Specifically, we first perform fine-grained cross-modal attention to highlight critical video objects invoked by global text c_l or local question q_l (or knowledge k_l) words while bridging the heterogeneous semantic gap:

$$o_{i,j}^{c} = o_{i,j} + \sum_{l=1}^{L_C} \frac{exp(o_{i,j}c_l)}{\Sigma_{r=1}^{L_C} exp(o_{i,j}c_r)} c_l, \tag{1}$$

$$o_{i,j}^{q} = o_{i,j} + \sum_{l=1}^{L_Q} \frac{exp(o_{i,j}q_l)}{\Sigma_{r=1}^{L_Q} exp(o_{i,j}q_r)} q_l, \tag{2}$$

$$o_{i,j}^{k} = o_{i,j} + \sum_{l=1}^{L_K} \frac{exp(o_{i,j}k_l)}{\Sigma_{r=1}^{L_K} exp(o_{i,j}k_r)} k_l. \tag{3}$$

ACCEPTED MANUSCRIPT / CLEAN COPY
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

Then, we can obtain three object-level confidence vectors by calculating the similarity between each sentence-level text feature ($\bar{c}$, $\bar{q}$, $\bar{k}$) and all object feature ($o_i^c = \{o_{i,j}^c\}_{j=1}^{N_o}$, $o_i^q = \{o_{i,j}^q\}_{j=1}^{N_o}$, $o_i^k = \{o_{i,j}^k\}_{j=1}^{N_o}$) in each frame:

$$S_i^c = Softmax(o_i^c \bar{c}^T), \tag{4}$$

$$S_i^q = Softmax(o_i^q \bar{q}^T), \quad S_i^k = Softmax(o_i^k \bar{k}^T), \tag{5}$$

where S_i^c and S_i^q (or S_i^k) represents the global text-aware and local question-aware (or knowledge-aware) confidence vector, respectively. To obtain the critical objects that are strongly related to the global semantic in the i-th frame, we sort all values in S_i^c in descending order, and extract the topN_α objects with the highest similarity while discarding the remaining weakly relevant objects. Formally, we formulate it as follows:

$$idx = Index(Sort_D(S_i^c)), \tag{6}$$

$$idx' = Sort_A(idx[:topN_\alpha]), \quad s.t.\ topN_\alpha = \lfloor \alpha N_O \rfloor \tag{7}$$

$$\hat{o}_i^c = o_i^c[idx'], \quad s.t.\ \hat{o}_i^c \in \mathbb{R}^{N_\alpha \times d} \tag{8}$$

where $Sort_A(\cdot)$ and $Sort_D(\cdot)$ represent ascending and descending operations, respectively, $Index(\cdot)$ represents getting index operation, $\lfloor * \rfloor$ represents rounding down the number $*$, $0 \leq \alpha \leq 1$ represent the pruning factor, $\hat{o}_i^c$ represents the topN_α critical objects. In Eq. (7), $Sort_A(\cdot)$ is used to preserve the original relative position relationship. Similarly, the topN_α objects of local question-aware and knowledge-aware can be denoted as $\hat{o}_i^q$ and $\hat{o}_i^k$, respectively. Benefiting from its training-free design, we can effectively alleviate the detrimental impact and additional computational overhead caused by negative objects.

Then, we propose a Cross-granularity Adaptive Aggregation (CAA) module, which adaptively aggregates local-aware object semantics into global-aware object semantics to reason the complex intrinsic correlations between critical video objects in a bottom-up manner. Specifically, we first calculate the cross-granularity correlation from each local-aware object $\hat{o}_{i,s}^q$ (or $\hat{o}_{i,s}^k$) to each global-aware object $\hat{o}_{i,j}^c$ as follows:

$$e_{i,j,s}^{c\leftarrow q} = \frac{exp(\phi_{oc}(\hat{o}_{i,j}^c)\phi_{oq}(\hat{o}_{i,s}^q))}{\sum_{l=1}^{N_\alpha} exp(\phi_{oc}(\hat{o}_{i,j}^c)\phi_{oq}(\hat{o}_{i,l}^q))}, \tag{9}$$

$$e_{i,j,s}^{c\leftarrow k} = \frac{exp(\phi_{oc}(\hat{o}_{i,j}^c)\phi_{ok}(\hat{o}_{i,s}^k))}{\sum_{l=1}^{N_\alpha} exp(\phi_{oc}(\hat{o}_{i,j}^c)\phi_{ok}(\hat{o}_{i,l}^k))}, \tag{10}$$

where $\phi_*(\cdot)$ denotes the FC layer. After that, we perform bottom-up cross-granularity adaptive aggregation guided by $\bar{c}$ from local-aware object semantics to global-aware object semantics based on the learned correlations:

$$\hat{o}_{i,j}^{c\leftarrow q} = \sigma(\sum_{s=1}^{N_\alpha} \phi_{q2c}(e_{i,j,s}^{c\leftarrow q}\ \hat{o}_{i,s}^q)), \tag{11}$$

$$\hat{o}_{i,j}^{c\leftarrow k} = \sigma(\sum_{s=1}^{N_\alpha} \phi_{k2c}(e_{i,j,s}^{c\leftarrow k}\ \hat{o}_{i,s}^k)), \tag{12}$$

$$g_o = Sigmoid(\phi_{con}([\hat{o}_{i,j}^{c\leftarrow q};\ \bar{c};\ \hat{o}_{i,j}^{c\leftarrow k}])), \tag{13}$$

$$\hat{o}_{i,j} = \hat{o}_{i,j}^c + g_o \odot \hat{o}_{i,j}^{c\leftarrow q} + (1 - g_o) \odot \hat{o}_{i,j}^{c\leftarrow k}, \tag{14}$$

where $\sigma(\cdot)$ is the $ReLU(\cdot)$ activation function, $\phi_*(\cdot)$ denotes the FC layers, $\odot$ represent the element-wise product. $\hat{O} = \{\hat{o}_i\}_{i=1}^{N_T} = \{\hat{o}_{i,j}\}_{i=1,j=1}^{N_T,N_\alpha}$ represents cross-granularity aggregated video objects. Next, we adopt a graph convolutional network to reason the correlations between them, and thus derive semantically rich objects $\tilde{O} = \{\tilde{o}_i\}_{i=1}^{N_T} = \{\tilde{o}_{i,j}\}_{i=1,j=1}^{N_T,N_\alpha}$ that are embedded with adjacency semantics:

$$\hat{E}_i^o = Softmax(\psi_1(\hat{o}_i)\psi_2(\hat{o}_i)^T), \tag{15}$$

$$\tilde{o}_i = \sigma((\hat{E}_i^o + I)\hat{o}_i W^o), \tag{16}$$

where $\psi_*(\cdot)$ represents the FC layer, I represents an identity matrix, W^o represents the learnable parameters. Finally, we employ a mean pooling operation to aggregate object-level features into frame-level features $O^F = \{o_i^F\}_i^{N_T}$.

