Multi-Granularity Contrastive Cross-Modal Collaborative Generation for End-to-End Long-Term Video Question Answering

Ting Yu, Kunhao Fu, Jian Zhang, Qingming Huang, *Fellow, IEEE*, and Jun Yu, *Senior Member, IEEE*

***Abstract*— Long-term Video Question Answering (VideoQA) is a challenging vision-and-language bridging task focusing on semantic understanding of untrimmed long-term videos and diverse free-form questions, simultaneously emphasizing comprehensive cross-modal reasoning to yield precise answers. The canonical approaches often rely on off-the-shelf feature extractors to detour the expensive computation overhead, but often result in domain-independent modality-unrelated representations. Furthermore, the inherent gradient blocking between unimodal comprehension and cross-modal interaction hinders reliable answer generation. In contrast, recent emerging successful video-language pre-training models enable cost-effective end-to-end modeling but fall short in domain-specific ratiocination and exhibit disparities in task formulation. Toward this end, we present an entirely end-to-end solution for long-term VideoQA: Multi-granularity Contrastive cross-modal collaborative Generation (MCG) model. To derive discriminative representations possessing high visual concepts, we introduce Joint Unimodal Modeling (JUM) on a clip-bone architecture and leverage Multi-granularity Contrastive Learning (MCL) to harness the intrinsically or explicitly exhibited semantic correspondences. To alleviate the task formulation discrepancy problem, we propose a Cross-modal Collaborative Generation (CCG) module to reformulate VideoQA as a generative task instead of the conventional classification scheme, empowering the model with the capability for cross-modal high-semantic fusion and generation so as to rationalize and answer. Extensive experiments conducted on six publicly available VideoQA datasets underscore the superiority of our proposed method.**

***Index Terms*— Video question answering, multi-granularity, contrastive learning, cross-modal collaborative generation, end-to-end modeling.**

Manuscript received 9 March 2023; revised 10 September 2023, 20 December 2023, and 13 February 2024; accepted 11 April 2024. Date of publication 24 April 2024; date of current version 29 April 2024. This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LY23F020005; and in part by the National Natural Science Foundation of China under Grant 62125201, Grant 62002314, and Grant 62020106007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chuang Gan. *(Corresponding author: Jun Yu.)*

Ting Yu, Kunhao Fu, and Jian Zhang are with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: yut@hznu.edu.cn; fukunhao@stu.hznu.edu.cn; jeyzhang@outlook.com).

Qingming Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qmhuang@ucas.ac.cn).

Jun Yu is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China, and also with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: yujun@hit.edu.cn).

Digital Object Identifier 10.1109/TIP.2024.3390984

I. Introduction

LEARNING to answer discretionary free-form questions based on long-term videos has been an increasingly popular and challenging research problem, emphasizing discriminative unimodal understanding and comprehensive cross-modal interaction to accurately infer answers. The complexity and multiplicity of long-term videos make it a more demanding task than conventional VideoQA [1], [2], [3], [4], [5]. Unlike short-term video clips with more straightforward semantics, untrimmed long-term videos tend to preserve significant redundancy and noise owing to their overly prolonged sequential frames, thus raising high requirements for models' capability and computation efficiency. To tackle the challenge, many long-term VideoQA approaches have emerged from myriad angles, e.g., discriminatory video-linguistic modeling [6], [7], [8], mighty sampling schemes [9], [10], and sufficient cross-modal interaction [11], [12] mechanisms.

Despite their considerable performance, most existing models share consistent bottlenecks: 1) Non-associative unimodal representation: Modality-independent offline feature extractors are employed to detour the costly computation overhead in unimodal representation modeling, especially for lengthy videos. The representations are learned intrinsically separated, ignoring the interplay and correlations between different modalities. 2) Asymmetric video-question paradigm: The relationship between a video and a question often exhibits a one-to-many structure in this asymmetric paradigm. Regardless of how the question changes, the representation of the referenced video stays immutable, which is counterintuitive and impedes performance enhancements. 3) Gradient blocking: As depicted in Figure 1(a), gradient flow is hindered between unimodal understanding and cross-modal interaction, rendering the model's overall optimization unviable and undermining the generation of reliable answers.

More recently, CLIP models [13], [14] have shown their powerful capability in joint representation learning by employing a unified visual-text encoder. Concurrently, Video-Language pretraining Models (VLMs) [15], [16], [17] have harnessed sparse sampling techniques to encourage affordable end-to-end modeling of raw cross-modal inputs, offering a promising inspiration to address the aforementioned bottlenecks. These models are developed through extensive pre-training endeavors, aiming at acquiring universal cross-modal knowledge to augment their understanding

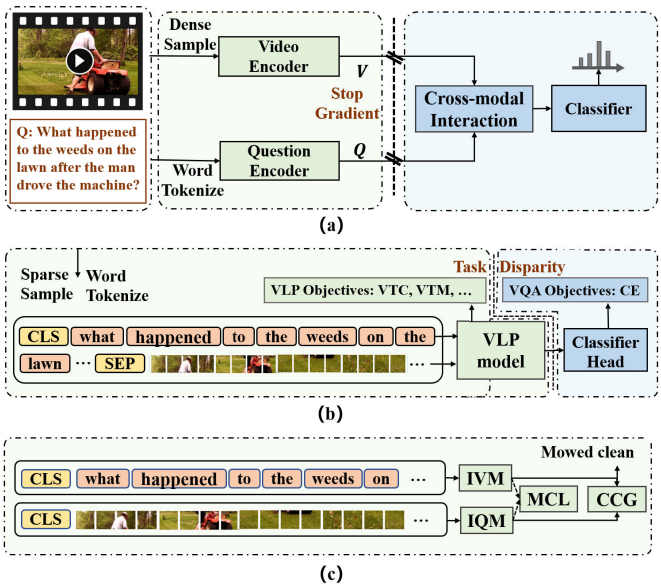


Fig. 1. Comparison of Existing VideoQA Paradigms with Our Approach. (a) Canonical models following a two-stage paradigm utilize offline feature extractors to mitigate computational overhead, yet suffer in domain-independent, modality-unrelated, gradient-blocking problems. (b) Video-language pre-training models facilitate affordable end-to-end modeling on the raw cross-modal inputs. However, the extra appended classifier head results in task disparity. (c) Our MCG embodies an entirely end-to-end generative paradigm.

of complex video and language data. As illustrated in Figure 1(b), the typical VLMs follow a two-phase paradigm: pre-training and fine-tuning. During the pre-training phase, these models undergo optimization through self-supervised tasks such as video-text matching and masked language modeling. Subsequently, they are fine-tuned to tailor their performance for specific downstream tasks, such as video question answering with task-specific objectives. Despite their promising performance, the inherent gap in task objectives between these two phases limits their generalization to downstream question-answering tasks effectively and raises pressing demand for numerous fine-tuning data. Furthermore, in comparison to the pre-trained bases with large model sizes and exposure to extensive data, the additional appended question-answering head typically exhibits a comparatively straightforward structure (e.g., a two-layer feedforward neural network). It primarily serves the purpose of adapting to downstream tasks, lacking in-depth domain-specific knowledge ratiocination, and inevitably brings unexpected extra parameters. The limited domain-specific ratiocination, along with formulation discrepancy, hinder these models from achieving superior performance.

In this paper, we reformulate VideoQA as a generative task with a novel Multi-granularity Contrastive cross-modal collaborative Generation (MCG) model without pulling in any extra heads. MCG operates as an entirely end-to-end framework for long-term VideoQA, indirectly taking raw videos and questions as inputs to generate answers. To derive discriminative representations possessing enriched visual concepts, we introduce joint unimodal modeling in a clip base [13], [14], and emphasize exploring intra-modal underlying instructive interaction between sub-components with the supervision of its complementary modality. To enhance the generation of high-quality unimodal semantics by capturing multi-granular correspondence, we propose an innovative multi-granularity contrastive learning strategy to activate the unimodal model by leveraging external web-sourced video-language pairs. This contrastive learning strategy incorporates both coarse-grained instance-level contrastive learning to ensure global semantic consistency and fine-grained token-level contrastive learning to concentrate on subtle but crucial cues that might otherwise be overlooked. To alleviate the task formulation discrepancy problem, we adapt VideoQA to a generative task instead of a classification task with a cross-modal collaborative generation module, incorporating a cross-modal fusor to facilitate deep multimodal interaction through cross-attention blocks and a video-grounded answer generator to produce answers.

In summary, the main contributions are listed as follows:

- We propose a multi-granularity contrastive cross-modal collaborative generation model, the first entirely end-to-end solution for long-term VideoQA in an open-ended generative formulation.
- We propose joint unimodal modeling to derive discriminative representations possessing high visual concepts and leverage a novel multi-granularity contrastive learning strategy to harness the intrinsically explicitly exhibited semantic correspondences.
- We propose a cross-modal collaborative generation module to reformulate VideoQA as a generative task, empowering the model with the capability for cross-modal high-semantic fusion and generation to rationalize and answer.
- Our approach achieves state-of-the-art results on four public VideoQA datasets: ActivityNet-QA, NExT-QA, MSRVTT-QA, and MSVD-QA, and successfully extends to multi-modal TVQA and diagnostic CLEVRER tasks, demonstrating consistent generalization and robustness.[1]

II. Related Work

This section provides a review of pivotal research in video question answering and video-language pre-training models, offering valuable context for our contributions.

A. Video Question Answering

VideoQA has earned increasing popularity in recent vision-language bridging research. As a straightforward but tougher extension of the ImageQA task, it targets exploring interactive intelligence to infer reliable answers by extensive communication with complicated real-world videos via natural language questions.

The earliest work [18] tried to employ a sequence-to-sequence framework directly extended from the ImageQA model [19] to answer multiple choice questions over the video frame sequences. To overwhelm the inadequacy of modeling the video temporal details, Zhao et al. [20] leveraged the hierarchical attention mechanism to capture frame and clip dual-level video dynamics. Subsequently, numerous attention

[1]The code is available at https://github.com/OpenMICG/mcg.

