Rethinking Multi-Modal Alignment in Multi-Choice VideoQA from Feature and Sample Perspectives

Shaoning Xiao[♡], **Long Chen**[♠,*], **Kaifeng Gao**[♡], **Zhao Wang**[♡], **Yi Yang**[♡], **Zhimeng Zhang**[♡], **and Jun Xiao**[♡]
[♡]Zhejiang University [♠]Columbia University
{shaoningx,kite_phone,zhao_wang,yangyics,zhimeng,junx}@zju.edu.cn,
zjuchenlong@gmail.com

Abstract

Reasoning about causal and temporal event relations in videos is a new destination of Video Question Answering (VideoQA). The major stumbling block to achieve this purpose is the semantic gap between language and video since they are at different levels of abstraction. Existing efforts mainly focus on designing sophisticated architectures while utilizing frame- or object-level visual representations. In this paper, we reconsider the multi-modal alignment in VideoQA from feature and sample perspectives to achieve better performance. From the view of feature, we break down the video into trajectories and first leverage trajectory feature in VideoQA to enhance the alignment between two modalities. Moreover, we adopt a heterogeneous graph architecture and design a hierarchical framework to align both trajectory-level and frame-level visual feature with language feature. In addition, we found that VideoQA models are largely dependent on language priors and always neglect visual-language interactions. Thus, two effective yet portable training augmentation strategies are designed to strengthen the cross-modal correspondence ability of our model from the view of sample. Extensive results show that our method outperforms all state-of-the-art models on the challenging NExT-QA benchmark.

1 Introduction

Given a video and a question about its content, Video Question Answering (VideoQA) aims to answer the question through multi-modal reasoning. Since it can benefit numerous multi-modal applications such as video retrieval (Zhang et al., 2021; Xiao et al., 2021c,b) and interaction with robot vision, VideoQA has received increasing attention in recent years. Compared to Image Question Answering (ImageQA) (Antol et al., 2015), VideoQA requires more complex reasoning. A naive extension of ImageQA methods (Chen et al., 2020, 2022) may not apply to the problem of VideoQA due to the extra information such as temporal object interactions.

Figure 1: An illustration of VideoQA. Existing VideoQA models utilize objects as finer clues. We further track the objects with same classes across time and first leverage video trajectory for better alignment.

As computing systems are more frequently intervening to improve people's work and daily lives, it is critical for machines to correctly comprehend the complex relationships between events such as causal and temporal relations. However, conventional VideoQA benchmarks (Xu et al., 2017; Yu et al., 2019) are constituted by descriptive questions, *e.g.*, asking the number, color or single action of the video elements. Recently, a benchmark called NExT-QA (Xiao et al., 2021a) was elaborately designed to challenge VideoQA models to reason about causal and temporal actions and understand the rich object interactions in daily activities. Instead of a fixed answer set, NExT-QA requires the models to pick the correct answer out of five candidates. Since this multi-choice setting allows concatenating question-answer pair to form a holistic language query and aligning it with videos, it can be easily generalized to any other multi-modal tasks besides question answering.

*Long Chen is the corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8188–8198
December 7-11, 2022

Multi-modal alignment, namely finding the correspondences between two modalities, is the foundation of multi-modal tasks. One of the major challenges to align two modalities is to fill in the semantic gap between language and video due to different levels of abstraction. Intuitively, the key to better alignment is to identify the direct relations between sub-elements (*e.g.*, words, frames, *etc.*) of instances from two different modalities. As a highly abstract vehicle of expression, almost every word in a natural language sentence can express a complete meaning, *e.g.*, an entity, an action or a time. In contrast, a video also contains a number of visual concepts but they are naturally indivisible unlike the word. Primitive VideoQA models (Li et al., 2019; Jang et al., 2017) utilize frame-level appearance feature as video representation and align it with language feature. These methods simply regard the video as a series of frames. To better capture dynamic change in videos, others (Gao et al., 2019; Fan et al., 2019; Jiang and Han, 2020) incorporate frame-level motion feature together with appearance feature. Recently, object feature as more granular information has been leveraged to strengthen the semantic correspondence ability (Huang et al., 2020; Le et al., 2020; Xiao et al., 2022). Although bringing in object information results in a better ability to reason about complex interactions in videos, there are still two unsolved problems: (1) Static object information is hard to model temporal-related relations. (2) Same object may take different actions during time and different objects with the same label can cause a mismatch. As shown in Figure 1, there are two boys and two balls in the video and both boys hold a ball. In this case, VideoQA models that only use object-level information may end up with misalignment, which leads to a wrong answer. Therefore, tracking objects across time is of vital importance.

In this paper, we explore the multi-modal alignment in VideoQA from feature and sample perspectives to better reason over videos' causal/temporal actions. From feature perspective, we decouple trajectories as salient entities from video and first leverage video trajectory feature in VideoQA. In order to model the rich interactions between trajectories, we propose a trajectory encoder using multi-head self-attention with temporal and semantic embeddings. Video trajectories are the essential ingredient for video relation detection task (Qian et al., 2019; Xie et al., 2020), which require tracking the same object from different frames along the temporal axis. Specifically, we first apply a pre-trained object detector to obtain bounding boxes. Then, an association algorithm named improved sequence NMS (Xie et al., 2020) is applied to obtain video trajectories that contain spatial-temporal information of the visual elements. We further align trajectory-level and frame-level feature with language feature by a cycle-attention module and adopt a heterogeneous graph architecture for implicit relation reasoning.

