Pairwise VLAD Interaction Network for Video Question Answering

Hui Wang[1,2,3], Dan Guo[1,2,3*], Xian-Sheng Hua[4], Meng Wang[1,2,3*]
[1] Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education
[2] Intelligent Interconnected Systems Laboratory of Anhui Province (HFUT), China
[3] School of Computer Science and Information Engineering School of Artificial Intelligence, Hefei University of Technology (HFUT)
[4] Damo Academy, Alibaba Group
wanghui.hfut@gmail.com,guodan@hfut.edu.cn,xiansheng.hxs@alibaba-inc.com,eric.mengwang@gmail.com

ABSTRACT

Video Question Answering (VideoQA) is a challenging problem, as it requires a joint understanding of video and natural language question. Existing methods perform correlation learning between video and question have achieved great success. However, previous methods merely model relations between individual video frames (or clips) and words, which are not enough to correctly answer the question. From human's perspective, answering a video question should first summarize both visual and language information, and then explore their correlations for answer reasoning. In this paper, we propose a new method called Pairwise VLAD Interaction Network (PVI-Net) to address this problem. Specifically, we develop a learnable clustering-based VLAD encoder to respectively summarize video and question modalities into a small number of compact VLAD descriptors. For correlation learning, a pairwise VLAD interaction mechanism is proposed to better exploit complementary information for each pair of modality descriptors, avoiding modeling uninformative individual relations (*e.g.*, frame-word and clip-word relations), and exploring both inter- and intra-modality relations simultaneously. Experimental results show that our approach achieves state-of-the-art performance on three VideoQA datasets: TGIF-QA, MSVD-QA, and MSRVTT-QA. Visualization results further validate the interpretability of our method.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Information systems** → **Question answering**.

KEYWORDS

Video question answering, VLAD, Pairwise interaction

ACM Reference Format:
Hui Wang[1,2,3], Dan Guo[1,2,3*], Xian-Sheng Hua[4], Meng Wang[1,2,3*] . 2021. Pairwise VLAD Interaction Network for Video Question Answering. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475620

*Corresponding authors.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475620

1 INTRODUCTION

Intelligent conversation referring to both vision and language has emerged as a prospective research topic in the computer vision community, such as visual question answering (VQA) [1, 37], visual dialog [4, 12, 13, 22] and video question answering (VideoQA) [18, 24, 31, 33]. Compared with image QA tasks where a question is just about a static image, VideoQA is more challenging due to two reasons: (1) a video tends to have substantial redundant information, and how to summarize core sequential clues in the video has been rarely studied; (2) answering a video-related question (*e.g.*, action counting and state transition) requires both appearance and motion information; and it is still difficult to establish the complex semantic connections between textual and various visual information.

For VideoQA, early works focus on sequential learning. A common practice [18] is to use RNN-based encoders to respectively encode video and word sequences, and then fuses the encoded features to reason the answer. However, RNNs have the weakness in utilizing long-distant information because of the gradient vanishing problem. To address this issue, Memory Networks [2, 7, 9] are introduced for VideoQA, which can cache different parts of sequence information in the memory slots and include long-term information.

Recent efforts towards VideoQA try to uncover latent correlations between video frames and question words with attention mechanisms. As illustrated in Fig. 1, previous attention approaches can be divided into three categories: (1) Self-attention [35] aggregates semantics inside each modality to model the intra-modality relations; (2) The co-attention mechanism [25] aggregates semantics from the other modality to model the inter-modality relations; (3) The intra-inter modality attention [26] models the relations within and across multiple modalities. Despite the success, these attention approaches were only proposed for modeling the relations between video frames and question words; thus, correlation learning requires large GPU memories because it needs to model relations between every pair. In addition, as videos contain substantial redundant information, simply modeling relations between individual video frames and words can be suboptimal for answer reasoning.

To model more complex multi-modal correlations, we propose a novel Pairwise VLAD Interaction Network (PVI-Net), to tackle

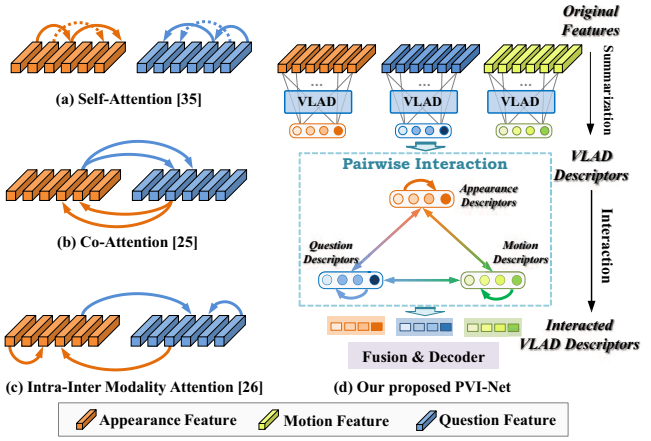


Figure 1: Different attention flows for VideoQA. In our solution (d), we first summarize each input modality into a small number of VLAD descriptors, and then simultaneously model intra- and inter-modality relations in a pairwise fashion.

the problem in two main steps. **First, summarizing core semantic clues in each modality stream (frame/clip/word feature sequences).** To this end, we extend the orderless clustering-based VLAD technique [19] for VideoQA. We propose a Long Short-term aware VLAD (LS-VLAD) encoder to achieve long-term temporal feature summarization with short-term temporal clues awareness. As shown in Fig. 1 (d), the LS-VLAD encodes each input modality into a small number of VLAD descriptors. Each descriptor can be formulated as the weighted aggregation over the entire feature sequence, which summarizes certain aspect of each modality from a global perspective and therefore contains richer context semantics compared with individual frame, clip, and word features. **Second, establishing multi-modal correlations on summarized VLAD descriptors.** As shown in Fig. 2, a pairwise interaction module is designed to model the relations between each pair of modality descriptors, and the outputs are divided into several interacted tensors according to the guided modality. Each interacted tensor captures both intra- and inter-modality relations, and is then fused with modality importance weights. The comprehensive interaction clues help the model achieve better answer reasoning.

Our main contributions can be summarized as follows: (1) We develop a learnable clustering-based VLAD encoder for modality summarization, which helps explore the multi-modal correlations from global perspectives and avoids capturing too much uninformative individual relations; (2) We propose a novel multi-modal interaction mechanism called pairwise VLAD interaction, which compute interactions between each pair of modality descriptors, capturing both intra- and inter-modality relations simultaneously; (3) Extensive experiments on TGIF-QA [18], MSVD-QA [41], and MSRVTT-QA [41] datasets demonstrate the superiority of the proposed PVI-Net compared against state-of-the-art methods. Ablation studies and qualitative visualizations also verify each component of PVI-Net.