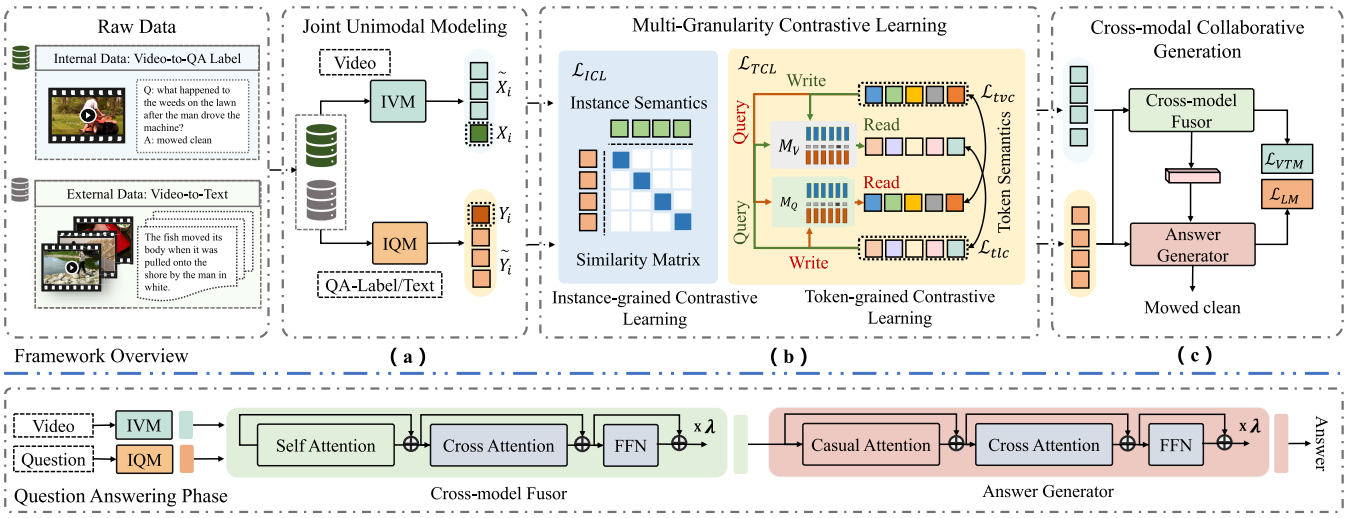


Fig. 2. **The proposed MCG framework** comprises three key components: (a) Joint Unimodal Modeling (JUM) operates to derive discriminative representations in the supervision of its complementary modality. (b) Multi-granularity Contrastive Learning strategy (MCL) leverages JUM to exploit multi-granular correspondence, enhancing the generation of high-quality unimodal semantics. MCL incorporates Instance-granularity Contrastive Learning (ICL) to capture global semantic consistency and Token-granularity Contrastive Learning (TCL) to concentrate on often overlooked yet crucial subtle cues. (c) The Cross-modal Collaborative Generation (CCG) module includes a cross-modal fusor enabling deep multimodal information interaction and an answer generator producing answers conditioned on referenced videos. During the question-answering phase, the joint unimodal encoder extracts discriminative unimodal semantics. Subsequently, these semantics pass through the cross-modal fusor to yield fused cross-modal reasoning evidence. Finally, absorbing this evidence, the answer generator generates the answer.

models flourished to study for a better focus on crucial linguistic-guided visual facts. From the spatial-temporal attention perspective, Jang et al. [21] effectively localized critical temporal frames from the video and figured out crucial spatial regions from the individual frame. Considering the significance of capturing far-distant dependency, Gao et al. [4] employed co-memory networks to model both appearance and motion evidence to infer accurate answers. To preciously associate correlated visual semantics in video with the intrinsic intention in question, Fan et al. [22] devised a heterogeneous memory network to integrate motion and appearance representations to the co-attention learning phase.

Inspired by the success of non-recurrent Transformer [23] in natural language processing, Li et al. [24] first attempted to involve Transformer structure to operate the video-question input sequences in parallel via positional self-attention blocks and perform cross-modal interaction with the co-attention mechanism. Depending on the Transformer, Peng et al. [9], [25] incorporated multilevel cross-modal interaction at different temporal scales. To strengthen the reasoning ability of the model, graph-structured models [11], [26] performed multi-step reasoning in a progressive mode and afforded considerable performance. Huang et al. [27] introduced a location-aware graph convolution structure designed to deduce not only the action categories but also the temporal locations. From the angle of conditional relation analysis, Le et al. [2] presented a general-objective reusable module to encapsulate and convert tensorial objects into conditioning representations. Gandhi et al. [28] stepped towards exploring compositional consistency in existing models with a question decomposition engine. Instead of involving the commonly-used empirical risk minimization objective, Li et al. [29] introduced invariant grounding to localize question-critical casual scenes to diminish the spurious correlations.

Neural-Symbolic solution [30] incorporated a neuro-symbolic concept learner capitalized on the intrinsic symbolic reasoning process to bridge the gap between visual concept acquisition and language semantic parsing. Drawing inspiration from neural-symbolic learning and reasoning methods [30], [31] in ImageQA, Yi et al. [32] developed a neuro-symbolic dynamic reasoning model to investigate the temporal and causal intricacies underlying synthetic object-oriented videos. This specialized framework, known as NS-DR, incorporates a video parser to obtain object-centric frame-level representations, a question parser to transform questions into functional programs, and a dynamic predictor to predict video dynamic scenes and ultimately yield insightful answers with a symbolic program executor. Building on this, Chen et al. [33] introduced a unified dynamic concept learner, strategically designed to anchor physical objects within dynamic scenes. Notably, Ding et al. [34] introduced ALOE, a more general neural network-based architecture, distinguished by its use of unsupervised techniques for object-centric representations [35], along with self-attention mechanisms and dynamics learning. Demonstrating robust performance in object-oriented scenes, ALOE exemplifies a leading-edge approach to physical dynamic visual reasoning. Subsequently, Ding et al. [36] introduced an innovative VRDP model, seamlessly integrating differentiable physics into the dynamic interaction process. These models, primarily focused on object-oriented synthetic scenes, have shown impressive dynamic visual reasoning capabilities but encounter challenges when applied to real-world scenarios with natural videos and open-ended questions.

Concurrently emerging with VideoQA [37], MovieQA presents distinctive challenges in reasoning across diverse modalities, involving not only the visual content but also additional textual resources like subtitles and plots of the movies [38], as well as TV shows [39], [40], among others. Approaches targeting both MovieQA and VideoQA share similar underlying principles. With the explosive enrichment of untrimmed web videos, research interests experience an evolution to longer and more complicated videos. Yu et al. [41] proposed the ActivityNet-QA dataset targeting understanding untrimmed videos with long duration. Zhao et al. [7], [42] employed a hierarchical reinforced network to answer questions on long-term videos. To capture long-term temporal dynamics, Yu et al. [10] proposed an action pooling stream as complementary to the uniform sampling stream to model video dynamics. Despite attaining strong performance, most of these works rely on the offline appearance (e.g., VGG [43] or ResNet [44]) and motion (e.g., C3D [45] or 3DResNet [46]) extractors, detached from the latter question-answering target, which has been a severe bottleneck and results in gradient-blocking inside the model. Unlike the aforementioned methods, we raise an entirely end-to-end solution to tackle open-ended question-answering on long-term videos.

B. Video-Language Pre-Training Models

With the commitment to acquiring universal cross-modal knowledge expressions, Video-Language Pre-training (VLP) has garnered remarkable success in various video-text downstream studies, including video retrieval [14], video captioning [47], and video question answering [17]. The representative VLP models mainly follow the encoder-only structure or encoder-decoder architecture, specifically possessing a joint encoder or dual encoder appending with a cross-modal fusion module. VideoBERT [48] steps the first attempt to explore video-language pre-training regarding video temporal dependencies and language sequential representations in encoder-only structure. ActBERT [49] proposed a joint video-text encoder to facilitate fine-grained semantical interaction between global and local visual evidence from video-text correspondences. HERO [50] performed cross-modal communication in a hierarchical mode, covering a cross-modal Transformer processing local contextualized information within individual frames and a temporal Transformer to derive the global semantic across the video. UniVL [51] presented as a unified video-text pre-training encoder-decoder framework for both multimodal expression and generation. To learn better video-text expressions, Miech et al. [52] presented a video-text encoder with contrastive learning from unlabelled and misaligned described videos. Patric et al. [53] incorporated a generative objective with the contrastive objective to train the video-text dual encoder to derive associated cross-modal representations. Bain et al. [14] introduced a dual encoder pre-trained on large-scale video and image captioning benchmarks for efficient video-text retrieval. BLIP [54] employed a captioner to effectively enhance pre-training language data quality and bootstrap image-text pre-training for unified vision-text learning. OmniVL [55] supported different visual modalities with functionality tasks unified in identical encoder-decoder architecture covering two visual-grounded decoders.

More recently emerging VLP models [15], [16], [17] incorporating sparse sampling techniques encourage affordable end-to-end modeling on the raw cross-modal inputs, offering a way to facilitate the settlement of the domain-disconnection and task-isolation problems caused by conventional two-stage VideoQA frameworks. CLIPBERT [15] presented the less-is-more theory and verified that pre-training with sparsely sampled clips is competent to reach higher accuracy compared to the conventional densely extracted features. On the top of CLIPBERT, SiaSamRea [17] embraced a siamese selection to extract several siamese video clips sparsely and explored inside knowledge among contextual clips with a siamese reasoning engine. To skillfully take benefit of region-entity visual knowledge and strengthen fine-grained alignment, ALPRO [16] introduced an entity prompts pre-training module encouraging instance-level and object-level alignment, greatly satisfying various downstream tasks. However, the additional task-specific heads involved in these models lack in-depth ratiocination and inevitably bring unexpected extra parameters. The inherent disparity in task formulation fails to deliver superior performance and raises surging demand for numerous video-question-answer data. To overcome the task formulation discrepancy problem, this paper adapts VideoQA to a generative task with a novel multi-granularity contrastive cross-modal collaborative generation model without pulling in any extra heads.