In addition, from sample perspective, we design two training strategies in order to enhance multi-modal alignment in feature space. To be specific, we first increase negative candidate answers when computing the matching score. This strategy forces the model to focus on the discriminative regions within a question-answer pair. We then add negative question-answer pairs that are attached to other videos. By doing so, the video and its affiliated language are drawn closer in feature space and the mismatched pairs are pulled away. Moreover, we found that these strategies can also solve the problem that VideoQA models are largely dependent on language priors and neglect visual-language interactions. Together with the proposed model, our method achieved state-of-the-art performance.

In summary, the main contributions of our work are listed as follows: (1) We first leverage video trajectory features in VideoQA to capture richer causal and temporal relations in the video. (2) We design two training strategies to strengthen the cross-modal correspondence ability of our model and further boost the performance. (3) We conducted extensive experiments on NExT-QA and the results demonstrate the effectiveness of our model.

2 RELATED WORK

2.1 Video Question Answering

We roughly summarise three kinds of VideoQA methods according to their utilized techniques, namely attention-based, memory-based, and graph-based models. Attention mechanism (Jang et al., 2017; Ye et al., 2017; Gao et al., 2019; Li et al., 2019; Jiang et al., 2020) is widely used in VideoQA. Jang et al. (2017) propose a dual-layer LSTM with spatial and temporal attention. Li et al. (2019) use the self-attention mechanism to encode each modality and utilize co-attention mechanism for alignment. Jiang et al. (2020) divide the semantic features generated from question into the spatial

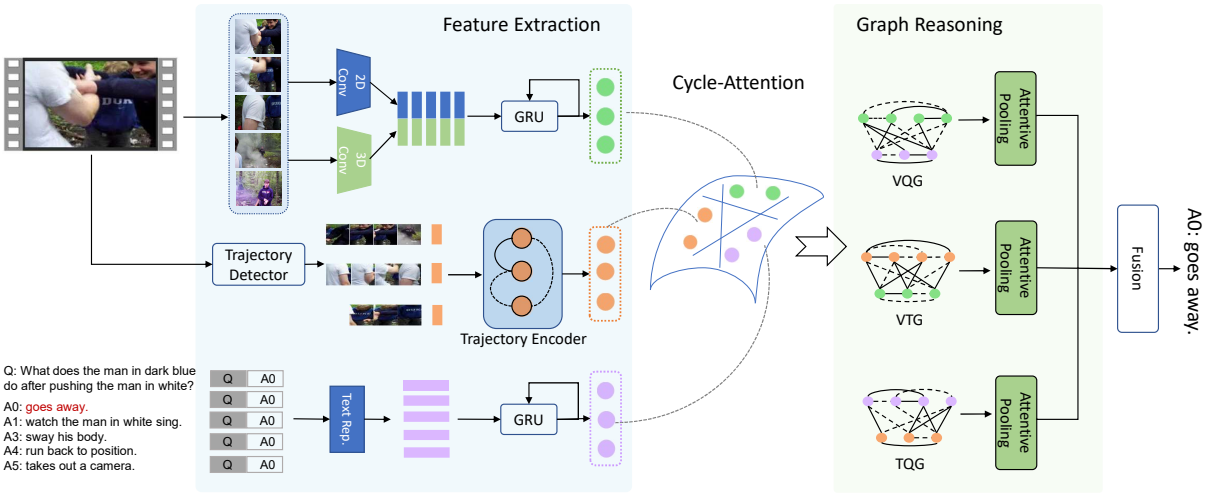

Figure 2: The overview of our model architecture for VideoQA. Firstly, the frame-level features, trajectory-level features and text representations are extracted. Then, the visual and language features are aligned in pairs by a cycle-attention module. At last, heterogeneous graphs are constructed and applied for reasoning.

part and the temporal part which guide the spatial and temporal attention of video, respectively. Memory based approaches (Xu et al., 2017; Gao et al., 2018; Fan et al., 2019) encode the input sources multiple cycles and use attention mechanism allowing the model to focus on different contents in each cycle. Xu et al. (2017) gradually refine the attention over the appearance and motion features of the video using the question as guidance. Gao et al. (2018) propose a co-memory attention module to extract useful cues from both appearance and motion memories to generate attention for motion and appearance separately. Fan et al. (2019) propose a read-write memory network that jointly encode the movie appearance and caption content.

Graph based models (Hu et al., 2019; Huang et al., 2020; Park et al., 2021; Jiang and Han, 2020; Xiao et al., 2022) advance the field by exploiting the ability of relation reasoning. Jiang and Han (2020) propose a heterogeneous graph alignment network to align and interact the inter- and intra-modality. Park et al. (2021) perform relation reasoning between appearance and motion information of the video with compositional semantics of the question. Although these methods achieve impressive performance using the graph structure, they only utilize frame-level video feature for alignment and thus suffer from a lack of fine-grained interaction. Some other methods leverage object information to enhance the fine-grained alignment. Hu et al. (2019) propose a graph network where each node represents an object, and conduct iterative message passing conditioned on the textual input. Huang et al. (2020) propose to represent video as a location aware graph and conduct graph convolution. Xiao et al. (2022) model video as a conditional hierarchical graph to align the video facts and textual cues on different levels. Instead of designing complicated models, we solve VideoQA task by leveraging a new trajectory feature and further boost performance from sample perspective.