2 RELATED WORK

Feature Aggregation for Videos. For video feature aggregation, several approaches have been investigated, such as simple aggregation – average or maximum pooling [36], recurrent neural networks (*e.g.*, RNN, LSTM, and GRU [17, 18]), and conventional clustering-based approaches (*e.g.*, BoVW [32], Fisher Vector [29], and VLAD [19]). Simple aggregations neglect the temporal structure of video. Recurrent aggregations have the weakness of gradient vanishing along temporal dimension. About clustering-based aggregations, recent works try to develop these techniques in a deep learning manner [11, 40, 44]. The technical point is to realize trainable aggregation neural networks via backpropagation rather than classical K-means clustering. In this paper, we investigate the VLAD aggregation with trainable networks for VideoQA.

Main Techniques for VideoQA. Recent approaches for VideoQA focused on sequential learning and correlation learning. We review the existing approaches as follows: (1) *RNN/Memory-based models* are typical methodologies for sequential learning, such as co-memory network [9], heterogeneous memory network [7], and augmented memory network [2]. (2) *Graph-based models* are recently popular methodologies for correlation learning among multi-modalities, such as location-aware graph [16] and heterogeneous graph [20]. (3) *Conditional Relation Network* (CRN) is another relational learning methodology [23], introducing a reusable CRN neural unit in a hierarchy stack to realize high-order relational learning and multi-step reasoning

Attention-based Approaches for VideoQA. Our work is mostly related to the attention-based approaches. Many correlation learning approaches using attention mechanisms to aggregate interaction information. Gao et al. [10] applied co-attention to model the relation between each word and video frame pairs. Li et al. [25] extended single-path co-attention to multi-path pyramid co-attention with diversity learning. Zha et al. [43] proposed spatio-temporal co-attention to focus on the question-relevant regions in frames and the relevant frames within videos. Transformer [35] proposed to use the self-attention mechanism to model the relationship inside each modality, which has been widely applied in various research topics. For VideoQA, Jin et al. [21] developed a multi-interaction network to capture both element-wise and segment-wise modality interactions. Li et al. [26] combined self-attention with co-attention to model relationships within and across video frames and question words.

Compared to previous methods, our proposed Pairwise VLAD Interaction Network (PVI-Net) has two distinctive characteristics: (1) PVI-Net does not model relations from the large number of individual visual-word pairs but from the small number of summarized VLAD descriptors, which can capture high-level cross-modality interactions with smaller modal capacity; (2) modality interactions are performed in a pairwise fashion, both intra- and inter-modality relations are learned simultaneously.

3 PROPOSED METHOD

The overall pipeline of the proposed PVI-Net is shown in Fig. 2, which consists of three components: (1) Modality Summarization via VLAD Encoder (Sec. 3.1), which aims to summarize each input modality into a few compact VLAD descriptors; (2) Correlation

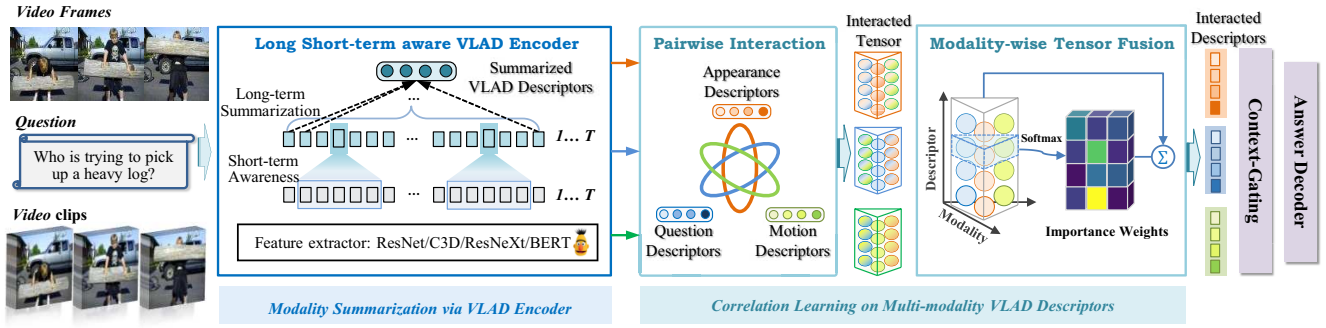


Figure 2: An overview of the proposed Pairwise VLAD Interaction Network (PVI-Net). Firstly, we use a generic VLAD encoder to summarize each input modality into a few VLAD descriptors. Subsequently, we explore correlations between each pair of modality descriptors and construct interacted tensor for each guided modality. A modality-wise tensor fusion is then executed to aggregate each tensor into a interacted descriptor. Finally, the interacted descriptors are gated fused for answer prediction.

Learning on Multi-modality VLAD Descriptors (Sec. 3.2), which is a two-stage procedure. In the first stage, in parallel to model interactions between each pair of modality descriptors and construct three interacted tensors. In the second stage, each tensor is weighted fused into interacted descriptors; and (3) Context-Gating & Answer Decoding (Sec. 3.3), which combines interacted descriptors and further recalibrates it with a learned gate for capturing context clues, which are most relevant to the answer.

3.1 Modality Summarization via VLAD Encoder

In this work, the summarization idea expands the temporal representation learning of both video and question. We adopt a clustering-based aggregation scheme - VLAD [19], to summarize core semantic clues across the entire feature sequence of each modality. The clustering characteristics of VLAD helps to eliminate redundant and uninformative clues in the feature sequence and ensure diversity with multiple to-be-learned descriptors (clusters).

Backbone. Conventional VLAD [19] is designed for aggregating isolate features, its cluster centers $\{c_k\}|_{k=1}^{K}$ are obtained in an unsupervised approach (*e.g.*, K-means), and maps each feature to the nearest $c_k \in \mathbb{R}^{d_x \times 1}$ as Eq. (1). However, its clustering manner leads to the non-differentiable problem. To address this issue, we redesign a RVLAD [6] for temporal feature summarization, which removes residual and uses a soft-assignment tactic. Given a feature sequence $X = [x_1, ..., x_T] \in \mathbb{R}^{d_x \times T}$, RVLAD aims to model the correlation between X and K clusters rather than learning the cluster centers themselves. RVLAD maps each feature of X to K latent clusters with a *Relevance* distribution matrix $\bar{R} \in \mathbb{R}^{T \times K}$, and obtains summarized descriptors $\bar{X} = [\bar{x}_1, \cdots, \bar{x}_K] \in \mathbb{R}^{d_x \times K}$. In *Relevance* $\bar{R}$, each element $\bar{r}_k(x_t)$ indicates a normalized relevance weight of original x_t to the k-th cluster. Each descriptor $\bar{x}_k$ is computed with weighted sum across the entire feature sequence as Eq. (2):

$$\mathrm{VLAD}_k(X) = \sum_{t=1}^{T} r_k(x_t) \cdot (x_t - c_k), \tag{1}$$

$$\begin{cases} \bar{r}_k(x_t) = \frac{e^{w_k^\top x_t + b_k}}{\sum_{j=1}^{T} e^{w_k^\top x_j + b_k}}, \\ \mathrm{RVLAD}_k(X) = \sum_{t=1}^{T} \bar{r}_k(x_t) \cdot x_t; \end{cases} \tag{2}$$

where $\{w_k\}$ and $\{b_k\}$ are sets of trainable parameters for each cluster k.