III. Method

The proposed multi-granularity contrastive cross-modal collaborative learning network is illustrated in Figure 2. Unlike the prior asymmetric format, this paper reformulates the VideoQA paradigm with a one-to-one symmetric pattern in a triplet (C, Q, A) fashion. Specifically, for an arbitrary question Q, the model stochastically picks out RGB frames sparsely in real-time to reconstruct a new augmented video clip C and performs multi-granularity cross-modal collaborative learning to generate the answer A correctly. Note that the model is optimized end-to-end, ensuring gradient flow support throughout the entire framework.

A. Joint Unimodal Modeling

Joint Unimodal Modeling (JUM) emphasizes exploring intra-modal underlying instructive interactions among sub-components with the supervision of another modality in a clip-base structure [13].

1) Intra-Video Model: To efficiently capture rich visual details from sparse frames while circumventing the expensive computation demands, we introduce an Intra-Video Module (IVM) based on TimeSformer [56] architecture. Considering continuous frames in a steady scenario always holding consistent semantics, the intra-video model performs a sparse head-tail sampling over the video, as depicted in Figure 3(a). For the sparsely sampled frames, we first partition the individual frame into non-overlapping patches

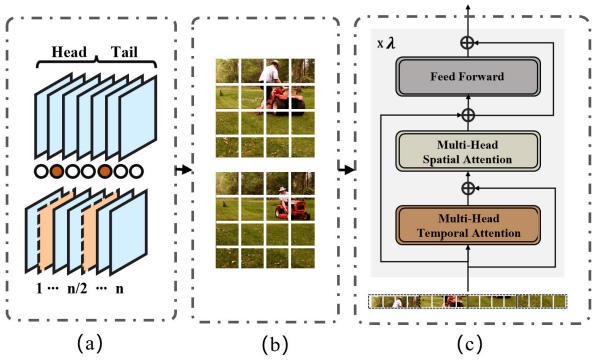


Fig. 3. Illustration of Intra-Video Model (IVM). (a) Sparse Head-Tail Sampling. (b) Frame Partition. (c) Sparsely Time-Space Divided Attention.

and then obtain a sequence of patch tokens through flattening and linear projection. Additional positional signals are also incorporated into the patch tokens. Next, the model sequentially performs multi-head temporal self-attention and multi-head spatial self-attention, focusing on temporal and spatial cues, respectively. This step-by-step attention mechanism effectively reduces computational complexity. Guided by temporal cues, the spatial self-attention module further explores spatial-range dependencies by analyzing instructive interactions across the spatial axis, leveraging a sparsely factorized spatial-dimensional attention mechanism. Finally, the spatial cue is input into the feed-forward network and combined with the temporal cue using a summation operation to generate the intra-modal video representation. To facilitate the formula expression, we present the output as $\mathcal{X} = \{x_{cls}, x_1, x_2, \ldots, x_{T_x}\} \in \mathbb{R}^{(T_x+1)\times d_x}$, where x_{cls} summarizes the video concepts.

2) Intra-Question Model: In a hierarchical collaborative parsing pattern, the Intra-Question Module (IQM) leverages a trainable multi-layer bidirectional transformer [57] to capture not only an instance-level summary but also fine-grained token-level semantics. We append an additional $[CLS]$ token to the beginning of the input to derive the global semantic and inject positional encodings into the input embeddings to memorize the order of tokens. Formally, when provided with a language input consisting of T_y words, the model produces $\mathcal{Y} = \{y_{cls}, y_1, y_2, \ldots, y_{T_y}\}$, where $y_{cls} \in \mathbb{R}^{d_y}$ preserve the instance-grained semantic, while $\{y_1, y_2, \ldots, y_{T_y}\} \in \mathbb{R}^{T_y \times d_y}$ holds the fine-grained token-level embeddings of d_y dimensionality.

B. Multi-Granularity Contrastive Learning

To harness multi-granular correspondence and facilitate the generation of high-quality intra-modal semantics covering broad visual concepts, we introduce a novel Multi-granularity Contrastive Learning (MCG) strategy to activate the joint unimodal model by leveraging external web-sourced video-language pairs. We incorporate contrastive learning from two aspects: coarse-grained instance-level contrastive learning and fine-grained token-level contrastive learning.

1) Instance-Grained Contrastive Learning: To capture the global semantic consistency, we incorporate Instance-grained Contrastive Learning (ICL) to encourage positive cross-modal pairs to be mapped nearby while negative pairs are as far apart as possible in the shared semantic space. Following [13], we first adopt two linear functions $g_x(\cdot)$ and $g_y(\cdot)$ to project the instance-wise semantics $\{x_{cls}\}$ and $\{y_{cls}\}$ into a normalized space. Then, we perform the similarity function on the intra-modal video semantic $\mathcal{X}$ and the intra-modal text semantic $\mathcal{Y}$ as follows:

$$sim(\mathcal{X}, \mathcal{Y}) = g_x(x_{cls}) \cdot g_y(y_{cls}) \tag{1}$$

We use the symmetric temperature-normalized contrastive learning strategy to maximize the interactions between the matched pairings. The two stream cross-modal inputs $\langle \mathcal{X}_i, \mathcal{Y}_i \rangle$ in a minibatch are alternately taken as queries and keys:

$$\begin{aligned} \ell_i^{x2y} &= -log \frac{exp(sim(\mathcal{X}_i, \mathcal{Y}_i)/\tau_1)}{\sum_{j=1}^{B} exp(sim(\mathcal{X}_i, \mathcal{Y}_j)/\tau_1)} \\ \ell_i^{y2x} &= -log \frac{exp(sim(\mathcal{Y}_i, \mathcal{X}_i)/\tau_1)}{\sum_{j=1}^{B} exp(sim(\mathcal{Y}_i, \mathcal{X}_j)/\tau_1)} \end{aligned} \tag{2}$$

where B represents the batch size and τ_1 denotes the learnable instance-wise temperature parameter. The symmetric temperature-normalized contrastive loss of instance-grained cross-modal module is formulated as:

$$\mathcal{L}_{ICL}(; \mathcal{T}_{vm}, \mathcal{T}_{qm}) = \frac{1}{2N} \sum_{i=1}^{N} (\ell_i^{x2y} + \ell_i^{y2x}) \tag{3}$$

where $\mathcal{T}_{vm}$,$\mathcal{T}_{qm}$ represent the parameters of the intra-video model and intra-question model. N denotes the whole number of the video-text pairs.

2) Token-Grained Contrastive Learning: Considering the fact that some vital subtle clues are feasible to be ignored during the instance-grained cross-modal optimization, we propose Token-grained Contrastive Learning (TCL) to explicitly execute in-depth interaction between the video patches and text tokens. Specifically, for the i^{th} video-text pair $\langle \mathcal{X}_i, \mathcal{Y}_i \rangle$, we reserve the video token semantics $\tilde{\mathcal{X}}_i = \{x_i^1, \ldots, x_i^j, \ldots, x_i^J\}$ into an internal memory buffer and mapped the k^{th} text token semantic y_i^k in $\tilde{\mathcal{Y}}_i = \{y_i^1, \ldots, y_i^k, \ldots, y_i^K\}$ to an internal state u_i^k of memory-dimension d_m. Then, we let the internal state u_i^k guide the knowledge extraction from memory $\mathcal{M}_i$ and learn the attention weights ρ_i to deliver the corresponding cross-modal video response r_i^k.

$$r_i^k = \mathcal{M}_i^T softmax(tanh(\mathcal{M}_i))^T tanh(u_i^k)) \tag{4}$$

where $\mathcal{M}_i \in \mathbb{R}^{J\times d_m}$ denotes the i^{th} memory buffer with J memory slots. Subsequently, the symmetric temperature-normalized contrastive loss is employed to pull y_i^k close to its corresponding video response r_i^k, but far away from other video tokens. The token-grained video-to-language contrastive loss can be formulated as follows:

$$\begin{aligned} \ell_i^{y2r} &= -log \frac{exp(sim(y_i^k, r_i^k)/\tau_2)}{\sum_{j=1}^{K} exp(sim(y_i^k, r_i^j)/\tau_2)} \\ \ell_i^{r2y} &= -log \frac{exp(sim(r_i^k, y_i^k)/\tau_2)}{\sum_{j=1}^{K} exp(sim(r_i^k, y_i^j)/\tau_2)} \end{aligned}$$

$$\mathcal{L}_{tvc} = \frac{1}{2NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_i^k(\ell_i^{y2r} + \ell_i^{r2y}) \tag{5}$$

where τ_2 represents the learnable token-grained temperature parameter, note that α_i^k is the saliency weight assigned to the k^{th} textual token, allowing us to distinguish the various significance of different word tokens. Simultaneously, for the j^{th} video token-grained semantics in the i^{th} pair, we calculate the cross-modal textual response $\tilde{r}_i^j$ and perform a contrastive operation between the video token-grained semantic x_i^j and $\tilde{r}_i^j$ in the same manner. The token-grained language-to-video contrastive loss is expressed as:

$$\mathcal{L}_{tlc} = \frac{1}{2NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\beta_i^j(\ell_i^{x2\tilde{r}} + \ell_i^{\tilde{r}2x}) \tag{6}$$

where β_i^j denotes the importance parameter assigned to the j^{th} visual patch. The final symmetric temperature-normalized contrastive loss of the token-grained contrastive learning is defined as:

$$\mathcal{L}_{TCL}(;\mathcal{T}_{vm},\mathcal{T}_{qm}) = \frac{1}{2}(\mathcal{L}_{tvc} + \mathcal{L}_{tlc}) \tag{7}$$

The multi-granularity contrastive objective can be represented as:

$$\mathcal{L}_{MCL}(;\mathcal{T}_{vm},\mathcal{T}_{qm}) = \theta_1 * \mathcal{L}_{ICL} + \theta_2 * \mathcal{L}_{TCL} \tag{8}$$

where θ_1 and θ_2 are hyperparameters that control the balance between the two levels of contrastive learning and are typically set to 1 by default.