2.2 Video Trajectory Detection

Video trajectory detection, as an essential component of video relation detection task (Qian et al., 2019; Xie et al., 2020; Gao et al., 2022, 2021), has attracted more and more attention. Detection of objects in static image has gained a great improvement in last few years. However, video trajectory detection is still a tough problem since it needs to tracking same object in different frames of a video clip. A popular scheme is tracking-by-detection, namely applying detection algorithm to each video frame and the detections are associated across frames to form trajectories. Seq-NMS (Han et al., 2016) takes detections from a state-of-the-art object detection method and associates over time by finding the highest scoring path. Improved Seq-NMS (Xie et al., 2020) improves seq-NMS by introducing a new linking mechanism to solve the missing connection problem caused by violent object movement. In this paper, we detect static objects using a pre-trained detector and utilize improved seq-NMS as trajectory tracking method to generate trajectories.

3 Approach

Formally, suppose we have a video $V = \{v_t\}_{t=1}^{T}$ which contains T frames and v_t denotes the t-th frame. Meanwhile, we have a natural language question $Q = \{w_l\}_{l=1}^{L}$, where w_l denotes the l-th word in the sentence and L represents the question length. VideoQA aims to predict the correct answer A^p to the question according to the relevant video content. In the multi-choice setting, the goal is to choose the correct answer A^p from n candidate answer set $S_A = \{A_1, A_2, ..., A_n\}$.

In this section, we sequentially introduce each component of our proposed model. The video and language encoding procedures are presented in Section 3.1. The alignment and reasoning modules are introduced in Section 3.2. The answer predictor is introduced in Section 3.3. In Section 3.4, we introduce two sample augmentation strategies.

3.1 Feature Encoding

3.1.1 Video Representations

We utilize both frame-level and trajectory-level video features for video representation since they naturally share complementary information.

Frame-level Features. Following previous works, we uniformly sample a fixed number N of clips for each video. We use a 2D ConvNet to extract video appearance feature $\boldsymbol{F_a} = \{f_i^a\}_{i=1}^{N}$, where $f_i^a \in \mathbb{R}^{d_a}$ and use a 3D ConvNet to extract video motion feature $\boldsymbol{F_m} = \{f_i^m\}_{i=1}^{N}$, where $f_i^m \in \mathbb{R}^{d_m}$.

Then, we apply a concatenation operation for the appearance and the motion feature with a fully-connected layer to obtain frame-level video feature, $\boldsymbol{F}_v = ReLU(FC([\boldsymbol{F}_a, \boldsymbol{F}_m]))$, where $\boldsymbol{F}_v \in \mathbb{R}^{N \times d}$ and $[\cdot]$ represent the concatenation operation along the feature dimension. Due to the temporal property of videos, we adopt a Gated Recurrent Units (Cho et al., 2014) to process the frame-level video feature, $\boldsymbol{V} = GRU(\boldsymbol{F}_v)$, where $\boldsymbol{V} \in \mathbb{R}^{N \times d}$ is the contextualized frame-level video features.

Trajectory-level Features. As mentioned before, we argue that the video and question are at different abstract levels due to their sub-components, *i.e.*, words and frames contain inconsistent semantic information. Thus, we utilize video trajectory feature to supplement the frame-level feature in order to enhance the feature alignment with language.

We take the tracking-by-detection strategy to generate video trajectories. We first sample the video and detect the objects from all the frames. To generate trajectories, we use improved seq-NMS (Xie et al., 2020) to associate bounding boxes along the time that belong to same object based on object detection results. Specifically, this algorithm links the bounding boxes that likely belong to the same object from consecutive frames to build a graph and it applies dynamic programming to repeatedly pick the path with the highest score. Then, we obtain a series of trajectories each of which contains a set of boxes, a predicted label and the start-end time points. For each trajectory, we apply average pooling to the associated objects features and normalize the start-end time with respect to the video length. To take advantage of semantic information, we project the trajectory label to semantic space using GloVe embeddings (Pennington et al., 2014). Thus, we obtain the visual feature t_v, semantic feature t_l and temporal position embedding t_p for each trajectory. Then, we project these three representations to the same space by fully-connected layers and add them together to get the final trajectory feature $tr_i \in \mathbb{R}^d$, as showed in Figure 3.

Given several trajectory features $\boldsymbol{F}_{tr} = \{tr_i\}_{i=1}^{N_t}$ in a video, where N_t is unequal for different videos, we employ a trajectory encoder with multi-head self-attention to model the rich trajectory-level interaction, $\boldsymbol{T} = MHSA(\boldsymbol{F}_{tr})$, where $\boldsymbol{T} \in \mathbb{R}^{N_t \times d}$ is the refined trajectory feature. As illustrated in Figure 3, our trajectory encoder consists of several multi-head self attention layers (Vaswani et al., 2017) and feed-forward layers with skip connection.

3.1.2 Language Representations

As for language, we use both GloVe features (Pennington et al., 2014) and fine-tuned BERT features (Devlin et al., 2019) in different experimental settings. A vocabulary set was pre-defined which is composed of top K most frequent words. For experiments with GloVe, each word in the set is initialized with word-level pre-trained GloVe representations. Following NExT-QA, we also use fine-tuned BERT feature which fine-tunes regular BERT on the dataset by maximizing the correct QA pairs' probability in each multi-choice QA. We extract token-wise sentence-level BERT features for each question-answer pair. For multi-choice setting, we concatenate the question Q with each candidate answer A_i to form a holistic *query*. In order to obtain well contextualized language representation, we apply another GRU to the word embeddings in the query feature $\boldsymbol{F}_q$, $\boldsymbol{Q}, \boldsymbol{F}_q^{global} = GRU(\boldsymbol{F}_q)$, where $\boldsymbol{Q} \in \mathbb{R}^{L \times d}$ and $\boldsymbol{F}_q^{global} \in \mathbb{R}^d$ is the global sentence