Long Short-term aware VLAD Encoder. In our solution, for temporal feature summarization, we approval the viewpoint of understanding the local temporal clues first and then capturing the long-range dependencies.

1) *Short-term Awareness.* To extract temporal clues at current timestamp, we observe the current frame, clip, or word x_t with a time sliding window $\mathbb{W} = 2\Delta T + 1$, namely backward and forward ΔT ranges, $[x_{t-\Delta T}, \cdots, x_t, \cdots, x_{t+\Delta T}]$. Here, to achieve short-term temporal clue awareness, we embed a separable convolution [3] operation into VLAD encoder, which implements $\mathbb{W} \times 1$ kernel size *depthwise Covn1D* to learn temporal-wise correlation with sliding window $\mathbb{W}$ at each fixed channel, and then use a 1×1 *pointwise Covn1D* to combine the above depth correlation. With the two-level decoupled convolution process, the local context feature x_t^{loc} at time t is calculated as follows:

$$\begin{aligned} x_t^{loc} &= \mathrm{SeparableConv}(x_t, \Delta T) \\ &= \begin{cases} x_t' = Conv1D(x_{t-\Delta T}, \cdots, x_t, \cdots, x_{t+\Delta T})|_{depthwise}; \\ x_t^{loc} = Conv1D(x_t')|_{pointwise}, \end{cases} \end{aligned} \tag{3}$$

where $x_t, x_t^{loc} \in \mathbb{R}^{d_x \times 1}$. We then add x_t^{loc} on the original feature x_t, so as to involve the local temporal clues:

$$x_t^S = \mathrm{LN}(x_t \oplus x_t^{loc}), \tag{4}$$

where LN denotes Layer Normalization, and $x_t^S \in \mathbb{R}^{d_x \times 1}$ is the output short-term aware feature.

2) *Long-term Summarization.* After acquiring the short-term aware sequence $X^S = [x_1^S, \cdots, x_T^S]$, we summarize X^S into K number of descriptors by using RVLAD, which implements K times global mappings of the entire sequence, promising diverse feature summarization from multiple global perspectives. As shown in Eq. (2), RVLAD maps the feature sequence into K clusters with a *Relevance* matrix $\bar{R} \in \mathbb{R}^{T \times K}$, which is calculated by *Conv1D* and

column-wise $softmax$. Thus, the whole LS-VLAD encoder is formulated as follows:

$$B_X = \text{LS-VLAD}(X, \Delta T) \Leftrightarrow \begin{cases} X^{loc} = \text{SeparableConv}(X, \Delta T); \\ X^S = \text{LN}(X \oplus X^{loc}); \\ \bar{X} = \text{RVLAD}(X^S); \\ \quad = X^S \cdot \bar{R}_{column-wise}; \\ \quad = X^S \cdot \text{softmax}(Conv1D(X^S)); \\ \quad = [\bar{x}_1, \cdots, \bar{x}_K], \\ B_X = \Phi(\bar{X}) = [W_1\bar{x}_1, \cdots, W_K\bar{x}_K]; \\ \quad = [b_1, \cdots, b_K], \end{cases} \tag{5}$$

where $Conv1D$ is applied to map the sequence length T to K clusters; a transform layer $\Phi(\cdot)$ is applied to project every $\bar{x}_k \in \mathbb{R}^{d_x \times 1}$ into a lower d_s-dim vector with parameters W_k. In our solution, the dimension of B_X is much smaller than X including both $K \ll T$ and $d_s \ll d_x$. The summarized VLAD descriptors is more compact than original feature sequence.

Multi-modality Summarization. We evenly sample T frames and segment T clips from each video, and extract appearance features $F_A = [f_1^A, \cdots, f_T^A] \in \mathbb{R}^{d_A \times T}$ and motion features $F_M = [f_1^M, \cdots, f_T^M] \in \mathbb{R}^{d_M \times T}$. For each question, we extract word-level features $F_Q = [f_1^Q, \cdots, f_L^Q] \in \mathbb{R}^{d_Q \times L}$. Here, we denote appearance modality summarization as $B_A = \text{LS-VLAD}(F_A) = [b_1^A, \cdots, b_K^A] \in \mathbb{R}^{d_s \times K}$, where b_k^A is the k-th summarized descriptor; the same are B_M and B_Q. Thus, we obtain the summarized VLAD descriptors of appearance, motion, and question, *i.e.*, $\{B_A, B_M, B_Q\}$. It is worth noting that the parameters of the VLAD encoder for each feature stream F_A, F_M, and F_Q are learned independently.

3.2 Correlation Learning on Multi-modality VLAD Descriptors

The summarized VLAD descriptors encode high-level information from one of the modalities. To reason the correct answer of the input video and question, it is important to understand the complex correlations among the multi-modalities. We therefore propose a pairwise interaction mechanism to establish the associations for all pairs of modalities.

Pairwise Interaction. To explore the multi-modal correlations, we extend the conventional multi-head self-attention [35] into a guided-attention fashion. Given a guided feature set X and a feature set Y, they are linearly projected into three new sequences: ***query*** X^Q, ***key*** Y^K, and ***value*** Y^V. The guided multi-head attention is performed to measures the relation of each element in X with Y as follows:

$$\begin{cases} \text{h}_i = \text{softmax}(\frac{X^Q(Y^K)^\top}{\sqrt{d_h}}) \cdot Y^V; \\ \text{MHA}(X, Y) = W_o[\text{h}_1; \text{h}_2; ...; \text{h}_h], \end{cases} \tag{6}$$

where h is the number of attention heads and h_i is the output of the i-th attention head.