C. Cross-Modal Collaborative Generation

The Cross-modal Collaborative Generation module (CCG) equips our model with the capability for cross-modal interaction and generation so as to reason and describe. We design a cross-modal fusor to enable the deep interaction of multi-modal information with cross-attention blocks and an answer generator to generate answers conditioned on the referenced video.

1) The Cross-Modal Fusor: Taking the video and linguistic semantics as inputs, the Cross-modal Fusor (CFor) is dedicated to generating a fused cross-modal reason result by exploring deeper informative interaction and communication with stacked transformer blocks. Each transformer block comprises a Self-Attention (SA) layer, a Cross-Attention (CA) layer, and a Feed-Forward Network (FFN). An additional [FUS] token is appended to deliver the fused cross-modal reason result. We adopt the widely applied video-text matching (VTM) loss $L_{VTM}(;\mathcal{T}_{cf},\mathcal{T}_{vm})$ to activate the CFor module for learning cross-modal fusion, capturing fine-grained interactions and alignments between different modalities. Afterward, we introduce a fully connected layer to the CFor output, generating a two-category probability, p^{vtm}. $\mathcal{H}$ calculates the binary cross-entropy between p^{vtm} and the ground-truth q^{vtm}.

$$\mathcal{L}_{VTM}(;\mathcal{T}_{cf},\mathcal{T}_{vm}) = \mathbb{E}_{(\mathcal{X},\mathcal{Y})\sim\mathcal{D}}\mathcal{H}(q^{vtm}, p^{vtm}(\mathcal{X},\mathcal{Y})) \tag{9}$$

2) The Answer Generator: The Answer Generator (AGor) targets generating open-ended answers. Conditioned on the referenced video and the fused reason evidence, the AGor plays the text generator roles by employing a CFor-similar transformer while replacing the SA layer with casual self-attention following [54] and [55]. Additionally, tokens [GEN] and [EOS] are separately added to signal the task and the end. Recent works [58] reveal that language modeling loss facilitates the model with the generalization ability to transform visual facts into coherent descriptions. Building on this inspiration, we activate the AGor module using Language Modeling Loss (LM), denoted as $\mathcal{L}_{LM}(;\mathcal{T}_{ag},\mathcal{T}_{ivm})$, to maximize the likelihood of the input text in an autoregressive mode.

$$\mathcal{L}_{LM}(;\mathcal{T}_{ag},\mathcal{T}_{vm}) = \mathbb{E}_{(x_i,y_i,z_i)\sim\mathcal{D}}\left[\sum_{l=1}^{L}\log P(z_i^l|z^{<l},x_i)\right] \tag{10}$$

where L refers to the length of the input sequence, while $\mathcal{T}_{ag}$ and $\mathcal{T}_{vm}$ represent the parameters of the AGor and intra-video model, respectively.

D. Overall Training Objectives

We jointly train our MCG framework with three loss functions: multi-granularity contrastive learning (MCL) loss, video-text matching (VTM) loss, and language modeling (LM) loss. Both internal VideoQA data and external web-sourced video-language pairs are utilized to optimize the model. During fine-tuning, we use only the LM loss. The overall training objective, combining Eqn. 8 through Eqn. 10, can be represented as:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{MCL} + \lambda_2 * \mathcal{L}_{VTM} + \lambda_3 * \mathcal{L}_{LM} \tag{11}$$

where λ_1, λ_2, and λ_3 are balanced hyper-parameters, with default values set to 1. Note that during the question-answering phase, we first convert the question and the video into discriminative unimodal semantics by joint unimodal modeling. Subsequently, we feed these unimodal semantics into the CFor module to deliver the fused cross-modal reason evidence. Finally, with this reasoning evidence incorporated, we employ the AGor to generate the answer.

IV. Experiments

A. Benchmark and Implementation Details

1) Benchmarks: We assess the performance of the proposed MCG on six publicly available datasets for video question answering, following standard data preprocessing, evaluation, and settings for each benchmark. Detailed statistics are presented in Table I.

- **ActivityNet-QA** [41] targets comprehending long-term videos through manually curated question-answer pairs. It challenges models with open-ended motion, temporal-spatial relationships, and standard description questions. It holds 5.8K untrimmed complex web videos with an average duration of 180 seconds and 58K question-answer labels with an average length of 8.7 and 1.9 words.

TABLE I
DETAILED STATISTICS OF THE SIX REPRESENTATIVE VIDEOQA DATASETS. VLEN. INDICATES THE AVERAGE DURATION OF THE VIDEOS, AND Q/A LEN. DENOTES THE AVERAGE LENGTH OF QUESTIONS AND OPEN-ENDED ANSWERS

Benchmark	#Video	#QA	VLen.	Q/A Len.
ActivityNet-QA [41]	5,800	58,000	180s	8.7/1.9
NExT-QA [59]	5,440	52,044	44s	11.6/2.6
MSRVTT-QA [60]	10,000	243,680	15s	7.4/1
MSVD-QA [60]	1,970	50,505	10s	6.6/1
TVQA [39]	21,793	152,545	76s	13.5/-
CLEVRER [32]	20,000	305,280	5s	11/-

- **NExT-QA** [59] is a newly presented dataset emphasizing causal and temporal video question reasoning beyond descriptive goals. It contains 5,440 videos with an average duration of 44 seconds and humanly annotated questions and answers with an average of 11.6 and 2.6 words. Both open-ended QA and multi-choice QA are supported.
- **MSRVTT-QA** [60] is automatically annotated to understand short-term video and focus on standard open-ended descriptive QA. It is large-scale, covering 10K trimmed video clips with a short average duration of 15s and 244K QA labels of straightforward syntactic structure.
- **MSVD-QA** [60] focus on open-ended descriptive QA tasks similar to MSRVTT-QA. However, It is small-scale, carrying only 2K videos with the shortest average duration of 10s. The automatically generated 51K QA labels primarily follow a simple grammatical format due to the natural language annotation algorithm.
- **TVQA** [39] targets multi-modal video question answering. This dataset poses unique challenges as it requires reasoning across diverse modalities, encompassing not only visual content from television but also supplementary resources like subtitles and speech.
- **CLEVRER** [32] is a unique diagnostic video question answering dataset, that emphasizes investigating the temporal and causal relationships within videos featuring simple visual objects, under a meticulously controlled environment.

2) Implementation Details: We implement the MCG model via the PyTorch deep-learning framework [61]. Specifically, the sparse spatial-temporal factorized attention block within the intra-video module is initialized with weights from the ViT-B/16 model pre-trained on ImageNet [62]. The intra-question module is initialized from the former six layers of BERT$_{base}$ [57]. The cross-modal fusor and generator are initialized with the last six-layer weights from the BERT$_{base}$ model. Following ALPRO [16], we construct external pre-training data utilizing the web-sourced dataset WebVid-2M [14], which contains 2.5 million video-text pairs, and the image-to-video preprocessed conceptual captions dataset CC-3M [63] featuring 3 million image-text pairs. Additionally, to get better pre-trained parameters for VideoQA adaption, we built internal VideoQA-specific pre-training data with TGIF-QA [21]. We pre-trained MCG for ten epochs and optimized it with AdamW [64] optimizer with a 0.001 weight decay. The learning rate was initially warmed up to $1e^{-4}$ and then linearly decayed to $1e^{-5}$. For variable-length diverse-resolution video input, we employed the sparsely head-tail sampling strategy to derive $4 * 224 * 224$ video clips. During the fine-tuning stage, we raise the frame resolution to $384 * 384$ and tweak the sparse sample number to 8. To accommodate different scales and domains, we employed dataset-specific training epochs and learning rates based on validation results. During testing, we observe the standard test split settings and assess the quality of generated answers with the top-1 Accuracy [60] and the Wu-Palmer similar (WUPS) [65] evaluation criteria.

B. State-of-the-Art Comparison

We evaluate the performance of our MCG model by comparing it against the state-of-the-art methods on four publicly available VideoQA benchmarks. Detailed experimental comparisons and analyses are presented below.

1) Comparison on AcitivityNet-QA: In Table II, we present a comprehensive comparison of the MCG model with existing state-of-the-art models, including RNN-based approaches like EVQA [18], memory networks such as CoMem [4], HME [22], and MHMAN [12], as well as modular attention networks like AMU [60] and CAN [10], all of which follow a two-stage paradigm. Additionally, we reproduce the typical video-language pre-training model ALPRO [16] on long-term videos as the representative pre-trained comparison model. Our experimental results yield several noteworthy findings: ALPRO consistently outperforms all existing frameworks and achieves a remarkable 2.7% improvement in overall accuracy over the state-of-the-art two-stage network MHMAN. This underscores the crucial role of external implicit knowledge acquired through pre-training in enhancing long-term video understanding and QA reasoning. However, as impressive as ALPRO's performance is, it falls short of our proposed MCG model. MCG achieves state-of-the-art results across all evaluation metrics, surpassing ALPRO by 3.5% in overall accuracy, 4.6% in WUPS@0.9, and 3.8% in WUPS@0.0. These results underscore the significance of end-to-end video rationale coupled with answer generation. A closer examination reveals that MCG exhibits a remarkable 5.3% improvement in motion-related tasks, a 4.8% improvement in tasks involving spatial relationships, and a 5.2% improvement in tasks related to temporal relationships when compared to the pre-trained SoTA model [16]. This notable performance enhancement in these critical aspects demonstrates MCG's ability to grasp both long video dependencies and spatial cues, which are particularly valuable in long-term VideoQA. This detailed comparison on the ActivityNet-QA benchmark underscores the superiority of the MCG model in handling complex video understanding and reasoning tasks.