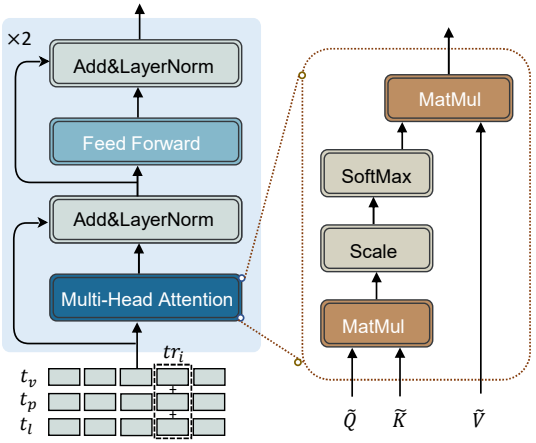

Figure 3: The architecture of our trajectory encoder. Right part is an illustration of dot-product attention.

feature from last hidden state.

3.2 Feature Alignment and Reasoning

Alignment. For a better alignment among language features, frame-level video features and trajectory features, we propose a cycle-attention module that aligns different features in a circular pattern. Firstly, we align the trajectory with the language feature,

$$\boldsymbol{Q}_{tq} = \underset{q\to t}{Atten}(\boldsymbol{Q},\boldsymbol{T},\boldsymbol{T}), \quad \boldsymbol{T}_{tq} = \underset{t\to q}{Atten}(\boldsymbol{T},\boldsymbol{Q},\boldsymbol{Q}), \quad (1)$$

where "$\to$" means "attend to". The $Atten$ operation is introduced in Appendix A. Then, we align the frame-level video feature with language feature,

$$\boldsymbol{Q}_{vq} = \underset{q\to v}{Atten}(\boldsymbol{Q},\boldsymbol{V},\boldsymbol{V}), \quad \boldsymbol{V}_{vq} = \underset{v\to q}{Atten}(\boldsymbol{V},\boldsymbol{Q},\boldsymbol{Q}). \quad (2)$$

We argue that the frame-level video feature and trajectory feature are complementary, since the trajectory feature decouples salient entities from the whole video and the frame-level video feature contains contextual information. In addition, there is a natural correspondence relationship between them that trajectory is made up of objects from frames. Thus, we also align them together,

$$\boldsymbol{T}_{vt} = \underset{t\to v}{Atten}(\boldsymbol{T},\boldsymbol{V},\boldsymbol{V}), \quad \boldsymbol{V}_{vt} = \underset{v\to t}{Atten}(\boldsymbol{V},\boldsymbol{T},\boldsymbol{T}). \quad (3)$$

Reasoning. After obtaining the aligned features, we conduct heterogeneous graphs, namely TQG, VQG and VTG as shown in Figure 2, for further reasoning. Taking the trajectory and question graph TQG as an example, we describe the details of our reasoning module. The nodes representations $\boldsymbol{X}_{tq}$ of TQG are the concatenation of token-wise language embeddings $\boldsymbol{Q}_{tq}$ and trajectories features $\boldsymbol{T}_{tq}$. Thus, each node either represents a word or a trajectory. We first calculate the value of graph edges represented by an adjacency matrix,

$$A_{tq} = softmax(f_{W_p}(\boldsymbol{X}_{tq}) f_{W_p}(\boldsymbol{X}_{tq})^T) + I, \quad (4)$$

where f_{W_p} denotes non-linear projection with learnable parameters W_p and I is an identity matrix for skip connection. Each element of A_{tq} means the correlation between the i-th and j-th node. Then, we apply graph convolution to aggregate and pass message over the nodes. Here we show a single-layer graph convolution operation,

$$\boldsymbol{X}_{tq}^{(l)} = \sigma(A_{tq}\boldsymbol{X}_{tq}^{(l-1)}W^{(l)}), \quad (5)$$

where l denotes the l-th layer of GCN and σ represents an activation function. To get the final multi-modal representation, we aggregate all the nodes in TQG by weighted pooling with a self-attention,

$$\boldsymbol{X}_{tq}^{final} = \sigma(f_{W_t}(\boldsymbol{X}_{tq}^{(L)}))\boldsymbol{X}_{tq}^{(L)}, \quad (6)$$

where f_{W_t} denotes non-linear transformation with learnable parameters W_t and $\boldsymbol{X}_{tq}^{(L)}$ is the output of the last GCN layer. Similarly, we construct VQG and VTG and conduct graph convolution operations on them to obtain $\boldsymbol{X}_{vq}^{final}$ and $\boldsymbol{X}_{vt}^{final}$.

3.3 Answer Prediction

Following the multi-choice setting in NExT-QA, we regard the VideoQA task as a multi-modal matching problem, which can easily extend to other multi-modal tasks. Specifically, the candidate answers are concatenated to the corresponding questions and the model scores the concatenated sentences based on the similarities to the video.

In order to bring in global semantic information, we enhance the three multi-modal features by fusion with the global query feature $\boldsymbol{F}_q^{global}$. Then, we calculate a score for each candidate question-answer pair using multi-modal compact bilinear fusion (MCB) (Fukui et al., 2016),

$$s_* = MCB(\boldsymbol{X}_*, \boldsymbol{F}_q^{global}), \quad (7)$$

where $*$ denotes tq, vq and vt. Next, we aggregate scores from different brunches by addition,

$$s = s_{vq} + \lambda_1 s_{tq} + \lambda_2 s_{vt}, \quad (8)$$

We adopt a Hinge loss that can maximize the margins between the correct and incorrect QA pairs,

$$L = \sum_{i=1}^{n-1} max(0, 1 + s_i^- - s^+), \quad (9)$$

where n is the number of candidate answers and s^+ and s^- represent the positive and negative samples.