1) *Interaction Unit.* Based on the guided multi-head attention, we propose an Interaction Unit to explicitly model the relation between a pair of VLAD descriptors, denoted by $\text{INT}(B_X, B_Y)$, where

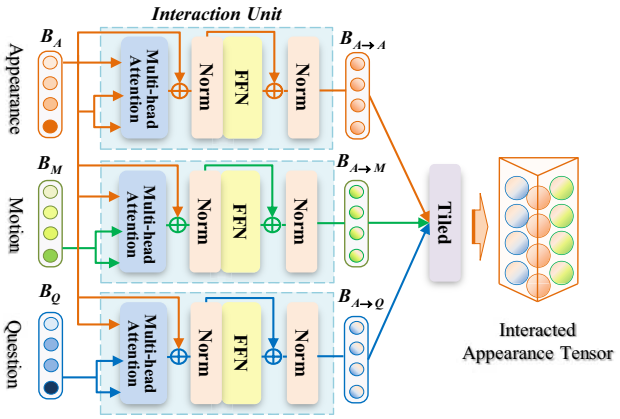


Figure 3: Illustration of the appearance guided pairwise interaction. The interacting results are tiled as a high-dimensional tensor to exploit both intra- and inter-modality semantics.

$B_X \in \{B_A, B_M, B_Q\}$ and $B_Y \in \{B_A, B_M, B_Q\}$. The interaction is formulated as:

$$B_{X \to Y} = \text{INT}(B_X, B_Y) \Leftrightarrow \begin{cases} \bar{B}_{X \to Y} = \text{LN}(\text{MHA}(B_X, B_Y) \oplus B_X); \\ B_{X \to Y} = \text{LN}(\text{FFN}(\bar{B}_{X \to Y}) \oplus \bar{B}_{X \to Y}), \end{cases} \tag{7}$$

where FFN(.) is a feed-forward network. Two residual connections are introduced to enhance the semantics of guided descriptors B_X and interacted descriptors $\bar{B}_{X \to Y}$, respectively. $B_{X \to Y} \in \mathbb{R}^{d_s \times K}$ is the output of the unit, which models the relation of $B_X \to B_Y$.

2) *Modality-interacted Tensor.* We adopt the above mentioned Interaction Unit to realize a modality guided pairwise interaction. As shown in Fig. 3, taking the appearance descriptors as the guided feature, we implement intra-modality interaction as $\text{INT}(B_A, B_A)$ and inter-modality interactions as $\text{INT}(B_A, B_M)$ and $\text{INT}(B_A, B_Q)$; then, the interacting results are tiled as a high-dimensional tensor:

$$\mathbb{A} = \text{Tile}\big(\text{INT}_1(B_A, B_A), \text{INT}_2(B_A, B_M), \text{INT}_3(B_A, B_Q)\big), \tag{8}$$

where $\mathbb{A} \in \mathbb{R}^{d_s \times 3 \times K}$ is the interacted appearance tensor, which models both intra- and inter-modality relations of $B_A \to \{B_A, B_M, B_Q\}$. With the same operation on motion and question descriptors, we obtain the interacted motion tensor $\mathbb{M} \in \mathbb{R}^{d_s \times 3 \times K}$ and question tensor $\mathbb{Q} \in \mathbb{R}^{d_s \times 3 \times K}$.

Modality-wise Tensor Fusion. Up to now, we have already modeled relations between each pair of summarized modality descriptors, and construct three interacted tensors $\{\mathbb{A}, \mathbb{M}, \mathbb{Q}\}$ according to the guided modality. The information in each tensor needs to be further aggregated before feeding it to answer decoder. Simple concatenation or pooling can achieve this purpose. Here, we perform a weighted fusion to attend important modality relations in each tensor.

Taking the appearance tensor $\mathbb{A}$ as an example, we use a position-wise fully-connected layer to predict the importance weight of each modality relation in $\mathbb{A}_k$, where $\mathbb{A}_k \in \mathbb{R}^{d_s \times 3}$ is the k-th descriptor layer of tensor $\mathbb{A}$. With the importance weight α_k^A, we aggregate all the interacting semantics of $\mathbb{A}_k$ into an interacted descriptor

$\widehat{\mathbb{A}}_k \in \mathbb{R}^{d_s \times 1}$. At last, we concatenate all descriptors $\{\widehat{\mathbb{A}}_k\}|_{k=1}^{K}$ and obtain the fused appearance descriptor v_A:

$$\begin{cases} \alpha_k^A = \text{softmax}_m(W_k^A \mathbb{A}_k + b_k^A); \\ \widehat{\mathbb{A}}_k = \sum_{m=1}^{3} \alpha_{k,m}^A \mathbb{A}_{k,m}; \\ v_A = [\widehat{\mathbb{A}}_1; \cdots; \widehat{\mathbb{A}}_K], \end{cases} \tag{9}$$

where $W_k^A \in \mathbb{R}^{1 \times d_s}$, $b_k^A \in \mathbb{R}^3$ are learnable parameters. $v_A \in \mathbb{R}^{d \times 1}$ is the output appearance descriptor – a relation aware descriptor, where $d = d_s \times K$. Similarly, we conduct the fusion operation on tensors $\mathbb{M}$ and $\mathbb{Q}$, and obtain the fused motion descriptor $v_M \in \mathbb{R}^{d \times 1}$ and question descriptor $q \in \mathbb{R}^{d \times 1}$, respectively. Thereafter, the new fused modality descriptors $\{v_A, v_M, q\}$ are used for answer reasoning.

3.3 Context-Gating & Answer Decoding

Context-Gating is designed to imitate humans adaptively selecting relevant context clues to reason the answer. Specifically, we concatenate the multi-modal descriptors $[v_A; v_M; q] \in \mathbb{R}^{3d \times 1}$, and feed it forward to a *Sigmoid* layer. We obtain a gating vector g, and apply g to retrieve useful context clues and suppress uninformative clues in $[v_A; v_M; q]$:

$$\begin{cases} g = \sigma(W_g[v_A; v_M; q] + b_g); \\ e = g \circ [v_A; v_M; q], \end{cases} \tag{10}$$

where $\circ$ denotes an element-wise multiplication, $W_g \in \mathbb{R}^{3d \times 3d}$ and $b_g \in \mathbb{R}^{3d}$ are learnable parameters. $e \in \mathbb{R}^{3d \times 1}$ is the gated multi-modal descriptor, which is further used for answer decoding. We follow the common settings for answer decoders [18] which project e using an MLP followed by a softmax to rank the possible answer choices. We train the model with cross-entropy loss for all tasks except *Count*, where Mean Square Error is used.