2) Comparison on MSRVTT-QA and MSVD-QA: We further investigate the performance of the MCG model on short-term videos by evaluating it on the MSRVTT-QA and MSVD-QA datasets. The results of this comparison are presented in Table III. From the experimental results, it is evident that MCG exhibits a significant improvement over previous state-of-the-art models on both benchmarks. Notably, it achieves an impressive 8.4% and 11.5% increase in overall accuracy, demonstrating its effectiveness in handling short-term

TABLE II

EXPERIMENTAL COMPARATIVE RESULTS OF THE PROPOSED MCG AND EXISTING SOTAS ON ACTIVITYNET-QA BENCHMARK. ALL RECORDS ARE LISTED AS A PERCENTAGE (%). "SPAT. REL" AND "TEMP. REL." ABBREVIATE SPATIAL RELATIONSHIP AND TEMPORAL RELATIONSHIP TASKS, RESPECTIVELY. "OBJ.," "LOC.," AND "NUM." CORRESPOND TO OBJECT, LOCATION, AND NUMBER TASKS, RESPECTIVELY. THE SYMBOL † DENOTES A REPRODUCTION VERSION IMPLEMENTED ON LONG-TERM VIDEOS, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	Accuracy (%)											WUPS (%)	
	Motion	Spat. Rel.	Temp. Rel.	Y/N	Color	Obj.	Loc.	Num.	Other	Free	All	@0.9	@0.0
EVQA[18]	2.5	6.6	1.4	52.7	27.3	7.9	8.8	44.2	20.6	34.3	25.1	29.3	53.5
CoMem[4]	16.1	13.8	2.9	58.3	29.8	15.4	18.1	44.4	29.3	40.3	31.5	34.3	56.1
AMU[60]	9.8	14.0	2.3	61.1	27.0	19.8	23.1	45.2	30.1	41.9	31.9	36.1	57.3
CAN[10]	21.1	17.3	3.6	62.6	31.1	20.1	30.6	48.0	33.3	44.5	35.4	40.5	60.0
HME[22]	19.4	14.1	3.0	59.7	31.6	16	19.4	45.4	30.0	41.4	32.7	36.7	57.1
MHMAN[12]	23.1	19.8	4.4	63.2	32.1	22.0	33.2	48.5	36.8	46.2	37.1	42.8	62.1
ALPRO†[16]	23.5	19.3	2.9	71.3	36.4	25.5	35.0	52.5	35.7	50.3	39.8	43.9	65.3
MCG	**28.8**	**25.1**	**8.1**	**73.1**	**38.9**	**30.8**	**38.9**	**55.6**	**38.7**	**53.0**	**43.3**	**48.5**	**69.1**

TABLE III

EXPERIMENTAL COMPARATIVE RESULTS OF THE PROPOSED MCG AND EXISTING SOTAS ON MSRVTT-QA (LEFT) AND MSVD-QA (RIGHT) BENCHMARKS. ALL VALUES ARE REPORTED AS PERCENTAGES (%). THE SYMBOL † DENOTES A REPRODUCTION VERSION CONDUCTED BY US. THE BEST OUTCOMES ARE EMPHASIZED IN BOLD

Method	MSRVTT-QA						MSVD-QA					
	What	Who	How	When	Where	All	What	Who	How	When	Where	All
EVQA [18]	18.9	38.7	83.5	70.5	29.2	26.4	9.7	42.4	83.8	72.4	53.6	23.3
STVQA [66]	24.5	41.2	78.0	76.5	34.9	30.9	18.1	50.0	83.8	72.4	28.6	31.3
CoMem [4]	23.9	42.5	74.1	69.0	42.9	32.0	19.6	48.7	81.6	74.1	31.7	31.7
AMU [60]	26.2	43.0	80.2	72.5	30.0	32.5	20.6	47.5	83.5	72.4	53.6	32.0
CAN [10]	26.7	43.4	83.7	75.3	35.2	33.2	21.1	47.9	**84.1**	74.1	57.1	32.4
TSN [67]	27.9	46.1	84.1	77.8	37.6	35.4	25.0	51.3	83.8	**78.4**	**59.1**	36.7
HME [22]	26.5	43.6	82.4	76.0	28.6	33.0	22.4	50.1	73.0	70.7	42.9	33.7
MHMAN [12]	28.7	47.1	85.1	77.1	35.2	35.6	23.3	50.7	**84.1**	72.4	53.6	34.6
HGA [26]	29.2	45.7	83.5	75.2	34.0	35.5	23.5	50.4	83.0	72.4	46.4	34.7
ALPRO†[16]	36.0	51.7	**85.7**	79.6	42.3	41.9	36.0	57.0	82.2	72.4	46.6	44.7
MCG	**38.1**	**53.9**	**85.7**	**81.2**	**45.2**	**44.0**	**39.4**	**60.6**	**84.1**	77.6	**57.1**	**48.2**

video comprehension tasks. While the performance improvements are relatively modest compared to long-term video datasets, these results underscore the remarkable capabilities of the entirely end-to-end MCG structure. To delve deeper into the reasons behind this achievement, we reproduce the representative VLP model ALPRO, which leverages external pre-training on 5.5 million video-text pairs. As expected, ALPRO substantially enhances performance, achieving 41.9% overall accuracy on MSRVTT-QA and 44.7% overall accuracy on MSVD-QA. This surpasses the hierarchical graph-based network HGA [26], equipped with sophisticated reasoning mechanisms, by a margin of 6.4% and 10.0% in accuracy. These findings highlight the value of external knowledge involving weak supervision pre-training is conducive to delivering a good performance on video question answering tasks. Surprisingly, our MCG model surpasses this newly pre-trained SoTA model by 2.1% on MSRVTT-QA and 3.5% on MSVD-QA. Moreover, when considering specific question types, MCG consistently achieves the highest accuracy on both MSRVTT-QA and MSVD-QA, illustrating its competence in short-term video descriptive analysis.

3) Comparison on NExT-QA: We conducted additional experiments on the challenging NExT-QA dataset, which places a strong emphasis on temporal and causal reasoning in a multi-choice task. To ensure a fair comparison, we adjusted the MCG model to adapt to the multi-choice paradigm by calculating the video-text matching score for each video and the answer candidates. As depicted in Table IV, the proposed MCG outperforms prior state-of-the-art models across all sub-questions, including causal, temporal, and descriptive

TABLE IV

EXPERIMENTAL COMPARATIVE RESULTS OF ACCURACY ON NEXT-QA DATASET. ALL RECORDS ARE LISTED AS A PERCENTAGE (%). THE "C", "T", AND "D" REFER TO THE CAUSAL, TEMPORAL, AND DESCRIPTIVE TASKS, RESPECTIVELY. THE BEST RECORDS ARE EMPHASIZED IN BOLD

Method	NExT-QA			
	Acc@C	Acc@T	Acc@D	Acc@A
EVQA [18]	43.27	46.93	45.62	44.92
STVQA [66]	45.51	47.57	54.59	47.64
CoMem [4]	45.85	50.02	54.38	48.54
HCRN [2]	47.07	49.27	54.02	48.89
HME [22]	46.76	48.89	57.37	49.16
HGA [26]	48.13	49.08	57.79	50.01
IGV [29]	48.56	51.67	59.64	51.34
JustAsk [68]	41.66	44.11	59.97	45.30
ATP [69]	53.10	50.20	66.80	54.30
VGT [11]	52.78	54.54	67.26	55.70
MCG	**57.31**	**56.45**	**68.19**	**58.83**

TABLE V
ABLATION STUDIES ON THE ACTIVITYNET-QA TEST SPLIT. ALL VALUES ARE PRESENTED AS PERCENTAGES (%). "SPA. R." AND "TEM. R." ABBREVIATE SPATIAL RELATIONSHIP AND TEMPORAL RELATIONSHIP, RESPECTIVELY

Methods	Motion	Spa.R.	Tem.R.	Free	All
Baseline	18.8	14.4	3.0	43.8	34.3
w/ JUM	24.4	21.9	3.5	47.3	38.2
w/ JUM&MCL	25.1	21.4	4.3	51.1	41.4
w/ JUM&MCL&CCG	28.8	25.1	8.1	53.0	43.3
MCG w/o TCL	23.8	20.0	4.3	49.9	39.8
MCG w/o ICL	25.0	23.8	4.4	52.3	41.9
MCG w/ ICL&TCL	28.8	25.1	8.1	53.0	43.3
MCG-CH	25.1	21.4	4.3	51.1	41.4
MCG	28.8	25.1	8.1	53.0	43.3

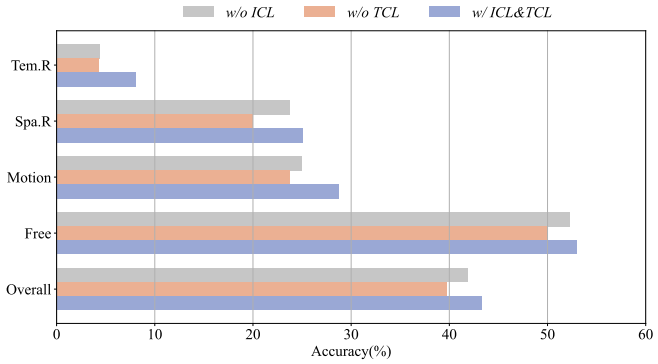


Fig. 4. In-depth ablation study comparing multi-granularity contrastive learning (w/ ICL&TCL) with uni-granularity contrastive learning (w/o ICL and w/o TCL) across various task types, including temporal relationship (Tem. R.), spatial relationship (Spa. R.), motion, and free tasks. Best viewed in color.

tasks. Specifically, compared to the two-stage SoTA model IGV [29], which is conditioned on invariant grounding without cross-modal pre-training, MCG achieves a remarkable 7.4% increase in overall accuracy and an incredible 8.7% improvement in the causal reasoning task. This improvement is attributed to the collaborative reasoning capability of MCG. We also conducted a detailed examination of prior SoTA models, namely VLP JustAsk [68], ATP [69], and VGT [11], which are pre-trained on large-scale external web-sourced data. The experimental results are presented in the second split of Table IV. Interestingly, we observe that these typically pre-trained models on NExT-QA have seen limited performance boosting in temporal and causal reasoning types. This limitation may be attributed to the domain gap between externally sourced data and QA-specific data. However, MCG, which incorporates both external and domain-specific data, effectively mitigates domain disconnection issues and consistently outperforms these competitive video-language pretraining counterparts [11], [68], [69]. It achieves absolute overall accuracy gains of 13.5%, 4.5%, and 3.3%, respectively, highlighting its effectiveness and causal reasoning advantage. These results underscore the superior performance and causal reasoning capabilities of the MCG model on the NExT-QA dataset.