Methods	Causal	Temp.	Descrip.	Overall
Random	20.52	20.10	19.69	20.08
Text Only	42.62	45.53	43.89	43.76
Text+Viusal	42.46	46.34	45.82	44.24
HGA	49.53	50.74	59.33	49.74
Human	87.61	88.56	90.40	88.38

Table 1: Some VideoQA baselines on NextQA.

3.4 Sample Augmentation

Although fine-tuned BERT features achieve remarkable results, it brings new problem that models answer the question excessively rely on the prior of the question-answer pairs without considering the video content. It is mainly because the fine-tuning goal is to maximize the probability of the correct QA pair in all the multi-choice QA pair candidates. A blind version of VideoQA model was studied by NExT-QA (Xiao et al., 2021a) which only considers the question-answer pairs and totally ignores the video inputs. As shown in Table 1, the performance of the Text-Only model is surprisingly comparable to the model incorporating the video information. We argue that the model devotes to estimating the rationality of question-answer combination or just memorizing the frequency of combinations. Recent state-of-the-art model HGA (Jiang and Han, 2020) achieved considerable improvement compared to both Text-Only and Text+Visual models in Table 1, but there is still a huge gap between the state-of-the-art model and human. Thus, although the elaborately designed architectures and features have the capacity of reasoning complex interactions, the models always get inferior results.

Based on this consideration, in order to capitalize on the full potential of the feature and model, we design two effective yet simple sample augmentation ways for better multi-modal alignment and further boost the VideoQA performance. To be specific, we first increase negative candidate answers (a^-) when computing matching score. This strategy forces the model to focus more on the minor difference between the correct question-answer pair and others. Meanwhile, the model can correspondingly focus on the discriminative video content. We then add negative question-answer pairs that are attached to other videos (qa^-). By cooperating with the hinge loss, the video and its affiliated language are drawn closer in feature space and the mismatched pairs are pulled away. In this way, we enhance the multi-modal alignment from a sample perspective. In addition, bringing in new negative samples break the models' excessive dependence on language prior, which partially solved the problem caused by the feature. In practice, we randomly sample M answers/QA pairs affiliated to other videos in the training set as negative samples.

4 EXPERIMENTS

4.1 Experimental Details

Dataset. NExT-QA (Xiao et al., 2021a) is a recently designed challenging VideoQA benchmark which advances video question answering from describing to reasoning. The dataset contains 5,440 videos where 3870 for training, 570 for validation and 1,000 for testing. The videos are selected from the relation dataset VidOR (Shang et al., 2019) which contains natural videos of daily life such as outdoor activities and social scenes. Thus they are richer in objects and interactions. It consists of 47,692 questions where 34,132, 4,996 and 8,564 for training, validation and testing, respectively. Almost half of the questions are causal questions which contain questions starting with "why" and "how", which is a great challenge for VideoQA models to reason about causality. Temporal questions of inferring temporal actions compose 29% of the dataset. Apart from causal and temporal questions, others are descriptive questions that focus on describing attributes, location and main events in videos. For multi-choice task that is to select one out of the five candidate answers, NExT-QA sampled four qualified candidates as distracting answers for each question to enhance the hard negatives. In a word, NExT-QA goes beyond descriptive QA to benchmark causal and temporal action reasoning in realistic videos and is also rich in object interactions. In addition, several recent state-of-the-art methods are examined on it.

Evaluation Metric. We report the accuracy of our model in all experiments which represents the percentage of correctly answered questions.

Implementation Details. For the training process, we set the number of hidden units d to 256. The batch size is set to 64 and Adam optimizer is used for optimization. The learning rate is set to 0.00005 for GloVe setting and 0.0001 for BERT-FT setting, respectively. For better performance, we reduce the learning rate when a metric has stopped improving. The dropout rate is set to 0.3. We set balance factors λ_1 and λ_2 to 0.5 for all the experiments.

We randomly sample 5 negative samples from the training set for each strategy. We utilize both

Methods	Text Rep.	Acc_C			Acc_T			Acc_D				ACC
		Why	How	All	P&N	Present	All	Count	Loc.	Other	All	
EVQA	GloVe	28.38	29.58	28.69	29.82	33.33	31.27	43.50	43.39	38.36	41.44	31.51
PSAC†	GloVe	35.03	29.87	33.68	30.77	35.44	32.69	38.42	71.53	38.03	50.84	36.03
Co-Mem	GloVe	36.12	32.21	35.10	34.04	41.93	37.28	39.55	67.12	40.66	50.45	38.19
ST-VQA	GloVe	37.58	32.50	36.25	33.09	40.87	36.29	45.76	71.53	44.92	55.21	39.21
HGA	GloVe	36.38	33.82	35.71	35.83	42.08	38.40	46.33	70.51	46.56	55.60	39.67
HME	GloVe	39.14	34.70	37.97	34.35	40.57	36.91	41.81	71.86	38.36	51.87	39.79
HCRN	GloVe	39.86	36.90	39.09	37.30	43.89	40.01	42.37	62.03	40.66	49.16	40.95
Ours	GloVe	**43.14**	**39.82**	**42.27**	**40.25**	**47.21**	**43.11**	**46.89**	**74.58**	**52.46**	**59.59**	**45.24**
EVQA	BERT-FT	42.31	42.90	42.46	46.68	45.85	46.34	44.07	46.44	46.23	45.82	44.24
ST-VQA	BERT-FT	45.37	43.05	44.76	44.52	51.73	49.26	43.50	65.42	53.77	55.86	47.94
Co-Mem	BERT-FT	46.15	42.61	45.22	48.16	50.38	49.07	41.81	67.12	51.80	55.34	48.04
HCRN*	BERT-FT	46.99	42.90	45.91	48.16	50.83	49.26	40.68	65.42	49.84	53.67	48.20
HME	BERT-FT	46.52	45.24	46.18	47.52	49.17	48.20	45.20	73.56	51.15	58.30	48.72
HGA	BERT-FT	46.99	44.22	46.26	49.53	52.49	50.74	44.07	72.54	55.41	59.33	49.74
Ours	BERT-FT	**52.81**	**47.44**	**51.40**	**51.11**	**53.70**	**52.17**	**46.89**	**75.25**	**58.03**	**62.03**	**53.30**