4 EXPERIMENTS

4.1 Experiment Setup

Datasets. We evaluate our method on three benchmark datasets: (1) **TGIF-QA** [18] dataset consists of 165K QA pairs from 72K animated GIFs. It is divided into four task types: *Action* - a multiple-choice task to recognize the action repeated for a specified times; *Trans.* - a multiple-choice task regarding temporal order of events; *Count* - an open-ended number task to retrieve number of occurrences of an action; and *FrameQA* - an open-ended word task in which the answers can be inferred from one of the frames in the video. (2) **MSVD-QA** [41] collected 50K QAs with 1,970 short movie videos. It is an open-ended word task, including five types of questions: *What*, *Who*, *How*, *When*, and *Where*. (3) **MSRVTT-QA** [41] chose 10K videos from MSRVTT [39] and collected 243K QA pairs. Similar to MSVD-QA, questions are of five types. Compared to the other two datasets, videos in MSRVTT-QA are longer and contain more complex scenes. For the evaluation metrics, we adopt Mean Square Error (**MSE**) for *Count* task and use **Accuracy** for other tasks.

Implementation Details. We systematic sample $T = 36$ frames and clips from each video in TGIF-QA dataset and $T = 20$ in MSVD-QA and MSRVTT-QA. Following [18], we adopt pre-trained ResNet [15] and C3D [34] to extract appearance and motion features on TGIF-QA; following [23], we utilize ResNet and ResNeXt [14] on MSVD-QA and MSRVTT-QA. Then, we tokenize questions. We pad or truncate each question to 35 words for all datasets. The pre-trained BERT [5] is applied to extract word-level features. After feature extraction, position embedding [35] is added to each feature sequence to incorporate the sequential information. The dimension of all the VLAD descriptors is set as $d_s = 128$, and the number of attention head h is set as 2. About training details, we set batch size as 64 for TGIF-QA and 32 for the others, and use Adamax optimizer with initial learning rate of 0.001. The learning rate is multiplied by 0.5 after every 5 epochs.

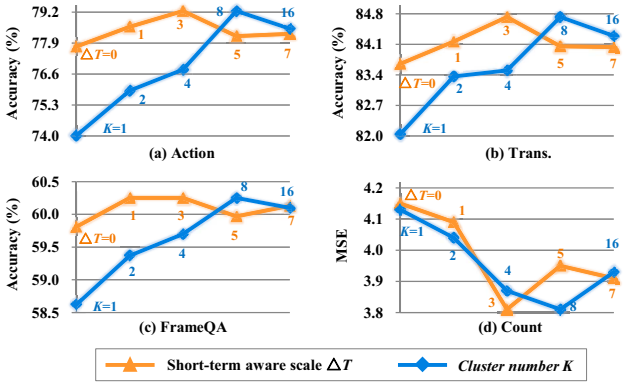


Figure 4: Ablation studies with short-term aware scale ΔT and cluster number K on TGIF-QA dataset. ΔT changes with fixed $K = 8$, and K changes with fixed $\Delta T = 3$.

Table 1: Ablation studies of different *Interaction* and *Tensor Fusion* strategies on TGIF-QA dataset.

Model	Task Type			
	Action↑	Trans.↑	FrameQA↑	Count↓
Intra-modality Interaction	78.36	84.31	59.94	3.88
Inter-modality Interaction	78.23	84.17	59.88	3.92
(Intra&Inter)-modality(Ours)	**79.24**	**84.72**	**60.25**	**3.79**
Concat Fusion	77.66	84.35	59.35	3.99
Sum Fusion	78.36	84.18	59.27	3.89
Weighted Fusion(Ours)	**79.24**	**84.72**	**60.25**	**3.79**

4.2 Ablation Study

Empirical parameters of LS-VLAD encoder. LS-VLAD contains two hyperparameters – **short-term aware scale** ΔT and **cluster number** K. As depicted in Fig. 4, the proposed PVI-Net achieves the best in $\Delta T = 3$. Note that the performances with $\Delta T > 0$ are always better than $\Delta T = 0$ (*i.e.*, removing separable convolution layer in LS-VLAD). It indicates considering the local temporal clues (in Sec. 3.1) improves the temporal summarization ability of VLAD. K reflects the number of summarized descriptors of each modality. Insufficient descriptors ($K < 8$) will be unable to capture different aspects of the input which deteriorates the overall performance. Too many descriptors ($K > 8$) will capture redundant clues. As

Table 2: Ablation studies of different components on TGIF-QA dataset.

Model	Task Type			
	Action↑	Trans.↑	FrameQA↑	Count↓
PVI w/o VLAD	74.71	82.09	55.74	4.12
PVI w/o PairAtt	76.91	82.89	58.87	4.15
PVI w/o Weighted Fusion	78.36	84.18	59.27	3.89
PVI w/o Context-Gating	77.89	84.10	59.59	3.92
PVI w/ GloVe	77.85	82.77	58.92	3.87
PVI-Net (Ours)	**79.24**	**84.72**	**60.25**	**3.79**

Table 3: Comparison with the state-of-the-art methods on TGIF-QA dataset. Visual features are: R(ResNet), C(C3D), F(FlowCNN), MR(Mask RCNN), and RX(ResNext).

Model	Task Type			
	Action	Trans.	FrameQA	Count
RNN/Memory-based models				
VQA-MCB (R) [8]	58.9	24.3	25.7	5.17
VIS+LSTM (R) [30]	46.8	56.9	34.6	5.09
CT-SAN (R) [42]	56.1	64.0	39.6	5.13
Co-memory (R+F) [9]	68.2	74.3	51.5	4.10
HME (R+C) [7]	73.9	77.8	53.8	4.02
FAM (R+C) [2]	75.4	79.2	56.9	**3.79**
Graph-based models				
LAG (R+MR) [16]	74.3	81.1	56.3	3.95
HGA (R+C) [20]	75.4	81.0	55.1	4.09
Conditional relation model				
HCRN (R+RX) [23]	75.0	81.4	55.9	3.82
Attention/Transformer-based models				
ST(R+C) [18]	60.8	67.1	49.3	4.40
STA (R) [10]	72.3	79.0	56.6	4.25
PSAC (R) [26]	70.4	76.9	55.7	4.27
MIN (R+MR) [21]	72.7	80.9	57.1	4.17
LAD-Net (R) [25]	72.0	80.7	58.2	4.24
ACR (R+F) [45]	75.8	81.6	57.7	4.08
PVI-Net (R+C)	**79.2**	**84.7**	**60.3**	**3.79**

shown in Fig. 4, $K = 8$ is the best. Thus, we set $\Delta T = 3$ and $K = 8$ in the following experiments.