Similarly, the semantically rich critical appearance representation $\tilde{F}$ and motion representation $\tilde{M}$ can also be obtained by employing similar reasoning operations as above. Specifically, a fusion operation is first introduced, which concatenates the aggregated frame-level object features and appearance (or motion) features, and map it into the unified d-dim layer space by a FC layer $\psi_f(\cdot)$ (or $\psi_m(\cdot)$):

$$\hat{F} = \psi_f([O^F; F]), \quad (or\ \hat{M} = \psi_m([O^F; M])). \tag{17}$$

Thereafter, the specific text-aware appearance (or motion) frames with the topN_α maximum similarities to global sentence-level semantics ($\bar{c}$) and local sentence-level semantics ($\bar{q}$ and $\bar{k}$) can be identified by Eqs. (1)-(8). Subsequently, we adaptively aggregate local-aware appearance (or motion) semantics into global-aware appearance (or motion) semantics through bottom-up cross-granularity aggregation Eqs. (9)-(14), and finally derive semantically rich appearance features $\tilde{F} = \{\tilde{f}_i\}_i^{N_\alpha}$ and motion features $\tilde{M} = \{\tilde{m}_i\}_i^{N_\alpha}$ for answers decoding Eqs. (15) and (16).

C. Answer Semantic Collaborated Decoding

In this subsection, we propose an Answer Semantic Collaborated Decoding (ASCD) module to collaboratively decode an implicitly predicted representation as output. Specifically, we first employ an adaptive gating mechanism to fuse semantically rich appearance and motion features into a unified video representation, which can be expressed as:

$$g_v = Sigmoid(\psi_{fm}([\tilde{F}; \bar{c}; \tilde{M}])), \tag{18}$$

$$V = g_v \odot \tilde{F} + (1 - g_v) \odot \tilde{M}, \tag{19}$$

where $\psi_{fm}(\cdot)$ is a FC layer. Then, we employ a Multi-Head Self-Attention (MHSA) [49] mechanism to capture the dynamic temporal relationship semantics between video frames V:

$$\tilde{V} = MHSA(V). \tag{20}$$

Similarly, we concatenate Q and K to further capture their contextual semantics $\tilde{C}$ through a multi-head attention mechanism similar to Eq. (20). Finally, we feed the video, question and knowledge semantics into a transformer block followed by a

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

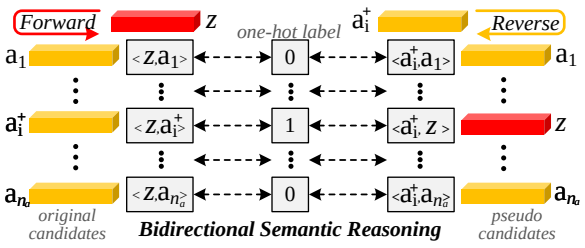


Fig. 4. The proposed BSR learning paradigm. It encourages the model to flexibly build the positive correlation from the predicted feature to the ground-truth answer while suppressing spurious correlations to all negative candidates without a prespecified margin via forward semantic reasoning (left). And it widens the semantic gap between the positive answer and all negative candidates to strengthen the unique positive semantic correlation from the positive answer to the predicted feature via reverse semantic reasoning (right).

mean pooling operation to collaboratively decode an implicitly predicted representation z:

$$z = Mean\_P(FFN(Cross\_Att(\tilde{V}, \tilde{C}, \tilde{C}) + \tilde{V})). \quad (21)$$

where $Cross\_Att(\cdot)$ represents a cross-attention module, $FFN(\cdot)$ represents a feed forward network and $Mean\_P(\cdot)$ is a mean pooling operation.

D. Candidate Learning with Adapters

In this subsection, we propose a Candidate Learning with Adapters (CLA) to learn the global semantic embedding of each candidate. Different from the previous methods, we freeze all parameters in the textual encoder, and only optimize lightweight adapters inserted in each transformer block. It not only significantly reduces computational overhead, but also effectively alleviates the negative impact caused by task gaps (upstream tasks are usually different from downstream tasks).

Specifically, we insert an adapter next to the multi-head attention and feed-forward layer in each transformer block to perform parallel computation. It is a bottleneck structure with a residual connection, consisting of a down-projection layer, a non-linear layer and an up-sampling layer. For the input T_i of the i-th layer, we have:

$$T_i^{ad} = \sigma(norm(T_i W_i^{dw}))W_i^{up}, \quad T_i^{ad} \leftarrow T_i^{ad} + T_i, \quad (22)$$

where $W_i^{dw} \in \mathbb{R}^{d \times d_b}$ and $W_i^{up} \in \mathbb{R}^{d_b \times d}$ represent the down- and up-projection matrix, respectively. d_b represents the bottleneck dimension, $norm(\cdot)$ represents the $Layernorm(\cdot)$ operation. Then, we add the output T_i^{ad} of the adapter to the output T_i^{mha} of the multi-head attention or T_i^{ff} of the feed-forward layer to perform subsequent operations. Finally, the last layer output cls_U is regarded as sentence-level candidate embedding a.