Table 2: Performance (%) comparisons of state-of-the-art methods on NExT-QA validation set. The best and the second results are bold and underlined respectively. † means to add motion feature and * means concatenation of question and answer to adapt to BERT representation.

Models	Causal	Temp.	Descrip.	Overall
ST-VQA	45.51	47.57	54.59	47.64
Co-Mem	45.85	50.02	54.38	48.54
HME	46.76	48.89	57.37	49.16
L-GCN	47.82	48.74	56.51	49.54
HGA	48.13	49.08	57.79	50.01
HCRN	47.07	49.27	54.02	48.89
HQ-GAU	49.04	**52.28**	59.43	51.75
Ours	**50.38**	50.88	**61.78**	**52.41**

Table 3: Performance(%) of on NExT-QA test set.

sample strategies for experiments using GloVe embedding and only use the second strategy for BERT-FT. Other details of implementation are given in Appendix B.

4.2 Compared Methods

In Table 2 and Table 3, we compared our model with other state-of-the-art methods on NExT-QA dataset. Among these methods, STVQA (Jang et al., 2017), PSAC (Li et al., 2019) are attention-based methods. Co-Mem (Gao et al., 2019) and HME (Fan et al., 2019) are memory-based methods. L-GCN (Huang et al., 2020), HGA (Jiang and Han, 2020) and HQ-GAU (Xiao et al., 2022) are graph-based methods.

Different from recent elaborately designed complex architectures for VideoQA, we consider the multi-modal alignment from feature and sample perspectives. We simply adopt a heterogeneous graph as the reasoning module and first leverage trajectory feature in VideoQA. We then design two effective yet easy-to-implement sample augmentation methods. Combining both of them, our model achieves the best performance.

Results. The results on NExT-QA validation set and test set are reported in Table 2 and Table 3, respectively. We can observe that our proposed method achieves new state-of-art performance over all kinds of questions. In particular, we observe that our method works well even on causal and temporal questions which require more complicated reasoning, *e.g.*, our method achieves a significant 5.14% absolute improvement on validation set compared to the second result on causal questions and 1.43% on temporal questions. It should be noticed that HGA also utilizes a heterogeneous graph model for alignment and reasoning which indicates that our trajectory-aware graph model with sample augmentation has the advantage to reason causal and temporal questions over others. L-GCN also utilizes a graph network with object feature and our model outperforms it by a large margin on test set as shown in Table 3. Recently proposed HQ-GAU also adopt a powerful hierarchical architecture with multi-granularity video features that leverages finer object interaction. Table 3 shows that our method outperforms HQ-GAU on causal and descriptive questions by 1.34% and 2.53%. For temporal questions, our method gets comparable result with HQ-GAU but is 1.4% lower than it. It is probably because HQ-GAU has a complicated structure with more parameters and adopts a more effective temporal position embedding.

4.3 Ablation Study

In this section, we report the results of ablative experiments with different variants to better investigate our approach. We first analyze the effect of

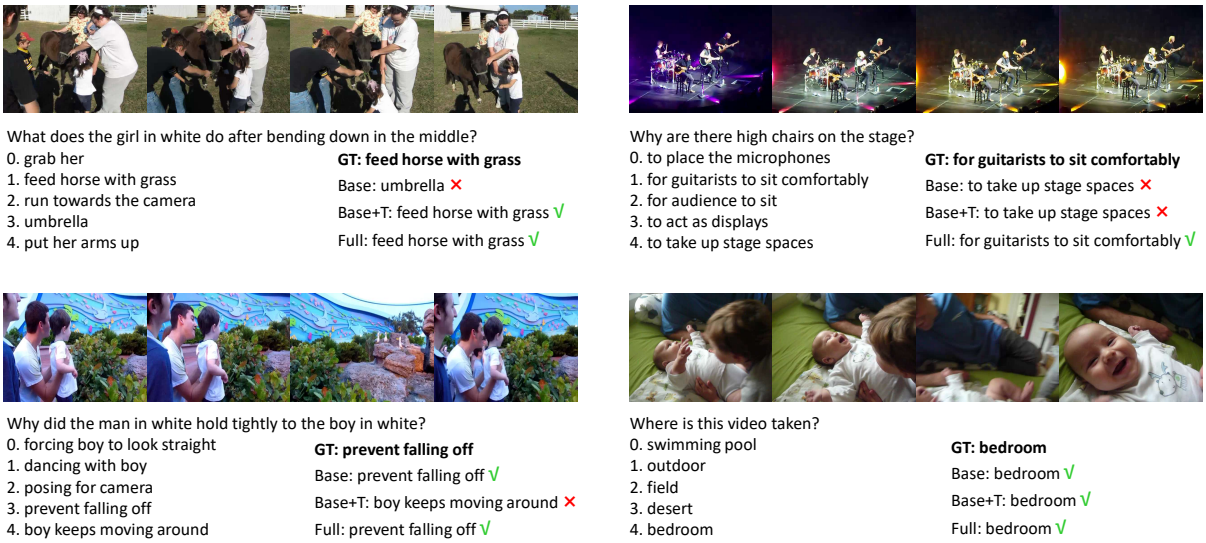

Figure 4: Some qualitative results of our model on NExT-QA validation set. Base: our model without trajectory and sample augmentation. Base+T: Base model with trajectory feature. Full: Base model with trajectory feature and sample augmentation.