C. Ablation Study

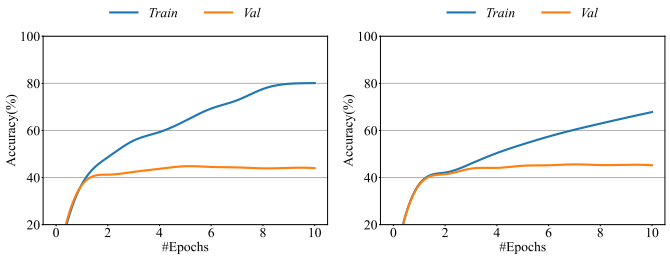
In this section, we conduct a comprehensive ablation analysis on the ActivityNet-QA dataset. Our objective is to investigate the effects of several critical components: joint unimodal modeling, multi-granularity contrastive learning, and the cross-modal collaborative generation module in the context of long-term videos. Additionally, we delve into the influence of varying video sampling rates for variable-length videos across different benchmarks, explore the impact of model parameters and pre-training data size, and assess MCG's generalization ability across different video scenarios.

1) MCG Versus Baseline: In the top split of Table V, we aim to assess the individual contributions of various components within the MCG model. To establish a baseline, we systematically remove or replace key components from the complete MCG system. These components include the JUM module, the MCL strategy, and the CCG module substituted with the AGor. Specifically, we replace the JUM with offline feature extractors [10], discard the MCL strategy, and replace the AGor in the CCG with a classification head in the system baseline. **JUM:** We begin by examining the impact of JUM, and the ablation results are presented in the first two rows of Table V. When compared to the system baseline, equipping the model with joint unimodal modeling (w/ JUM) in clip fashion [13] leads to a consistent improvement in performance. Specifically, the overall accuracy increases by 3.9%, with notable performance enhancements observed across all sub-questions. Notably, tasks related to motion, spatial relationships, and free-form tasks see substantial improvements, with scores increasing by 5.6%, 6.5%, and 3.5%, respectively. These results underscore the effectiveness of joint unimodal modeling in expressing discriminative representations with the supervision of complementary modalities, thereby playing a vital role in enhancing video question-answering capabilities.

MCL: To assess the performance of MCL, we enhance the baseline by incorporating MCL into the JUM module (w/ JUM&MCL). The experimental results demonstrate a notable 7.1% improvement in overall accuracy when MCL is combined with JUM. This outcome validates our hypothesis that MCL has the capability to harness hierarchical intrinsic semantic correspondences, thereby facilitating cross-modality understanding.

CCG: We further explore the impact of the CCG module. When we add the collaborative generation module to the baseline system equipped with JUM and MCL, denoted as w/ JUM&MCL&CCG, our model experiences significant performance improvement, achieving the best performance. Quantitatively, the CCG module contributes to an average 3.7% gain in motion-related, spatial relationship-related, and temporal relationship-related tasks, ultimately resulting in an overall accuracy of 43.3%. These substantial enhancements underscore the effectiveness of the CCG module.

2) Multi-Granularity Versus Uni-Granularity: In the middle part of Table V, we conduct an in-depth analysis of MCL by comparing it to uni-granularity contrastive learning. Firstly, we remove the TCL loss from the overall training objectives, effectively setting θ_2 to 0 in Eqn.8, and retain only the



(a) MCG-CH (b) MCG

Fig. 5. Visualization of training and validation accuracy trends across various training epochs for (a) MCG-CH, a model variant employing a Classifier Head, and (b) MCG, the proposed model with language generation capability. Optimal viewing experience in color.

TABLE VI

ABLATION STUDIES INVESTIGATING THE IMPACT OF DIFFERENT VIDEO FRAME SAMPLING RATES ON FOUR VIDEO QUESTION ANSWERING BENCHMARKS. ALL RECORDS ARE REPORTED AS A PERCENTAGE (%)

#Frames	2	4	8	12	16
ActivityNet-QA	39.8	41.6	43.3	44.1	44.3
NExT-QA	54.3	56.9	58.8	59.6	59.9
MSRVTT-QA	42.1	42.9	44.0	44.2	44.3
MSVD-QA	46.3	47.1	48.2	48.3	48.3

uni-granularity ICL loss. This configuration is represented as MCG w/o TCL. The comparison results clearly demonstrate that eliminating the TCL loss leads to a noticeable drop in performance across all sub-tasks. Similarly, we also experiment with removing the ICL loss from the full model by setting the hyperparameter θ_1 to 0 in Eqn.8, referred to as MCG w/o ICL. Once again, the results show significant performance degradation compared to the multi-granularity contrastive model. Detailed comparisons are visualized in Figure 4. In summary, both uni-granularity contrastive losses (i.e., ICL or TCL) play a crucial role in performance enhancement, while multi-granularity contrastive learning, which incorporates both of them, achieves the best performance.

3) Generation Versus Classification: We study MCG's variant by replacing the answer generator with a classifier head. Specifically, the output of the [FUS] signal from the cross-modal fusion module is directed to a K-way classifier following [15], [16], and [17]. As illustrated in the bottom block of Table V, this classification variant model, referred to as MCG-CH, results in a significant drop in performance, highlighting that the task-specific classifier head lacks the depth of reasoning required for video question answering. To further verify our assumption that incorporating a task-disparity classifier head inevitably brings in unexpected extra parameters and may lead to severe over-fitting in video question answering, we conducted in-depth experiments to analyze the performance on both training and validation data, respectively. Figure 5 provides a visual representation of training and validation scores across various training epochs. The experimental results indeed confirm the presence of overfitting in the classifier variant model MCG-CH. Moreover, it's evident from the trends in the MCG curves that MCG with a generative solution can mitigate the overfitting problem to some extent.

TABLE VII

COMPARISON WITH EXISTING VIDEO-LANGUAGE PRE-TRAINING MODELS CONCERNING MODEL PARAMETER SCALE AND PRE-TRAINING DATA SIZE

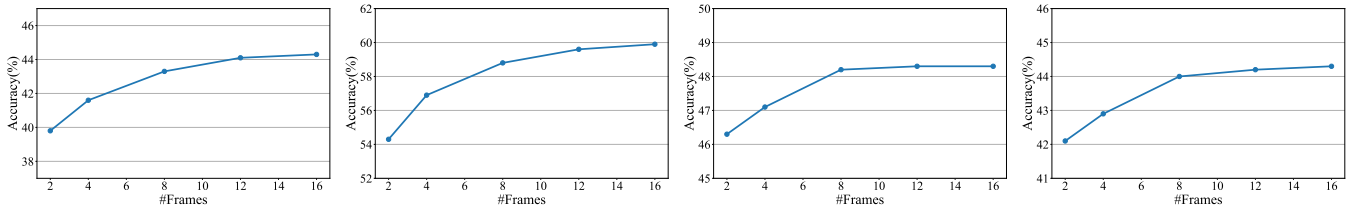
Method	PT Data	Model Size
ClipBERT [15]	COCO[70]&VG[71](5.6M)	137M
JustAsk[68]	HTVQA[68](69M)	600M
ATP [69]	WebIT[13](400M)	428M
VGT [11]	WV[14](0.18M)	511M
ALPRO[16]	WV[14]&CC[63](5.5M)	240M
FrozenBiLM [72]	WebVid10M[14](10M)	890M
MCG	WV[14]&CC[63]&TGIF[21](5.6M)	297M

In summary, the ablation results underscore the superiority of solving question-answering tasks in a generative paradigm and simultaneously subtly alleviate the task formulation discrepancy.

4) Sparse Sampling Versus Dense Sampling: In Table VI, we investigate the influence of video frames sampling rate across all four video question answering benchmarks. Intuitively, increasing the number of frames should bring in more information and potentially facilitate the delivery of better performance. Figure 6 illustrates the performance trends of the model as the number of sampled frames varies from sparse to dense. From the evaluation curve trends, we observe that the model consistently experiences rapid performance improvement across all benchmarks when using sparse samples, ranging from 2 to 8 frames. However, as the number of sampled frames gradually increases from 8 to 16, the performance tends to reach saturation, especially for short-term videos. Even for the ActivityNetQA dataset, which features prolonged videos, the model's performance approaches saturation when the number of samples reaches eight, with only slight accuracy gains beyond this point. Considering the balance between computational efficiency and model performance, we have consistently set the sampling number to 8 for our final MCG across all benchmarks. Sparse sampling not only enables efficient end-to-end modeling but also benefits video question answering with high quality and efficiency.

5) Model Parameters and Pre-Training Data Size: In Table VII, we present a comprehensive comparison of MCG with other video-language pre-trained models, focusing on model size and pre-training data size. It is generally understood that larger models, supported by extensive pre-training data, typically yield enhanced performance. Remarkably, our MCG model demonstrates superior results when compared to models like ALPRO, which are similar in scale concerning both model size and pre-training data size. MCG even surpasses larger-scale architectures such as JustAsk, which has 600 million parameters and benefits from a large, domain-specific pre-training dataset. This demonstrates that MCG's performance is not merely a function of its parameter count or data scale.

We further extend our analysis to Image-Language pre-trained Models (ILMs) [13], [54], [73], [75] versus Video-Language pre-training models (VLMs) [11], [15], [16], [68], as visualized in Figure 7. Since the increased accessibility of image data and the comparatively simpler data structure,



(a) AcitivityNet-QA (b) NExT-QA (c) MSRVTT-QA (d) MSVD-QA

Fig. 6. Variations in test accuracy by employing different video frame sampling rates on (a) ActivityNet-QA, (b) NExT-QA, (c) MSRVTT-QA, and (d) MSVD-QA. The number of sampling frames ranges from 2 to 16. A consistent performance boost is observed when the number of samples varied from 2 to 8 while gradually reaching saturation from 8 to 16. Best viewed in color.