Different Interaction and Tensor Fusion Strategies. We inspect the intra- and inter-modality interactions separately. As shown in Table 1, only modeling the intra- or inter-modality relations is not enough to reason the correct answer. PVI-Net performs modality interaction in a pairwise fashion to model both intra- and inter-modality relations, improving the performance with a clear margin. We also compare the effects of different tensor fusion strategies, and our **weighted fusion** (in Sec. 3.2) performs better than simple **concatenation** and **sum pooling**. It indicates the modality relations in each tensor produce unique clues of different importance.

Main Components in PVI-Net. Among the variants of PVI-Net in Table 2, the worst performance occurs on **PVI w/o VLAD** that directly feed original features into pairwise interaction module without modality summarization, deteriorating the performance by -4.53% on *Action*, -2.63% on *Trans.*, -4.51% on *FrameQA*, and -8.71% on *Count*. The capability of summarizing core sequential clues is rapidly dropped without VLAD encoding. The second worse performance occurs on **PVI w/o PairAtt** (*i.e.*, -2.33% on *Action*, -1.83% on *Trans.*, -1.38% on *FrameQA*, and -9.49% on *Count*). It indicates modeling the multi-modal correlations is important for VideoQA. The comparison result between **PVI w/o Weighted Fusion** and PVI-Net indicates that performing modality-wise tensor fusion is contributive in our solution. **Context-Gating** offers further performance gain in all tasks, by adaptively selecting relevant clues through gating mechanism. In addition, we provide the results of PVI-Net using GloVe [28] text representation (**PVI w/ GloVe**).

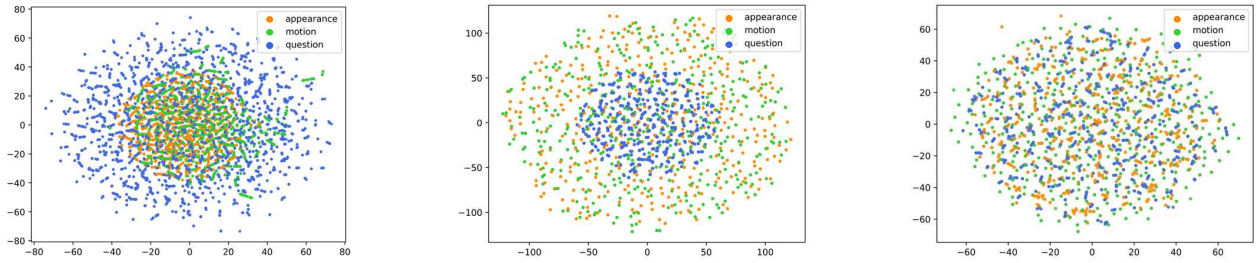
Table 4: Comparison with the state-of-the-art methods on MSVD-QA dataset. Visual features are: V (VGG), R(ResNet), C(C3D), MR(Mask RCNN), and RX(ResNext).

Model	Question Type					
	What	Who	How	When	Where	All
E-MN (V) [38]	12.9	46.5	80.3	70.7	50.0	26.7
Co-memory (V+C) [9]	19.6	48.7	81.6	74.1	31.7	31.7
HME (V+C) [7]	22.4	50.1	73.0	70.7	42.9	33.7
FAM (V+C) [2]	23.1	51.6	82.2	71.4	51.9	34.5
LAG (R+MR) [16]	-	-	-	-	-	34.3
HGA (V+C) [20]	23.5	50.4	83.0	72.4	46.4	34.7
E-SA (V) [38]	15.0	45.1	**83.8**	65.5	32.2	27.6
DLAN (V) [47]	21.2	46.0	83.2	72.4	50.0	31.8
GRAAM (V+C) [38]	20.6	47.5	83.5	72.4	53.6	32.0
AA-Net (R+C) [46]	21.3	48.3	82.4	70.7	53.6	32.6
STCA (V+C) [43]	24.3	49.6	83.0	74.1	53.6	35.0
MIN (V+MR) [21]	24.2	49.5	**83.8**	74.1	53.6	35.0
HCRN (R+RX) [23]	-	-	-	-	-	36.1
PVI-Net (R+RX)	**31.7**	**55.1**	83.5	**74.1**	**57.1**	**41.5**

4.3 Comparison with State-of-the-Arts

We compare our PVI-Net with state-of-the-art methods as follows: RNN/Memory-based models [2, 7–9, 30, 42], graph-based models [16, 20], conditional relation model [23], and attention-based models [10, 18, 21, 25, 26, 43, 45–47]

Results on TGIF-QA. As shown in Table 3, PVI-Net performs prominent superior to all the other methods. **(1) FAM** [2] achieves the best performance in the RNN/Memory-based models. PVI-Net performs significantly better than it, *i.e.*, 79.2% vs. 75.4% on *Action*, 84.7% vs. 79.2% on *Trans.*, and 60.3% vs. 56.9% on *FrameQA*. **(2)** In Graph-based models, **HGA** [20] builds a heterogeneous graph to model the relations among all video shots and words. Compared with HGA, our PVI-Net does not explore interactions from the large number of individual visual-word pairs but from the multi-modal summarization descriptors, improving the results by 3.8% on *Action*, 3.7% on *Trans.*, 5.2% on *FrameQA*, and 7.3% on *Count*. **(3)** For conditional relation model, **HCRN** [23] merely models frame-level object relations conditioned on the motion and linguistic cues. In contrast, PVI-Net comprehensively models relations between each pair of modalities and outperforms HCRN on all tasks. **(4) ACR** [45] reports the recent best results in attention-based models, which extracts action-aware frame features with action encoder [27] and



(a) Original features (b) VLAD descriptors (c) Interacted VLAD descriptors

Figure 5: t-SNE plots for visualizing the embedding distribution of various features. Original features are directly derived from feature extractors, VLAD descriptors are outputs of LS-VLAD encoder as in Eq. (5), and interacted VLAD descriptors are obtained after pairwise VLAD interaction as in Eq. (9). Orange, green, and blue points denote appearance, motion, and question features, respectively.

Table 5: Comparison with the state-of-the-art methods on MSRVTT-QA dataset. Visual features are: V (VGG), R(ResNet), C(C3D), MR(Mask RCNN), and RX(ResNext).