frame.	traj.	aug.	Causal	Temp.	Discrip.	Overall
	✓		46.18	48.08	57.27	48.52
✓			46.49	48.76	58.94	49.16
✓	✓		47.18	51.18	59.33	50.36
✓		✓	50.44	51.30	59.85	52.18
✓	✓	✓	51.40	52.17	62.03	53.30

Table 4: Performance (%) on validation set in ablative experiments of trajectory and sample augmentation.

Aug.	Ablation	Causal	Temp.	Descrip.	Overall
No	w/o cyatten.	46.91	48.70	59.33	49.42
	w/o VQG	46.18	48.08	57.27	48.52
	w/o TQG	46.49	48.76	58.94	49.16
	w/o VTG	46.30	50.74	57.27	49.44
	Full	47.18	51.18	59.33	50.36
Yes	traj. GRU	50.63	53.54	60.36	53.08
	traj. MHSA	51.40	52.17	62.03	53.30

Table 5: Performance (%) on validation set in ablative experiments of model components.

trajectory feature and sample augmentation method. Then, we introduce an ablation study conducted on components of our model. All the variants in this section are evaluated on NExT-QA validation set.

Ablation on trajectory. To exploit the effect of the trajectory, we compared the performance of the models with and without trajectory feature in Table 4. By comparing the second line with the third line, we notice that utilizing trajectory feature improves the accuracy by 1.20%. Comparing the last two lines in Table 4, the model using trajectory feature outperforms the other by 1.12% overall accuracy even though the score has already been improved a lot by sample augmentation. These results demonstrate the necessity of employing the trajectory feature. In addition, we give the results of the model only using trajectory feature as visual representation on line 1 in Table 4. The results indicate that the frame-level feature also plays an important role in multi-modal reasoning. The main reason is that frame-level features provide contextual information which some questions heavily rely on.

Strategy	Text Rep.	Causal	Temp.	Descrip.	Overall
none	GloVe	36.86	37.59	54.70	39.87
a^-		41.66	43.61	57.01	44.68
qa^-		40.20	41.07	58.43	43.31
both		42.27	43.11	59.59	45.24
none	BERT-FT	47.18	51.18	59.33	50.36
a^-_*		47.22	49.88	59.33	49.96
qa^-		51.40	52.17	62.03	53.30

Table 6: Comparisons of different sample augmentation strategies.

Ablation on sample augmentation. We analyzed the effectiveness of sample augmentation methods in Table 4. By comparison of line 2 and line 4 (line 3 vs. line 5), we notice an almost 3% absolute overall improvement, which indicates that our augmentation methods can boost the performance. Considering that it's harder to improve in a higher score range, this indicates the trajectory feature and sample augmentation method could promote each other for better multi-modal alignment.

We also explore the different augmentation strategies in Table 6. The a^- represents that

we sample negative answers from other questions and concatenate them with the original question. The qa^- means that we sample negative question-answer pairs attached to other videos. For experiments with GloVe embedding as text representation, we can find that each strategy improved the accuracy by a large margin and using both strategies can further boost the performance. With regards to BERT-FT as text representation, we cannot directly apply strategy a^- because BERT features are sentence-level holistic feature for question-answer pairs where the question parts vary across different pairs. So we averaged the question features as a unified question representation and concatenated randomly sampled answers (a^-_* in Table 6). However, we can observe that the accuracy barely changed. This may be because the average operation harms the integrity of sentence representation especially the sentence matching information that [CLS] embedding contains. Thus we only used the second strategy qa^- for BERT-FT and the performance is improved a lot even so. We also analyzed the effect of the number of negative samples. The performance grows when the number of negative samples increases. When the number is more than 15, the performance would barely change.

Ablation on trajectory encoder. On the bottom section of table 5, we studied the effect of trajectory encoder MHSA. By replacing our MHSA with a GRU with temporal and semantic embedding, the performance drops by 0.77% on causal questions and 1.67% on descriptive questions making a overall 0.22% decline, which demonstrates the global interactions modeling ability of MHSA. For temporal questions, there is a 1.37% improvement which indicates that the RNN architecture is better to capture sequential information.

Ablation on model components. We analyzed the model components on the top part of Table 5. By removing each graph and cycle-attention, performance of the model all dropped. The results demonstrate that all parts of the architecture play an important role in alignment and reasoning.

4.4 Qualitative Analysis

We show some qualitative results on NExT-QA validation set in Figure 4. The results of three models with different configurations are visualized, *i.e.*, Base: the baseline without trajectory feature and sample augmentation, Base+T: add trajectory feature to Base, and Full: our full model with both trajectory and sample augmentation. We notice that Base+T and Full model perform better than Base in most cases, which demonstrates that both feature and strategies are helpful. In the bottom-left case, we surprisingly found that Base+T predicted a wrong answer whereas Base answered correctly. However, the candidate answer 4 "boy keeps moving around" seems hardly to be a wrong answer to the question.