E. Bidirectional Semantic Reasoning

We propose a Bidirectional Semantic Reasoning (BSR) learning paradigm to drive model learning on both OE and MC tasks, which is shown in Fig. 4. Specifically, we perform forward semantic reasoning from the implicitly predicted feature z to all candidates $A = \{a_1, a_2, \cdots, a_{n_a}\}$ and obtain the similarity S_{z2A}, where n_a denotes the number of candidates. And then, we use the MSE function to calculate the forward reasoning loss between S_{z2A} and the ground-truth Y. Formally, it can be expressed as:

$$loss_{FSR} = MSE(S_{z2A}, Y), \ S_{z2A} = Cos(z, A^T), \quad (23)$$

where Y is a one-hot vector transformed from the ground-truth label, $Cos(\cdot,\cdot)$ represents the Cosine similarity between the two vectors. Additionally, the pseudo candidate list $\hat{A}$ can be easily obtained by replacing the positive answer feature a_i^+ in A with the predicted implicit feature z, and the reverse semantic reasoning loss can be expressed as:

$$loss_{RSR} = MSE(S_{a_i^+ 2\hat{A}}, Y), \quad (24)$$

$$S_{a_i^+ 2\hat{A}} = Cos(a_i^+, \hat{A}^T), \ \hat{A} = \{z \cup \{A/a_i^+\}\}. \quad (25)$$

By integrating Eq. (23) and Eq. (24) into a unified learning framework, the overall objective function can be formulated as:

$$\begin{aligned} loss = MSE(S_{z2A}, Y) + MSE(S_{a_i^+ 2\hat{A}}, Y) \\ + MSE(S_{z2A}, S_{a_i^+ 2\hat{A}}), \end{aligned} \quad (26)$$

where the last item in Eq. (26) is used to constrain the consistency of S_{z2A} and $S_{a_i^+ 2\hat{A}}$.

With the proposed BSR learning strategy, our method not only flexibly builds a positive correlation from the implicitly predicted feature to the ground-truth answer while suppressing spurious correlations to all negative candidates without a prespecified margin via forward semantic reasoning, but also explicitly widens the semantic gap between the positive answer and all negative candidates while reinforcing the unique positive correlation from the ground-truth answer to the implicitly predicted feature via reverse semantic reasoning. It explicitly alleviates the negative impact caused by overlapping semantic co-occurrences that exist between question and negative candidates. Additionally, benefiting from a consistent learning strategy with shared model parameters on both OE and MC tasks, our method can achieve seamless transfer between these two reasoning tasks.

IV. Experiments

In this section, we conduct comparison experiments with state-of-the-art baselines based on common VideoQA datasets to verify the superior performance of our method, and design the ablation experiments to validate the effects of each learning module.

A. Evaluation Datasets

MSVD-QA [30] contains 1,970 videos with 50,505 question-answer pairs. The average video length is 10 seconds and the average question length is 6 words. It is a representative VideoQA dataset for **Open-Ended (OE)** tasks, which contains 5 question types: *what*, *who*, *how*, *when* and *where*. To keep the consistence with the experimental settings of the previous methods, we use 61% of the VQ pairs as the training

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

TABLE II
BASIC STATISTICS OF THE EXPERIMENTAL DATASETS. OE MEANS THAT THE CURRENT DATASET IS AN OPEN-ENDED TASK DATASET, AND MC MEANS THAT IT IS A MULTI-CHOICE TASK DATASET. (CANDIDATES) INDICATES THE NUMBER OF CANDIDATES.

Datsets	Reasoning Challenge	Video Duration	Train Pairs	Validation Pairs	Test Pairs	Task(Candidates)
MSVD-QA [30]	Visual Recognition	10s	30,933	6,415	13,157	OE(1,852)
MSRVTT-QA [30]	Visual Recognition	15s	158,581	12,278	72,821	OE(4,000)
NExT-QA [50]	Causal & Temporal Reasoning	44s	34,132	4,996	8,564	MC(5)
Causal-VidQA [51]	Evidence & Commonsense Reasoning	9s	112,656	16,170	32,574	MC(5)
NExT-OOD [52]	Out-Of-Distribution Setting (VA, QA, QA&VA)	44s	34,132	19,320/848/5,810	25,870/1,373/8,010	MC(5)
ActivityNet-QA [53]	Long-form Spatial-Temporal Action Reasoning	180s	32,000	18,000	8,000	OE(1,654)
EgoSchema [54]	Long-form Egocentric Zero-Shot Reasoning	180s	0	0	5,031	MC(5)

set, 13% and 26% of the pairs for the validation set and test set, respectively.

MSRVTT-QA [30] is a larger VideoQA dataset than MSVD-QA with more complex video scenarios, which consists of 10,000 videos with 243,680 question-answer pairs for **Open-Ended (OE)** task. The average video length is 15 seconds and the average question length is 7 words. We use 65% of the VQ pairs as the training set, 5% and 30% of the pairs for the validation set and test set, respectively.

NExT-QA [50] is a representative dataset for **Multi-Choice (MC)** tasks, which mainly focuses on causal and temporal reasoning. It contains 5,440 videos and 47,692 questions, and each question is followed by 5 candidates. The average video length is 44 seconds, and the average question and candidate lengths are 12 words and 3 words, respectively. Obeying the official division, we set 34,132 VQ pairs as the training set, 4,996 and 8,564 pairs as the validation set and test set, respectively.

Causal-VidQA [51] is a large-scale **Multi-Choice (MC)** task dataset that highlights evidence reasoning and commonsense reasoning. It contains 26,900 videos, and each video contains 6 question-candidate pairs (161,400 in total). The average video length is 9 seconds, and the average question and candidate length are both 10 words. With the same experimental setting as the previous methods, we set 112,656 VQ pairs as the training set, 16,170 and 32,574 pairs as the validation set and test set, respectively.

NExT-OOD [52] tends to verify the generalization ability of the VideoQA model under Out-Of-Distribution setting, which is a challenging **Multi-Choice (MC)** task dataset. It contains three sub-datasets of NExT-OOD-VA, NExT-OOD-QA and NExT-OOD-VQA for comprehensive analysis of VA bias, QA bias and VA&QA bias, respectively. With the same experimental setting as the previous method [52], we perform experiments on these three sub-datasets to test the generalization performance of the proposed method.

ActivityNet-QA [53] is a long-form dataset for evaluating VideoQA models reasoning about complex spatial-temporal actions. It is a challenging dataset for **Open-Ended (OE)** tasks, which contains a total of 58,000 long-form action videos. Following the official division, we set 32,000 VQ pairs as the training set, 18,000 and 8,000 pairs as the validation set and test set, respectively.

EgoSchema [54] proposes a long-form, egocentric diverse dataset for performing zero-shot evaluation. It is a challenging dataset for **Multi-Choice (MC)** tasks, and all video-question pairs are considered as testing samples. Therefore, we first perform pre-training on the NExT-QA training set, and then verify the zero-shot reasoning performance on EgoSchema.

For convenience, the basic statistics of the seven experimental datasets are listed in Table II.