Model	Question Type					All
	What	Who	How	When	Where	
E-MN (V) [38]	23.4	41.8	83.7	70.8	27.6	30.4
Co-memory (V+C) [9]	23.9	42.5	74.1	69.0	**42.9**	32.0
HME (V+C) [7]	26.5	43.6	82.4	76.0	28.6	33.0
FAM (V+C) [2]	26.9	43.9	82.8	70.6	31.1	33.2
HGA (V+C) [20]	29.2	45.7	83.5	75.2	34.0	35.5
E-SA (V) [38]	22.0	41.6	79.6	73.1	33.2	29.3
DLAN (V) [47]	25.4	42.8	81.0	72.1	31.2	32.0
GRAAM (V+C) [38]	26.2	43.0	80.2	72.5	30.0	32.5
STCA (V+C) [43]	27.4	45.4	83.7	74.0	33.2	34.2
MIN (V+MR) [21]	29.5	45.0	83.2	74.7	42.4	35.4
HCRN (R+RX) [23]	-	-	-	-	-	35.6
PVI-Net (R+RX)	**32.8**	**48.9**	**84.8**	**79.3**	38.8	**39.0**

models frame-to-frame interplays with relation transformer. PVI-Net utilizes VLAD to summarize the core semantics of each input and models both intra- and inter-modality relations, achieving new state-of-the-art performance.

Results on MSVD-QA and MSRVTT-QA. Tables 4 and 5 show that PVI-Net still significantly outperforms existing methods, achieving 41.5% and 39.0% on *All* accuracy which improves 5.4 and 3.4 points on MSVD-QA and MSRVTT-QA, respectively. The results demonstrate the robustness of our method on different VideoQA tasks.

4.4 Comparison of Interaction Complexity

Table 6 shows a comparison of interaction complexity and GPU memory cost. PSAC [26] and LAD-Net [25] released codes, and we reproduced them in the same experiment environment as ours. LAD-Net is a typical co-attention model in which each word calculates an attention matrix from each video frame and vice versa; thus, its interaction complexity is $O(2 \times T \times L_2)$. PSAC combines self-attention with co-attention, the complexity of self-attention is $O(T \times T + L_1 \times L_1)$, and the total interaction complexity of PSAC is $O((T+L_1)\times(T+L_1))$. The quadratic number of interactions requires large GPU memories. For our proposed PVI-Net, it first summarizes

Table 6: Comparison of interaction complexity and required GPU memory on *Action* task of TGIF-QA. $T = 36$, $L_1 = 20$, and $L_2 = 25$ are respective sequence length of video and question features. $K = 8$ is the number of summarized descriptors

Model	Complexity	GPU(Mb)	Acc on *Action*↑
PSAC [26]	$O((T + L_1) \times (T + L_1))$	7817	70.4
LAD-Net [25]	$O(2 \times T \times L_2)$	5120	72.0
PVI-Net (Ours)	$O(9 \times K \times K)$	4441	79.2

each input modality into K descriptors, and then performs modality interaction in a pairwise fashion, its total interaction complexity is $O(9 \times K \times K)$. Notably, LAD-Net and PSAC merely consider the interaction between appearance and question features, while PVI-Net further takes motion features into account. Compared with LAD-Net and PSAC, PVI-Net realizes more complex multi-modality interactions with much fewer interaction complexity and GPU memories, and achieves better performance. The efficiency of PVI-Net is due to the summarized VLAD descriptors, *i.e.*, $K \ll T$, L_1, and L_2.

4.5 Qualitative Results

Fig. 5 shows the feature distribution of each module in PVI-Net. As shown in Fig. 5 (a), original visual features, including appearance and motion, are close to each other. In contrast, the textual features of question appear at the outer-ring of visual features. After VLAD encoding, as shown in Fig. 5 (b), question descriptors (blue points) appear in the center of the feature space. The distribution of visual descriptors greatly changes. Visual descriptors are scattered in the feature space to capture different aspects of the video from global perspectives. Turning to the interacted VLAD descriptors in Fig. 5 (c), multi-modal correlations have been established after pairwise interaction; thus, different modality descriptors get close to each other in the feature space.

Fig. 6 (a) displays an example from *Action* task, in which our VLAD successfully summarize different core clues of each modality stream. The summarized question descriptors separately focus on words {'head', 'man', '2 times', 'put hand on head'}, and the visual descriptors summarize the video clues with different temporal spans.