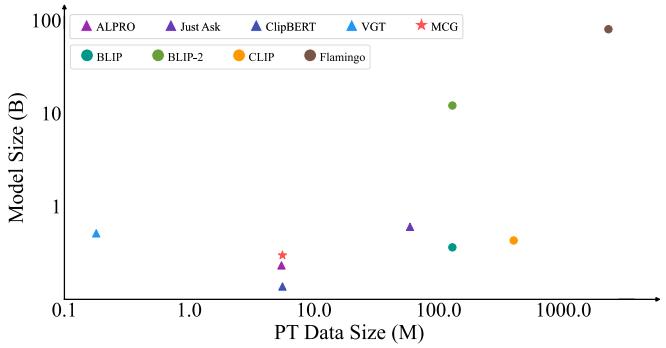


Fig. 7. Comparison of Image-Language and Video-Language pre-training models concerning model parameter scale and pre-training data size.

TABLE VIII
COMPARISON OF DIRECT IMAGE-LANGUAGE MODEL ADAPTATION TO VIDEO QUESTION ANSWERING ON ACTIVITYNET-QA

Method	#Total Params	#Trainable Params	#PT Data	Acc.
BLIP [54]	305M	305M	129M	24.2
BLIP-2 [73]	3.1B	104M	129M	35.4
MCG	297M	297M	5.6M	43.3

ILMs have advanced more rapidly than VLMs. As the figure illustrates, VLMs generally exhibit smaller scales in terms of model and pre-trained data sizes. This discrepancy arises from the cost and complexity associated with obtaining video-level annotations due to the temporal nature of videos and their richer content. The notable gap between ILMs and VLMs highlights the urgent need for more efficient tuning methods and strategies for adapting powerful image-language models to VideoQA tasks.

To provide preliminary insights into the adaptability of image-language pre-trained models to video question answering, we conducted two sets of exploratory experiments. Table VIII presents a comparative performance of BLIP (ViT-B) [54], BLIP-2 (ViT-L-OPT2.7B) [73], and our MCG model on ActivityNet-QA. To adapt BLIP to the video, we randomly selected a single frame per video to ensure a straightforward end-to-end training process without modifying the model's architecture. We adapted BLIP-2 to video with a concatenation setting, where the uniformly sampled eight frames were processed by the ViT-L [13] and the Q-former, with OPT [76] taking the concatenated visual features as the prefix. The results from these experiments have been insightful, revealing a performance gap due to domain differences and the lack of temporal modeling when directly adapting image-language pre-training models to VideoQA. Despite this, BLIP-2's performance improved significantly compared to BLIP, underscoring the potential of image-to-video transfer learning and parameter-efficient tuning techniques for complex video understanding.

6) Generalization on TVQA: Table IX evaluates MCG's generalization performance in multi-modal television scenarios with the TVQA [39] dataset. To emphasize the model's understanding of inter-video elements and the reasoning behind their correlations with questions, MCG intentionally excludes the utilization of television subtitles or audio information, focusing solely on visual content to derive answers. To ensure a fair comparison, methods utilizing additional multi-modal television inputs are grayed out. The experimental results reveal several key insights. First, the existing SoTA FrozenBiLM exhibits a significant 24.5% performance gap between scenarios with and without speech input. This observation suggests that the model's comprehension is influenced by factors beyond visual content alone, possibly including cues from speech. When comparing MCG with the leading FrozenBiLM model, MCG exhibits a promising 2.6% performance gain. This comparison signifies MCG's robustness in capturing intricate inter-modal relationships and reasoning within televisions. Moreover, focusing solely on video visual information, MCG attains an impressive 59.8% accuracy, closely approaching the human accuracy of 61.9%. This experimental achievement illuminates MCG's certain ability to generalize across film and television scenarios.

7) Generalization on CLEVRER: Table X investigates MCG's generalization ability in synthetic object-oriented videos with the diagnostic CLEVRER [32] dataset. We gray out object-centric or neural-symbolic reasoning approaches that incorporate symbolic program execution bridging video concept acquisition and language semantic parsing. We observed that MCG outperforms existing VideoQA methods, demonstrating a notable 11.5% increase in overall per-task accuracy. In open-ended descriptive tasks, MCG shows a significant 5.9% improvement, underscoring its ability to interpret video content and engage in temporal reasoning. Additionally, MCG displays strong performance in explanatory and predictive tasks with improvements of 13.2% and 16.1%, respectively, indicating its certain ability in causal inference and future event anticipation.

Fig. 8. Visulization of question answering results from the proposed MCG for both open-end tasks and multi-choice tasks. The top two rows display four examples from ActivityNet-QA targeting open-end QA, including spatial relationship, temporal relationship, motion, and free tasks. The bottom half illustrates the multi-choice instances from NExT-QA, covering casual, temporal, and description tasks. Correct answers and Ground Truth (GT) are highlighted in green, while incorrect responses are marked in red.

Yet, when compared with the ALOE [34] model, MCG's limitations become apparent, particularly in the counterfactual domain where ALOE's discrete object processing strategy prevails. ALOE's strengths mainly stem from its focused object-centric representations, fine-grained attention, along with self-supervised dynamics learning, which are particularly effective for the synthetic, structured object-oriented scenes where understanding the causal relationships and interactions between objects is crucial. In contrast, taking raw videos as input, MCG's sparse sampling approach with spatial-temporal blocks may overlook essential details pivotal for the comprehensive question-answering tasks that CLEVRER presents. Moreover, the cross-modal general knowledge that MCG acquired through large-scale pretraining on noisy video-text pairs might not transfer as effectively to the physical object-specific scenarios. These insights encourage future works toward evolving a more generalized model with interpretability and logical causal robustness behind question-answering ratiocination.

D. Qualitative Results

Figure 8 illustrates typical examples to qualitatively investigate the performance of the proposed MCG on both open-ended and multi-choice tasks. The top four present open-ended instances from ActivityNet-QA, including spatial relationship, temporal relationship, motion, and free tasks. To provide a comprehensive analysis of MCG's effectiveness and drawbacks, we combine the competitive SOTA model ALPRO (leveraging the same magnitude of external training data with MCG) for joint analysis. According to the results, we draw several important observations: 1) Both ALPRO

TABLE IX
EXPERIMENTAL COMPARATIVE RESULTS OF THE PROPOSED MCG AND EXISTING SOTAS ON TVQA. OVERALL ACCURACY IS LISTED AS A PERCENTAGE (%). MODELS THAT TAKE EXTRA MULTI-MODAL TELEVISION INPUTS ARE GRAYED OUT. THE BEST OUTCOMES ARE EMPHASIZED IN BOLD

Method	Accuracy
w/ multi-modal television inputs	
Human [39]	89.4
HCRN [2]	71.3
HERO [50]	73.6
FrozenBiLM [72]	82.0
w/ only visual inputs	
Human [39]	61.9
ALPRO†[16]	57.2
FrozenBiLM [72] w/o speech	57.5
MCG	**59.8**

TABLE X
EXPERIMENTAL COMPARATIVE RESULTS OF THE PROPOSED MCG AND EXISTING SOTAS ON CLEVRER BENCHMARK. ALL RECORDS ARE LISTED AS A PERCENTAGE (%). DESCRIP., EXPLA., PREDICT., AND COUNT. REPRESENT DESCRIPTIVE, EXPLANATORY, PREDICTIVE, AND COUNTERFACTUAL TASKS, RESPECTIVELY. NEURAL-SYMBOLIC AND OBJECT-CENTRIC METHODS ARE GRAYED OUT

Method	Descrip.	Expla.	Predict.	Count.	Overall
Neural-Symbolic or Object-Centric Methods					
IEP(V) [74]	52.8	14.5	9.7	3.8	20.2
NS-DR [32]	88.1	79.6	68.7	42.4	69.7
DCL-O [33]	91.4	82.0	82.1	46.9	75.6
ALOE [34]	94.0	96.0	87.5	75.6	88.3
VRDP [36]	93.4	91.9	91.4	84.3	90.3
Q-type	29.2	8.1	25.5	10.3	18.3
HME [22]	54.7	13.9	33.1	7.0	27.2
HCRN [2]	55.7	21.0	21.0	**11.5**	27.3
MCG	**61.6**	**34.2**	**49.2**	10.5	**38.8**

and MCG demonstrate the ability to comprehend questions of varying complexity and locate even small visual details, such as the watch in the first case and the nail in the second case. 2) MCG exhibits an advantage in generating longer answers, while ALPRO remains to face challenges in generating multi-word answers, as evident in the second and fourth cases. 3) The failure case highlights that MCG may still struggle to handle why-type questions or ambiguous answers. These observations further demonstrate that the difficulty of open-ended QA stems from question-answering reasoning and answer-generation challenges.

Furthermore, we present typical multi-choice examples from NExT-QA in the bottom half of Figure 8 to illustrate MCG's effectiveness and causal reasoning capabilities. Our observations from these visualizations are as follows: 1) MCG exhibits its potential to deliver reliable answers, even in the face of complex causal questions (fifth instance) or questions requiring common-sense comprehension (sixth instance). 2) For the tough description question with substantial background interference (seventh instance), our model demonstrates robustness by correctly distinguishing a tiny blue object from a sizeable yellow background. 3) As for the failure case, we find that both ALPRO and the proposed MCG inevitably suffer the unobservable and elusive confounding effect misled by the data bias, which leads the model to focus on the spurious correlations, e.g., connecting "two ladies" with "pose for photographs" in the last case. This phenomenon indicates the necessity of causality study in video question answering.

V. CONCLUSION AND FUTURE WORK

We introduce MCG, an end-to-end multi-granularity contrastive cross-modal collaborative generative model for long-term VideoQA. MCG employs joint unimodal modeling to derive discriminative representations possessing high visual concepts and leverages a novel multi-granularity contrastive learning strategy to harness the intrinsically explicitly exhibited semantic correspondences. At its core, MCG formulates VideoQA as a generative task to reconcile existing discrepancies in VideoQA task formulations with a cross-modal collaborative generation module. It empowers MCG with the capability for cross-modal high-semantic fusion and generation to rationalize and answer, ensuring a more intuitive and effective approach to generating answers. MCG sets new benchmarks on four VideoQA datasets and shows strong generalization across diverse tasks.