B. Implementation Details

Following the existing methods [17], [59], we split each video into N_f = 32 frames on the seven datasets, and then utilize the pre-trained 3D ResNeXt-101 [43], [44] and ResNet-101 [42] as backbones to extract motion features and appearance features of videos, respectively. Additionally, we use Faster R-CNN [45] to detect the 20 regions with the highest confidence in each frame on NExT-QA and NExT-OOD, and 10 regions on other datasets. The hidden dimension of the proposed model is set as d=768. We set α to 0.5 on MSVD-QA and MSRVTT-QA, and 0.6 on other datasets. As for the text modality, we utilize the Bert [48] model to encode knowledge, question and candidate answers into their corresponding text embeddings. The bottleneck dimension is set as d_b=256. At the stage of training, we fix the batchsize as 64, and consistently set the learning rate as 10^{-5} with the AdamW optimizer on all experimental datasets. All the experiments are performed on a workstation with one NVIDIA RTX 3090 GPU.

C. Comparison with State-of-the-Art Methods

We carefully perform the experiments of the proposed method on the seven VideoQA datasets, and list the accuracy comparison results with the SOTAs in Tables III, IV, V and VI, respectively. Based on comparison results, we arrive at the following analyses:

1) Results on MSVD-QA and MSRVTT-QA: As shown in Table III, the proposed method consistently outperforms SOTAs on two Open-Ended datasets. Specifically, our method outperforms the second-best method by 5.9% and 2.1% on MSVD-QA and MSRVTT-QA, respectively. The superior performance of our method can be attributed to the following: (1) Our method collaboratively reasons and decodes complex intrinsic correlations between more question-aware and knowledge-aware positive video scenes than between individually question-focused video scenes. It effectively expands the positive clues for the ground-truth answer than question-oriented methods, thereby achieving better reasoning performance. (2) The proposed BSR learning strategy encourages the learned implicit feature to be semantically close to the

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

TABLE III
COMPARISONS WITH STATE-OF-THE-ART BASELINES ON THREE COMMON MSVD-QA, MSRVTT-QA AND NEXT-QA DATASETS. ACC@C, @T, AND @D REPRESENT THE SUB-ACCURACY FOR CAUSALITY, TEMPORALITY, AND DESCRIPTIVE QUESTIONS, RESPECTIVELY. THE BEST RESULTS IN EACH COLUMN ARE MARKED WITH **BOLD** AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE BELOW IS THE SAME.

Methods	MSVD-QA	MSRVTT-QA	NExT-QA			
			ACC@ALL	ACC@C	ACC@T	ACC@D
Co-Mem [55]	34.6	35.3	48.5	45.9	50	54.4
HME [12]	33.7	33.0	49.2	–	–	–
HGA [32]	34.7	35.5	50.0	48.1	49.1	57.8
HCRN [14]	36.1	35.6	48.8	47.1	49.3	54
DualVGR [31]	39.0	35.5	–	–	–	–
B2A [13]	37.2	36.9	49.6	47.4	49.0	58.3
MSPAN [56]	40.3	37.8	50.9	48.6	49.8	60.4
IGV [15]	40.8	38.3	51.3	48.6	51.7	59.6
ERM [57]	38.4	37.1	–	–	–	–
MGIN [58]	39.7	38.2	–	–	–	–
HQGA [17]	41.2	38.6	51.8	49.0	52.3	59.4
VGT [59]	–	39.7	53.7	51.6	51.9	63.7
MCR [40]	–	–	52.4	49.2	52.0	62.3
VCSR [18]	–	38.9	54.1	53.0	51.5	62.3
DRV [60]	42.2	40.0	55.8	54.1	54.8	63.2
KPI [41]	43.3	40.0	55.0	–	–	–
Ours	**49.2**	**42.1**	**60.3**	**59.3**	**58.0**	**67.9**

ground-truth answer rather than being forced to classify into discrete class label, which facilitates our model to learn a more robust predicted representation and achieves higher reasoning accuracy on the test set.

2) Results on NExT-QA: According to the comparison results reported in Table III, we can clearly find that the proposed method achieves consistent performance improvements on both total accuracy (ALL: 4.5%) and all sub-accuracies (C: 5.2%, T: 3.2% and D: 4.2%). The analysis can be summarized as follows: (1) We adopt an off-the-shelf VLM to generate coarse question-related knowledge, which facilitates the model to deeply understand the complex video content and enhances the collaborative mining of critical video scenes and reasoning of complex intrinsic correlations through the proposed CACR module. (2) The proposed BSR learning strategy encourages the model to flexibly build a unique positive semantic correlation between the predicted feature and the ground-truth answer, while alleviating spurious correlations by widening the semantic gap between all negative candidates and both the positive answer and predicted feature via forward and reverse semantic reasoning. Benefiting from this, our method can effectively alleviate the negative impact caused by overlapping semantic co-occurrences and learn discriminative semantic correlations to achieve superior reasoning performance.

3) Results on Causal-VidQA: By observing the experimental results reported in Table IV, we can clearly find that the proposed method achieves significant performance improvements over existing competitors. Specifically, the reasoning accuracy of our method is 3.1% higher than that of the second-best method, which mainly benefits from the significant performance improvement in the sub-tasks of evidence reasoning (E: +1.4%) and commonsense reasoning (P: +6.0% and C: +2.4%). For the commonsense reasoning sub-task, we introduce a Δ to evaluate the consistency of model reasoning by calculating the numerical difference between Q→A and Q→AR. It reveals the performance of the model in reasoning the answer correctly but the rationale incorrectly, that is, inconsistency reasoning. Therefore, a lower $\Delta\downarrow$ reflects better reasoning consistency of the VideoQA model. The comparison results demonstrate that the proposed method can not only reason the correct answer (Q→A) more accurately, but also feedback a more reasonable rationale (Q→AR) to explain why the inferred answer is correct, which proves that the proposed method has better consistent reasoning ability.

4) Results on NExT-OOD: As reported in Table V, the proposed method significantly exceeds the existing competitors on three test subsets under out-of-distribution setting. Specifically, the proposed method achieves 6.1%, 11.5% and 6.9% performance improvements over the second-best method on the NExT-OOD-QA, NExT-OOD-VA and NExT-OOD-VQA datasets, respectively. These comparison results verify that our method achieves more robust reasoning performance under the VA (long-tail) bias and QA (overlapping words) bias settings, which can be attributed to the following: (1) Benefiting from additional knowledge semantic and collaborative aware reasoning strategy, our method strives to progressively reasoning the real logical interaction between critical video scenes, which effectively alleviates the negative impact of VA long-tail bias. (2) The proposed BSR learning strategy explicitly widens the semantic gap between all negative candidates and both the ground-truth answer and the implicitly predicted feature to alleviate the challenges caused by overlapping semantic co-occurrences. Therefore, our model is encouraged to learn the robust inherent correlation between the implicit feature and the ground-truth answer, achieving higher performance in QA bias setting.