5 CONCLUSION

In this paper, we explored multi-modal alignment in VideoQA from feature and sample perspectives. From the view of feature, we first leverage video trajectory features in VideoQA to bridge the semantic gap between the sub-components of the video and the language. Moreover, in order to better utilize the trajectory feature, we propose a graph-based model which is capable of alignment and reasoning over heterogeneous representations. From the view of sample, we propose two sample augmentation strategies to further enhance the cross-modal correspondence ability of our model. The promising results on challenging NExT-QA dataset have exhibited the causal and temporal reasoning ability of our method. In the future, we will further explore a better way to take advantage of trajectory information considering its significant potential.

Limitations

Although video trajectories are effective on VideoQA and other video understanding tasks, the model is sensitive to the quality of trajectories. The object detection and tracking methods are of vital important to the quality of trajectories. Using a weak tracking method may introduce noises which can be harmful to the performance. The training augmentation strategies are naturally suitable for multi-choice setting, however for open-ended setting, further work needs to be done to adapt.

Acknowledgements

This work was supported by the National Key Research & Development Project of China (2021ZD0110700), the National Natural Science Foundation of China (U19B2043, 61976185), Zhejiang Natural Science Foundation (LR19F020002), Zhejiang Innovation Foundation (2019R52002), and the Fundamental Research Funds for the Central Universities (226-2022-00087).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *ICCV*, pages 2425–2433.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, pages 10800–10809.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *ECCV*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

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

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468.

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

Kaifeng Gao, Long Chen, Yifeng Huang, and Jun Xiao. 2021. Video relation detection via tracklet based visual transformer. In *ACM MM*, pages 4833–4837.

Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, and Jun Xiao. 2022. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *CVPR*, pages 19497–19506.

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. 2019. Structured two-stream attention network for video question answering. In *AAAI*, pages 6391–6398.

Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S. Huang. 2016. Seq-nms for video object detection.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *ICCV*, pages 10293–10302.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 1359–1367.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *AAAI*, pages 11101–11108.

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

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset.

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

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

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering. In *CVPR*, pages 15526–15535.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video relation detection with spatio-temporal graph. In *ACMMM*, pages 84–93.

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

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *ICMR*, pages 279–287. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

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

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*.

Shaoning Xiao, Long Chen, Jian Shao, Yueting Zhuang, and Jun Xiao. 2021b. Natural language video localization with learnable moment proposals. In *EMNLP*.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021c. Boundary proposal network for two-stage natural language video localization. In *AAAI*, pages 2986–2994.

Wentao Xie, Guanghui Ren, and Si Liu. 2020. Video relation detection with trajectory-aware multi-modal features. In *ACMMM*, pages 4590–4594. ACM.

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

Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *SIGIR*, pages 829–832.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134.

Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Video corpus moment retrieval with contrastive learning. In *SIGIR*, pages 685–695.

Appendix

A Attention Operation

Attention can be generalised to compute a weighted sum of the values dependent on the query and the corresponding keys. Since the query determines which values to focus on, we can say that the query attends to the values. Given a query $\tilde{Q}$ and a set of key-value pairs ($\tilde{K}$, $\tilde{V}$), dot-product attention adopted by Transformer (Vaswani et al., 2017) computes the alignment weights using dot-production of $\tilde{Q}$ and $\tilde{K}$ as shown in the right part of Figure 3,

$$Atten(\tilde{Q}, \tilde{K}, \tilde{V}) = softmax(\frac{(\tilde{Q}W_h^{\tilde{Q}})(\tilde{K}W_h^{\tilde{K}})^T}{\sqrt{d_h}})\tilde{V}W_h^{\tilde{V}}, \tag{10}$$

where $W_h^{\tilde{Q}}$, $W_h^{\tilde{K}}$ and $W_h^{\tilde{V}}$ are trainable projection matrices and $\sqrt{d_h}$ is a scaling factor that prevents softmax function from excessively large with keys of higher dimensions.

B Implementation Details

Frame-level feature details. We uniformly split each video into 16 segments and each segment has 16 consecutive frames. We utilize a ResNet-101 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) to extract per-frame appearance feature of 2048-D. As for the 2048-D motion feature, we utilize an I3D ResNeXt-101 (Hara et al., 2018) pre-trained on Kinetic (Kay et al., 2017) as mainstream framework.

Trajectory-level feature details. We sample video at a rate of 1fps. We adopt Faster-RCNN (Ren et al., 2015) trained on open-Images as the object detector, which uses Inception Resnet V2 as the image feature extractor, containing 600 classes. We use a dynamic programming algorithm improved from sequence NMS to associate bounding boxes that belong to the same object and generate trajectories. This tracking method consists of two steps: graph building and trajectory selection and we refer readers to (Xie et al., 2020) for more details.

Language representation details. We first extract tokens from sentences. Then we employ the GloVe (Pennington et al., 2014) pre-trained on Wikipedia to obtain 300-D embedding for each word token. The maximum length of question-answer pairs is set to 37. We truncated the sentences longer than the max length and padded the shorter ones with zeros. For the BERT-FT setting, we directly utilized finetuned BERT feature provided by NExT-QA (Xiao et al., 2021a). Each answer is appended to the question as a global sentence. A BERT build-in tokenizer is used to obtain the tokenized representation of the sentence. Then the tokens are organized by the format: [CLS] question [SEP] candidate answer [SEP].