Limitations and Future Directions: While MCG demonstrates robust performance in analyzing natural scenes, it encounters difficulties in synthetic scenes characterized by simple visuals yet complex dynamics, such as those found in the CLEVRER dataset. This challenge points to a broader need for enhanced video understanding and causal relationship reasoning. Moving forward, our future works will focus on developing a more generalized model with enhanced interpretability and logical causal robustness behind question-answering ratiocination. Moreover, we plan to leverage parameter-efficient tuning techniques to maximize VideoQA performance without proportionally increasing computational demands and explore transfer capabilities from large-scale foundation models to dynamic video scenarios, thereby enhancing the model's ability to interpret complex temporal relationships.

REFERENCES

[1] M. Gu, Z. Zhao, W. Jin, R. Hong, and F. Wu, "Graph-based multi-interaction network for video question answering," *IEEE Trans. Image Process.*, vol. 30, pp. 2758–2770, 2021.

[2] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9972–9981.

[3] Y. Liu, X. Zhang, F. Huang, B. Zhang, and Z. Li, "Cross-attentional spatio-temporal semantic graph networks for video question answering," *IEEE Trans. Image Process.*, vol. 31, pp. 1684–1696, 2022.

[4] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6576–6585.

[5] P. Zeng, H. Zhang, L. Gao, J. Song, and H. T. Shen, "Video question answering with prior knowledge and object-sensitive learning," *IEEE Trans. Image Process.*, vol. 31, pp. 5936–5948, 2022.

[6] W. Jin, Z. Zhao, X. Cao, J. Zhu, X. He, and Y. Zhuang, "Adaptive spatio-temporal graph enhanced vision-language representation for video QA," *IEEE Trans. Image Process.*, vol. 30, pp. 5477–5489, 2021.

[7] Z. Zhao et al., "Long-form video question answering via dynamic hierarchical reinforced networks," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5939–5952, Dec. 2019.

[8] Y. Zhuang et al., "Multichannel attention refinement for video question answering," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–23, Jan. 2020.

[9] M. Peng, C. Wang, Y. Gao, Y. Shi, and X.-D. Zhou, "Multilevel hierarchical network with multiscale sampling for video question answering," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1276–1282.

[10] T. Yu, J. Yu, Z. Yu, and D. Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Trans. Image Process.*, vol. 29, pp. 1204–1218, 2020.

[11] J. Xiao, P. Zhou, T.-S. Chua, and S. Yan, "Video graph transformer for video question answering," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 39–58.

[12] T. Yu, J. Yu, Z. Yu, Q. Huang, and Q. Tian, "Long-term video question answering via multimodal hierarchical memory attentive networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 931–944, Mar. 2021.

[13] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[14] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1728–1738.

[15] J. Lei et al., "Less is more: CLIPBERT for video-and-language learning via sparse sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7331–7341.

[16] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. H. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4953–4963.

[17] W. Yu et al., "Learning from inside: Self-driven Siamese sampling and reasoning for video question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26462–26474.

[18] K.-H. Zeng et al., "Leveraging video descriptions to learn video question answering," in *Proc. AAAI*, 2017, pp. 4334–4340.

[19] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 4534–4542.

[20] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical dual-level attention network learning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1050–1058.

[21] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: Toward spatio-temporal reasoning in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1359–1367.

[22] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1999–2007.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[24] X. Li et al., "Beyond RNNs: Positional self-attention with co-attention for video question answering," in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 8, 2019, pp. 8658–8665.

[25] M. Peng, C. Wang, Y. Gao, Y. Shi, and X.-D. Zhou, "Temporal pyramid transformer with multimodal interaction for video question answering," 2021, *arXiv:2109.04735*.

[26] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11109–11116.

[27] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proc. AAAI*, 2020, pp. 11021–11028.

[28] M. Gandhi, M. O. Gul, E. Prakash, M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Measuring compositional consistency for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5036–5045.

[29] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Invariant grounding for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2928–2937.

[30] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–28.

[31] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1039–1050.

[32] K. Yi et al., "CLEVRER: Collision events for video representation and reasoning," in *Proc. ICLR*, 2020, pp. 1–19.

[33] Z. Chen et al., "Grounding physical concepts of objects and events through dynamic visual reasoning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–20.

[34] D. Ding, F. Hill, A. Santoro, M. Reynolds, and M. Botvinick, "Attention over learned object embeddings enables complex visual reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 9112–9124.

[35] C. P. Burgess et al., "MONet: Unsupervised scene decomposition and representation," 2019, *arXiv:1901.11390*.

[36] M. Ding, Z. Chen, T. Du, P. Luo, J. B. Tenenbaum, and C. Gan, "Dynamic visual reasoning by learning differentiable physics models from video and language," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 887–899.

[37] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, and T.-S. Chua, "Video question answering: Datasets, algorithms and challenges," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 6439–6455.

[38] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: Understanding stories in movies through question-answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4631–4640.

[39] J. Lei, L. Yu, M. Bansal, and T. Berg, "TVQA: Localized, compositional video question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1369–1379.

[40] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering," 2019, *arXiv:1904.11574*.

[41] Z. Yu et al., "ActivityNet-QA: A dataset for understanding complex web videos via question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9127–9134.

[42] Z. Zhao et al., "Open-ended long-form video question answering via adaptive hierarchical reinforced networks," in *Proc. Twenty-Seventh Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 8.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 4489–4497.

[46] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[47] K. Lin et al., "SwinBERT: End-to-end transformers with sparse attention for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17949–17958.

[48] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7464–7473.

[49] L. Zhu and Y. Yang, "ActBERT: Learning global-local video-text representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8746–8755.

[50] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "HERO: Hierarchical encoder for video+language omni-representation pre-training," 2020, *arXiv:2005.00200*.

[51] H. Luo et al., "UniVL: A unified video and language pre-training model for multimodal understanding and generation," 2020, *arXiv:2002.06353*.

[52] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9879–9889.

[53] M. Patrick et al., "Support-set bottlenecks for video-text representation learning," in *Proc. ICLR*, 2021, pp. 1–18.

[54] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. ICML*, 2022, pp. 12888–12900.

[55] J. Wang et al., "OmniVL: One foundation model for image-language and video-language tasks," 2022, *arXiv:2209.07526*.

[56] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021, *arXiv:2102.05095*.

[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[58] Z. Wang, J. Yu, A. Wei Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," 2021, *arXiv:2108.10904*.
[59] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "NExT-QA: Next phase of question-answering to explaining temporal actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9772–9781.
[60] D. Xu et al., "Video question answering via gradually refined attention over appearance and motion," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1645–1653.
[61] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[62] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.
[63] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.
[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[65] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014.
[66] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, and G. Kim, "Video question answering with spatio-temporal reasoning," *Int. J. Comput. Vis.*, vol. 127, no. 10, pp. 1385–1412, Oct. 2019.
[67] T. Yang, Z.-J. Zha, H. Xie, M. Wang, and H. Zhang, "Question-aware tube-switch network for video question answering," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1184–1192.
[68] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2021, pp. 1686–1697.
[69] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the 'video' in video-language understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2917–2927.
[70] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.
[71] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, Feb. 2017.
[72] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-shot video question answering via frozen bidirectional language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 124–141.
[73] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.
[74] J. Johnson et al., "Inferring and executing programs for visual reasoning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3008–3017.
[75] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23716–23736.
[76] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

Ting Yu received the M.Sc. degree from Zhejiang University, Zhejiang, China, in 2013, and the Ph.D. degree from the School of Computer Science and Technology, Hangzhou Dianzi University, Zhejiang, in 2021. She is currently an Associate Professor with the School of Information Science and Technology, Hangzhou Normal University, Zhejiang. She has published high-quality papers in prestigious journals and at top conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, AAAI, and CVPR. Her research interests include cross-media analysis, multimodal learning, and trustworthy visual-language reasoning.

Kunhao Fu is currently pursuing the M.Eng. degree with the School of Information Science and Technology, Hangzhou Normal University, Zhejiang, China. His research interests include multimodal learning, prompt learning, and video question answering.

Jian Zhang received the Ph.D. degree from Zhejiang University, China. From 2009 to 2011, he was with the Department of Mathematics, Zhejiang University, as a Postdoctoral Research Fellow. Since 2016, he has been doing research on machine learning with Simon Fraser University (SFU) as a Visiting Scholar. He is currently a Professor with the School of Information Science and Technology, Hangzhou Normal University. His research interests include multimodal learning and medical image analysis. He serves as a reviewer for several prestigious journals in his research domain.

Qingming Huang (Fellow, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, China, in 1988 and 1994, respectively. He is currently a Professor with the University of Chinese Academy of Sciences and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has authored or coauthored more than 400 academic papers in prestigious international journals and top-level international conferences. His research interests include multimedia computing, image processing, computer vision, and pattern recognition. He has served as the General Chair, the Program Chair, the Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, and PSIVT. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Acta Automatica Sinica*. He is a reviewer of various international journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON MULTIMEDIA.

Jun Yu (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China. He was an Associate Professor with the School of Information Science and Technology, Xiamen University, Xiamen, China. From 2009 to 2011, he was with Nanyang Technological University, Singapore. From 2012 to 2013, he was a Visiting Researcher with Microsoft Research Asia (MSRA). He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. He has authored or coauthored more than 100 scientific articles. Over the past years, his research interests include multimedia analysis, machine learning, and image processing. He has served as a Program Committee Member for top conferences, including CVPR, ACM MM, AAAI, and IJCAI. In 2017, he received the IEEE SPS Best Paper Award. He has (co-)chaired several special sessions, invited sessions, and workshops. He has served as an Associate Editor for prestigious journals, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Pattern Recognition*.