5) Results on ActivityNet-QA: Our method achieves a significant performance improvement according to the comparison results in Table VI. Specifically, our method achieves leads of 18.2%, 9.1%, 63.2%, and 3.9% on the subtasks *Motion*, *Spatial*, *Temporal*, and *Free*, respectively, and improves the overall reasoning performance by 11.1%. These comparison results verify that the proposed method can effectively reason about complex spatiotemporal relationships of diverse actions

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART BASELINES ON THE COMMON CAUSAL-VIDQA DATASETS. ACC@D, @E, @P AND @C REPRESENT THE SUB-ACCURACY FOR DESCRIPTION, EXPLANATION, PREDICTION, AND COUNTERFACTUAL QUESTIONS, RESPECTIVELY. Q→A, Q→R AND Q→AR REPRESENT THE SUB-ACCURACY OF ANSWERING THE ANSWER, THE REASON, AND BOTH OF THEM, RESPECTIVELY. Δ MEANS Q→A MINUS Q→AR.

Methods	ACC@D	ACC@E	ACC@P				ACC@C				mean$\Delta\downarrow$	ACC@ALL
			Q→A	Q→R	Q→AR	$\Delta\downarrow$	Q→A	Q→R	Q→AR	$\Delta\downarrow$		
Co-Mem [55]	60.1	62.8	51.0	50.4	31.4	19.6	51.6	53.1	32.6	19.0	19.30	47.7
HME [12]	63.4	61.5	50.3	47.6	28.9	21.4	50.4	51.7	30.9	19.5	20.45	46.2
HGA [32]	65.7	63.5	49.4	50.6	32.2	17.2	52.4	55.9	34.3	**18.1**	**17.65**	48.9
HCRN [14]	65.4	61.6	51.7	51.3	32.6	19.1	51.6	53.4	32.7	18.9	19.00	48.1
B2A [13]	66.2	62.9	49.0	50.2	31.2	17.8	53.3	56.3	35.2	**18.1**	17.95	49.1
IGV [15]	65.9	62.1	52.8	53.5	35.0	17.8	50.7	52.3	31.2	19.5	18.65	48.6
MCR [40]	67.5	65.6	56.5	56.4	37.8	18.7	52.4	54.1	33.4	19.0	18.85	51.1
VSCR [18]	66.0	65.4	60.9	58.5	41.2	19.7	53.4	54.4	34.1	19.3	19.50	51.7
VGT [59]	70.8	70.3	55.2	56.9	38.4	**16.8**	61	59.3	42	19.0	17.90	55.4
KPI [41]	73.0	72.4	58.2	57.1	38.8	19.4	61.7	61.1	42.6	19.1	19.25	56.7
Ours	**73.1**	**73.8**	**64.0**	**62.8**	**47.2**	**16.8**	**63.5**	**61.5**	**45.0**	18.5	**17.65**	**59.8**

TABLE V
COMPARISONS WITH STATE-OF-THE-ART BASELINES ON NEXT-OOD. QA, VA AND VQA REPRESENT THE SUBSETS FOR VERIFYING AND ANALYZING QA BIAS, VA BIAS AND VQA BIAS, RESPECTIVELY.

Methods	QA	VA				VQA			
		N1	N2	N5	Avg.	N1	N2	N5	Avg.
PSAC [61]	27.2	32.8	33.8	34.2	33.6	23.9	25.2	24.7	24.6
STVQA [62]	30.0	34.3	35.2	36.8	35.4	30.4	29.2	29.5	29.7
HCRN [14]	31.1	35.0	36.5	39.1	36.9	30.0	30.9	31.1	30.6
HGA [32]	33.0	37.1	37.7	39.5	38.1	34.6	32.3	31.3	32.7
HQGA [17]	33.0	35.5	39.3	40.7	38.5	36.0	31.7	31.4	33.1
GCS [52]	36.6	39.6	41.5	45.5	42.3	35.8	32.7	34.5	34.3
Ours	**42.7**	**52.1**	**52.8**	**56.6**	**53.8**	**42.4**	**39.4**	**41.9**	**41.2**

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART BASELINE ON TWO LONG-FORM VIDEOQA DATASETS, ACTIVITYNET-QA AND EGOSCHEMA.

Methods	ActivityNet-QA					EgoSchema	
	Motion	Spatial	Temporal	Free	ALL	Sub	Total
VGT [59]	20.4	15.5	3.4	64.1	36.1	25.0	25.4
Ours	**38.6**	**24.6**	**66.6**	**68.0**	**47.2**	**39.8**	**36.7**

even on long-form video reasoning tasks.

6) Results on EgoSchema: In Table VI, the proposed method achieves 14.8% and 11.3% higher performance than VGT on the *Sub* set and *Total* set, respectively. The superior performance improvement can be attributed to the fact that the proposed method can effectively mine and model the complex correlation semantics between critical scenarios to reason about the correct answer, which alleviates the interference of spurious correlations on model reasoning and achieves better zero-shot performance.

D. Ablation Studies

In this subsection, we perform ablation studies to verify the effect of each learning module in the proposed method. Taking MSVD and NExT-QA as examples, we list all ablation experimental results in Table VII. Specifically, "w/o $*$" means removing the "$*$" module from our proposed method. We summarize the ablation analysis as follows:

1) The variant (*w/o CPM*) suffers from suboptimal reasoning performance compared to our method, which demonstrates its effectiveness in mining critical video scenes and discarding noisy scenes for reasoning. Intuitively, we plot the Acc-α curve in Fig. 5(a), and have the following analysis: when α is too small, a large number of critical video scenes are regarded as noisy scenes to be discarded, resulting in insufficient intrinsic semantic reasoning; when α is too large, more noisy video scenes are regarded as critical scenes, which inevitably introduces more detrimental information into subsequent reasoning and decoding, and thus degrades the reasoning performance. Additionally, we remove the CAA module by concatenating global and local text features and feeding them into a vanilla GCN module. We can find that the variant (*w/o CAA*) falls into significant performance degradation compared to our method, which verifies that the proposed CAA module can indeed effectively enhance complex intrinsic correlation reasoning between critical video scenes to improve the model's performance. Further, we design two variants (*w/o G-aware* and *w/o L-aware*) to verify the effectiveness of local text and global text in guiding video reasoning. The experimental results demonstrate that both are indispensable for achieving superior reasoning performance. Finally, the variant (*w/o CACR*) suffers from severe performance degradation due to the lack of collaborative mining of critical video scenes and reasoning of complex intrinsic correlations, which verifies the

TABLE VII
ABLATION RESULTS ON MSVD-QA AND NEXT-QA.

Methods	MSVD-QA	NExT-QA			
		ACC-ALL	ACC-C	ACC-T	ACC-D
w/o CACR	40.1	54.7	52.9	52.7	64.1
w/o CPM	48.4	60.0	59.1	56.9	68.6
w/o CAA	45.3	58.7	58.2	55.3	66.6
w/o G-aware	45.2	57.8	57.2	54.7	65.6
w/o L-aware	47.9	59.2	58.5	55.8	67.6
w/o ASCD	44.9	56.7	56.7	54.2	67.1
w/o BSR	47.5	58.8	57.3	57.0	67.2
w/o Adapter	48.9	60.6	59.6	58.5	67.9
w/o Ques.	23.1	48.3	46.9	43.9	61.4
w/o Know.	41.2	51.3	48.8	49.5	62.5
w/o Video	41.7	54.5	53.6	50.6	65.1
mPLUG-Owl*	–	23.0	23.6	20.9	25.2
Ours	49.2	60.3	59.3	58.0	67.9
Ours+	**51.1**	**61.0**	**59.8**	**58.9**	**69.1**

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

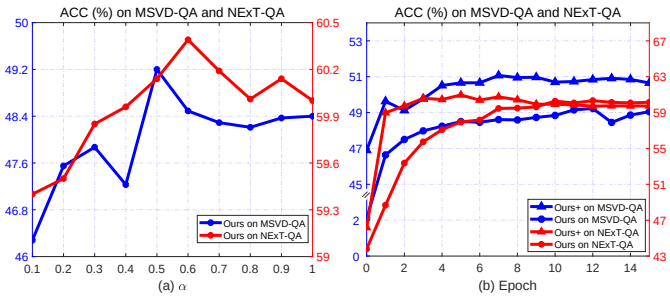


Fig. 5. The ACC-α curve (a) and ACC-Epoch curve (b).

effectiveness of the proposed CACR module.

2) We design a variant (*w/o ASCD*) that replaces the proposed ASCD module with a two-layer MLP to decode video and question semantics into an implicit representation. The experimental results verify the effectiveness of our adaptive gated fusion and collaborative semantic decoding strategy. Additionally, we design a variant (*w/o BSR*) by replacing the proposed BSR with the common classification strategy on OE task or matching strategy on MC task with different task-specific decoders adopted by previous methods [14], [17]. From Table VII, our method consistently outperforms the variant (*w/o BSR*) on both OE and MC tasks, verifying that the proposed BSR learning strategy can indeed learn discriminative semantic correlations to achieve higher reasoning performance. Finally, we remove the additional introduced adapter and optimize all network parameters in the text encoder. The comparison results indicate that our method achieves competitive performance compared to the variant (*w/o Adapter*) on both OE and MC tasks, which verifies that the introduced adapter can effectively alleviate the task gap with less computational overhead (9M vs 110M).

3) To verify the effect of different modalities on reasoning performance, we design several variants and report their performance in Table VII. Despite being equipped with generated knowledge semantics, the variant (*w/o Ques.*) in which the question modality is removed still suffers from severe performance degradation, especially on MSVD-QA. This is reasonable as it is indeed logically challenging for the model to reason satisfactory answers in the absence of question semantics. Additionally, after removing the generated knowledge semantics, the variant (*w/o Know.*) degenerates into a question-oriented video reasoning model, and falls into significant performance degradation. The comparison results prove that the generated coarse knowledge can effectively improve the complex intrinsic correlation reasoning between critical video scenes in the CACR model and enhance the decoding of positive semantics in the ASCD module to achieve better reasoning performance. Ultimately, the comparison result (*w/o Video vs Ours*) verifies that the rich semantic information in videos cannot be entirely replaced by shallow knowledge semantics, and remains one of the core evidences to achieve higher reasoning performance.

4) We design a variant (*mPLUG-Owl**) to verify the reasoning performance of the adopted mPLUG-Owl module on NExT-QA by prompting it to reason a prediction directly. Specifically, we reorganize the prompt into ***"Please reason the correct answer from the following candidates based on the given video and*** $\{Q\}$***. Candidates: [0]***$\{A_0\}$ ***[1]***$\{A_1\}$ ***[2]***$\{A_2\}$ ***[3]***$\{A_3\}$ ***[4]***$\{A_4\}$***."***, and input it into mPLUG-Owl along with each video-question-candidates pair. $\{Q\}$ and $\{A_i\}$ represent the placeholder of the question and the i-th candidate ($i \in$ [0,1,2,3,4]), respectively. According to the results in Table VII, it is clear that the variant exhibits significantly unsatisfactory reasoning performance. This is because mPLUG-Owl mainly utilizes existing image-caption pairs to enhance the multi-modal capabilities of large language models by aligning image and text semantics. Despite performing extensive pre-training on massive image-text pairs, it scarcely takes complex causal or temporal reasoning tasks into account, which significantly limits its multi-modal reasoning performance on video question answering tasks.

5) To verify that our method can achieve seamless task transfer, we designed a variant (*Ours+*), which successively trains on the MC (OE) task based on the pre-trained model on the OE (MC) task instead of training from scratch. The experimental results demonstrate that seamless transfer can further advance our method to achieve better reasoning performance on both OE and MC tasks. Further, we plot the ACC-Epoch curve in Fig. 5(b), and summarize the analysis as follows: (1) When the training epoch is 0, the variant method achieves better reasoning performance (*Ours+ vs Ours*), especially on MSVD-QA, which demonstrates that it can effectively transfer the learned reasoning ability between different reasoning tasks. (2) The variant method achieves faster convergence and better accuracy on both datasets with fewer training epochs, which demonstrates that seamless model transfer indeed benefits our method to achieve better performance on new reasoning tasks.

E. Qualitative Analysis

We visualize several cases to study the reasoning evidence of the proposed method and its variants, including three correct reasoning cases and one incorrect reasoning case. In the first video, our method not only captures the question-aware 2nd, 3rd and 4th frames but also the knowledge-aware 5th and 6th frames, and successfully reasons the causal correlation between "boy in blue run" and "flying kite" with the highest confidence. In contrast, the variant (*w/o CACR*) focuses more on question-oriented video scenes and unfortunately falls into spurious correlations between "boy" and "children". In the second video, compared to the variant (*w/o ASCD*), our method correctly reasons the object state ("running" in A_2 instead of "stopped" in A_5) by adaptively fusing static appearance and dynamic motion features, and then collaboratively decodes the positive representation semantically related to the correct answer (0.26 *vs* 0.19) together with the coarse knowledge semantics. For the third video, by observing the comparison results reported with the variant, our method can effectively widen the semantic gap between all negative candidates and both the ground-truth answer and the implicitly predicted feature to alleviate the impact of overlapping semantic co-occurrences, and strengthens the semantic correlation (0.27 *vs* 0.04) between the predicted feature and the ground-truth

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

Fig. 6. Qualitative analysis on NExT-QA. The critical video scenes are framed, and the ground-truth answer to each question is colored green. ◯, ◯, ◯ and ◯ represent the prediction of our method and variants *w/o CACR*, *w/o ASCD*, and *w/o BSR*, respectively. "—" or "—" underline the repeated words between the negative candidates and the question or between the negative candidates and the ground-truth answer. In addition, the reasoning correlations of our method and variants are quantified with *Softmax(·)* regularization and shown on the right.

answer. Finally, our method derives an incorrect reasoning result in the last video. The reason may be attributed to our model focusing excessively on "colorful" appearance information and failing to effectively capture local fine-grained motion information ("break to pull"), ultimately leading to spurious correlation reasoning between "walk" and "move".

V. Conclusion

In this paper, we propose a novel collaborative aware bidirectional semantic reasoning framework for VideoQA. With the proposed CACR module, we collaboratively mine multi-granularity text-aware video scenes to reason the complex intrinsic correlations among them via bottom-up cross-granularity adaptive aggregation, and finally derive a set of semantically rich critical video features. Then, we adopt a unified ASCD module to collaboratively decode the video, question and knowledge semantics into an implicit representation. Finally, a novel BSR learning strategy is proposed to bridge and reinforce unique semantic correlations between the learned predicted representation and the ground-truth answer, and to explicitly alleviate the detrimental impact caused by overlapping semantic co-occurrences. The proposed method shares the same model structure and learning strategy on both Open-Ended and Multi-Choice tasks, which enables it to be seamlessly transferred between these two reasoning tasks. Extensive experimental results on seven benchmarks demonstrate the superiority of the proposed method.

References

[1] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4631–4640.

[2] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, and T.-S. Chua, "Video question answering: Datasets, algorithms and challenges," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 6439–6455.

ACCEPTED MANUSCRIPT / CLEAN COPY

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

[3] Z. Chen, L. Wang, P. Wang, and P. Gao, "Question-aware global-local video understanding network for audio-visual question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 4109–4119, 2023.

[4] K. Guo, D. Tian, Y. Hu, C. Lin, Z. Qian, Y. Sun, J. Zhou, X. Duan, J. Gao, and B. Yin, "Cfmmc-align: Coarse-fine multi-modal contrastive alignment network for traffic event video question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[6] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[7] Z. Wang, X. Xu, J. Wei, N. Xie, Y. Yang, and H. T. Shen, "Semantics disentangling for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 33, pp. 2226–2237, 2024.

[8] Z. Guo, Z. Zhao, W. Jin, Z. Wei, M. Yang, N. Wang, and N. J. Yuan, "Multi-turn video question generation via reinforced multi-choice attention network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1697–1710, 2020.

[9] Z. Wang, Z. Gao, Y. Yang, G. Wang, C. Jiao, and H. T. Shen, "Geometric matching for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.

[10] L. Zhu, X. Wu, J. Li, Z. Zhang, W. Guan, and H. T. Shen, "Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8838–8851, 2022.

[11] Z. Wang, Z. Gao, M. Han, Y. Yang, and H. T. Shen, "Estimating the semantics via sector embedding for image-text retrieval," *IEEE Transactions on Multimedia*, pp. 1–12, 2024.

[12] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1999–2007.

[13] J. Park, J. Lee, and K. Sohn, "Bridge to answer: Structure-aware graph interaction network for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 526–15 535.

[14] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9972–9981.

[15] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Invariant grounding for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2928–2937.

[16] L. Peng, S. Yang, Y. Bin, and G. Wang, "Progressive graph attention network for video question answering," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 2871–2879.

[17] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T.-S. Chua, "Video as conditional graph hierarchy for multi-granular question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2804–2812.

[18] Y. Wei, Y. Liu, H. Yan, G. Li, and L. Lin, "Visual causal scene refinement for video question answering," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 377–386.

[19] J. Zhang, J. Shao, R. Cao, L. Gao, X. Xu, and H. T. Shen, "Action-centric relation transformer network for video question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 63–74, 2020.

[20] T. Yu, J. Yu, Z. Yu, Q. Huang, and Q. Tian, "Long-term video question answering via multimodal hierarchical memory attentive networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 931–944, 2020.

[21] M. Liu, F. Zhang, X. Luo, F. Liu, Y. Wei, and L. Nie, "Advancing video question answering with a multi-modal and multi-layer question enhancement network," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 3985–3993.

[22] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 800–10 809.

[23] L. Chen, Y. Zheng, Y. Niu, H. Zhang, and J. Xiao, "Counterfactual samples synthesizing and training for robust visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[24] Z. Wen, G. Xu, M. Tan, Q. Wu, and Q. Wu, "Debiased visual question answering from feature and sample perspectives," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 3784–3796.

[25] Q. Si, Z. Lin, M. yu Zheng, P. Fu, and W. Wang, "Check it again: Progressive visual question answering via visual entailment," in *Proceedings of the International Joint Conference on Natural Language Processing*, 2021, pp. 4101–4110.

[26] Z. Wen, S. Niu, G. Li, Q. Wu, M. Tan, and Q. Wu, "Test-time model adaptation for visual question answering with debiased self-supervisions," *IEEE Transactions on Multimedia*, vol. 26, pp. 2137–2147, 2023.

[27] S. Na, S. Lee, J. Kim, and G. Kim, "A read-write memory network for movie story understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 677–685.

[28] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo, "Progressive attention memory network for movie story question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8337–8346.

[29] X. Nie, S. Xu, X. Liu, G. Meng, C. Huo, and S. Xiang, "Bilateral memory consolidation for continual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 026–16 035.

[30] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the ACM International Conference on Multimedia*, 2017, pp. 1645–1653.

[31] J. Wang, B.-K. Bao, and C. Xu, "Dualvgr: A dual-visual graph reasoning unit for video question answering," *IEEE Transactions on Multimedia*, vol. 24, pp. 3369–3380, 2021.

[32] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 109–11 116.

[33] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 021–11 028.

[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the International Conference on Learning Representations*, 2016.

[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proceedings of the International Conference on Learning Representations*, 2018.

[36] F. Liu, J. Liu, W. Wang, and H. Lu, "Hair: Hierarchical visual-semantic relational reasoning for video question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1698–1707.

[37] L. H. Dang, T. M. Le, V. Le, and T. Tran, "Hierarchical object-oriented spatio-temporal reasoning for video question answering," *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021.

[38] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," in *Proceedings of the International Conference on Learning Representations*, 2019.

[39] X. Yang, H. Zhang, G. Qi, and J. Cai, "Causal attention for vision-language tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9847–9857.

[40] C. Zang, H. Wang, M. Pei, and W. Liang, "Discovering the real association: Multimodal causal reasoning in video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 027–19 036.

[41] J. Li, L. Niu, and L. Zhang, "Knowledge proxy intervention for deconfounded video question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 2782–2793.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

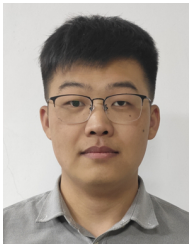
[43] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[44] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

ACCEPTED MANUSCRIPT / CLEAN COPY

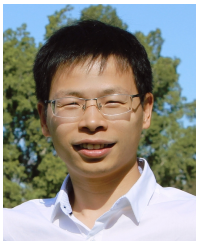
This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3490665

[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 28, 2015.
[46] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Jiang, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qian, J. Zhang, and F. Huang, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.
[47] Q. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," *arXiv preprint arXiv:2311.04257*, 2023.
[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.
[50] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9777–9786.
[51] J. Li, L. Niu, and L. Zhang, "From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 273–21 282.
[52] X. Zhang, F. Zhang, and C. Xu, "Next-ood: Overcoming dual multiple-choice vqa biases," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 1913–1931, 2023.
[53] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.
[54] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, 2024.
[55] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6576–6585.
[56] Z. Guo, J. Zhao, L. Jiao, X. Liu, and L. Li, "Multi-scale progressive attention network for video question answering," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 973–978.
[57] F. Zhang, R. Wang, F. Zhou, and Y. Luo, "Erm: Energy-based refined-attention mechanism for video question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1454–1467, 2022.
[58] Y. Wang, M. Liu, J. Wu, and L. Nie, "Multi-granularity interaction and integration network for video question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7684–7695, 2023.
[59] J. Xiao, P. Zhou, T.-S. Chua, and S. Yan, "Video graph transformer for video question answering," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 39–58.
[60] J. Liu, G. Wang, J. Xie, F. Zhou, and H. Xu, "Video question answering with semantic disentanglement and reasoning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3663–3673, 2023.
[61] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8658–8665.
[62] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, and G. Kim, "Video question answering with spatio-temporal reasoning," *International Journal of Computer Vision*, vol. 127, pp. 1385–1412, 2019.



Xize Wu is currently a Ph. D candidate at the School of Computer Science and Engineering, Southeast University, China. His research interests include multimedia content analysis, cross-modal retrieval and visual reasoning.

Jiasong Wu (Member, IEEE) received the B.S. degree in Biomedical Engineering from the University of South China, Hengyang, China, in 2005, and joint Ph.D. degree with the Laboratory of Image Science and Technology (LIST), Southeast University, Nanjing, China, and Laboratoire Traitement du signal et de l'Image (LTSI), University of Rennes 1, Rennes, France in 2012. He is now working in the LIST as an associate professor. His-research interest mainly includes deep learning, fast algorithms of digital signal processing and its applications. Dr. Wu received the Eiffel doctorate scholarship of excellence (2009) from the French Ministry of Foreign Affairs and also the Chinese government award for outstanding self-financed student abroad (2010) from the China Scholarship Council.



Lei Zhu (Senior Member, IEEE) received his B.Eng. and Ph.D. degrees from Wuhan University of Technology in 2009 and Huazhong University Science and Technology in 2015, respectively. He was a Research Fellow at the University of Queensland (2016-2017). His research interests are in the area of large-scale multimedia content analysis and retrieval. Zhu has co-/authored more than 100 peer-reviewed papers, such as ACM SIGIR, ACM MM, IEEE TPAMI, IEEE TIP, IEEE TKDE, and ACM TOIS. His publications have attracted more than 4,400 Google citations. At present, he serves as the Associate Editor of IEEE Transactions on Big Data and Information Sciences. He has served as the Area Chair, Senior Program Committee or reviewer for more than 40 well-known international journals and conferences. He won ACM SIGIR 2019 Best Paper Honorable Mention Award, ADMA 2020 Best Paper Award, ChinaMM 2022 Best Student Paper Award, ACM China SIGMM Rising Star Award, Shandong Provincial Entrepreneurship Award for Returned Students, and Shandong Provincial AI Outstanding Youth Award.

Lotfi Senhadji (Senior Member, IEEE) received the Ph.D. degree from the University of Rennes 1, Rennes, France, in signal processing and telecommunications in 1993. He is a Professor and the Head of the INSERM Research Laboratory LTSI. His is also Co-Director of the French-Chinese Laboratory CRIBs "Centre de Recherche en Information Biomédicale Sino-Français". His main research efforts are focused on nonstationary signal processing with particular emphasis on wavelet transforms and timefrequency representations for detection, classification, and interpretation of biosignals. He has published more than 80 research papers in journals and conferences, and he contributed to five handbooks. Dr. Senhadji is a senior member of the IEEE EMBS and the IEEE Signal Processing Society.

Huazhong Shu (Senior Member, IEEE) received the B.S. degree in applied mathematics from Wuhan University, China, in 1987, and the Ph.D. degree in numerical analysis from the University of Rennes 1, Rennes, France, in 1992. He is currently a Professor with the LIST Laboratory and also the Co-Director of CRIBs. His recent work concentrates on the image analysis, pattern recognition, and fast algorithms of digital signal processing. He is a Senior Member of the IEEE Society.

ACCEPTED MANUSCRIPT / CLEAN COPY