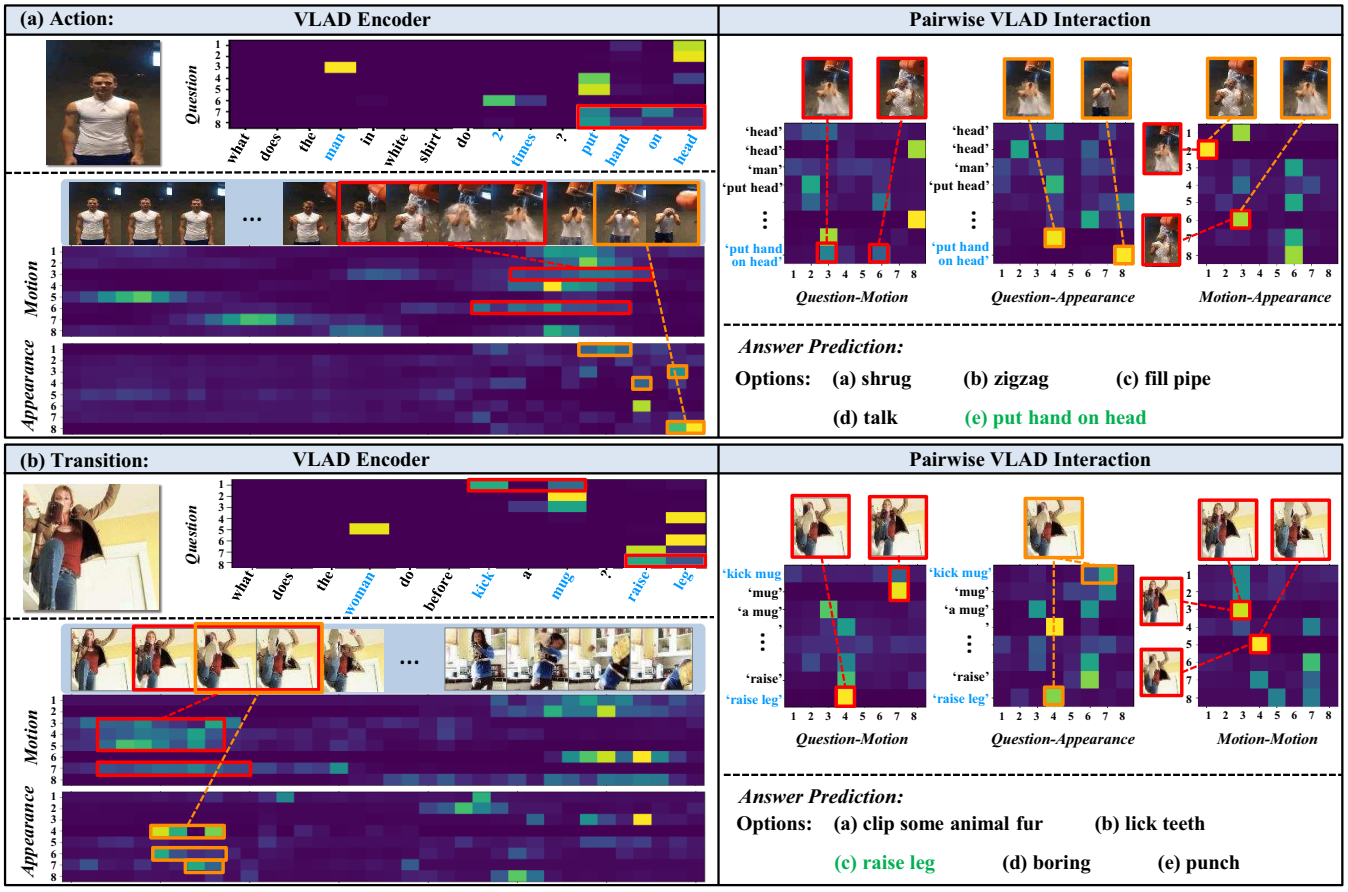


Figure 6: Visualization of two examples in TGIF-QA dataset. In the VLAD Encoder stage, the visualization of relevance matric $\bar{R}$ of each modality shows the mapping of feature summarization, where Y-axis denotes the i-th (1$\sim K$) descriptor summarization, X-axis corresponds to the original feature sequence $\{x_t\}|_{t=1}^{T}$. In the Pairwise VLAD Interaction stage, some selected attention maps from the interaction unit reflecting the correlations between different modality pairs.

In the interaction stage, we select three attention maps to verify the effectiveness of multi-modal correlation learning. For question-to-visual interactions, the 8-th question descriptor covering key words 'put hand on head', and its most relevant visual clues are the {3, 6}-th motion descriptors and the 8-th appearance descriptor, which successfully capture the action clues in two temporal spans (red and orange boxes). The motion-to-appearance interaction further highlights relevant clues in the visual descriptors. Fig. 6 (b) shows another example from *Transition* task, the question descriptors focus on two action-relevant phrases 'kick mug' (the 1-st descriptor) and 'raise leg' (the 8-th descriptor). For question-to-visual interactions, the 1-st question descriptor ('kick mug') correctly focuses on the 7-th motion descriptor and the {6, 7}-th appearance descriptors; turning to the 8-th question descriptor ('raise leg'), the responsive 4-th motion descriptor and 4-th appearance descriptor are completely correct too. The motion-to-motion interaction successfully models the intra-modality relations. To summarise, our VLAD encoder guarantees the diversity and discrimination of the summarized VLAD descriptors; and the Pairwise Interaction establishes complex associations among the multi-modal descriptors. The effectiveness of these two modules helps the model correctly reason the answer.

5 CONCLUSION

In this paper, we propose a novel Pairwise VLAD Interaction Network (PVI-Net) for VideoQA. We develop a clustering-based VLAD encoder to summarize each input modality into a few descriptors and realize modality interactions in a pairwise fashion, exploring both intra- and inter-modality relations for answer reasoning. We evaluate our method on three benchmark datasets and conduct extensive ablation studies to verify the effectiveness of PVI-Net. Experiments show the superiority of our method.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (NSFC) under grants 61725203, 62020106007, 61876058, and Fundamental Research Funds for the Central Universities under grant JZ2020HGTB0020.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. 2425–2433.

[2] Jiayin Cai, Chun Yuan, Cheng Shi, Lei Li, Yangyang Cheng, and Ying Shan. 2020. Feature Augmented Memory with Global Attention Network for VideoQA. In *IJCAI*. 998–1004.

[3] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*. 1800–1807.

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *CVPR*. 1080–1089.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[6] Matthijs Douze, Jérôme Revaud, Cordelia Schmid, and Hervé Jégou. 2013. Stable Hyper-pooling and Query Expansion for Event Detection. In *ICCV*. 1825–1832.

[7] 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*. 1999–2007.

[8] 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*. 457–468.

[9] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-Appearance Co-Memory Networks for Video Question Answering. In *CVPR*. 6576–6585.

[10] 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*. 6391–6398.

[11] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan C. Russell. 2017. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification. In *CVPR*. 3165–3174.

[12] Dan Guo, Hui Wang, and Meng Wang. 2019. Dual Visual Attention Network for Visual Dialog. In *IJCAI*. 4989–4995.

[13] Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative Context-Aware Graph Inference for Visual Dialog. In *CVPR*. 10052–10061.

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *CVPR*. 6546–6555.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[16] 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*. 11021–11028.

[17] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 2016. A Hierarchical Deep Temporal Model for Group Activity Recognition. In *CVPR*. 1971–1980.

[18] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*. 1359–1367.

[19] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *CVPR*. 3304–3311.

[20] Pin Jiang and Yahong Han. 2020. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In *AAAI*. 11109–11116.

[21] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. 2019. Multi-interaction Network with Object Relation for Video Question Answering. In *ACM MM*. 1193–1201.

[22] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog. In *NAACL*. 582–595.

[23] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical Conditional Relation Networks for Video Question Answering. In *CVPR*. 9968–9978.

[24] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*. 1369–1379.

[25] Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. 2019. Learnable Aggregating Net with Diversity Learning for Video Question Answering. In *ACM MM*. 1166–1174.

[26] 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*. 8658–8665.

[27] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *ECCV*. 3–21.

[28] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. 1532–1543.

[29] Florent Perronnin and Christopher R. Dance. 2007. Fisher Kernels on Visual Vocabularies for Image Categorization. In *CVPR*. 1–1.

[30] Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. Exploring Models and Data for Image Question Answering. In *NeurIPS*. 2953–2961.

[31] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision* 123, 1 (2017), 94–120.

[32] Josef Sivic and Andrew Zisserman. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*. 1470–1477.

[33] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*. 4631–4640.

[34] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*. 4489–4497.

[35] 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*. 5998–6008.

[36] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*. 20–36.

[37] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020. On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. In *CVPR*. 10123–10132.

[38] 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 *ACM MM*. 1645–1653.

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*. 5288–5296.

[40] Youjiang Xu, Yahong Han, Richang Hong, and Qi Tian. 2018. Sequential Video VLAD: Training the Aggregation Locally and Temporally. *IEEE Transactions on Image Processing* 27, 10 (2018), 4933–4944.

[41] 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*. 829–832.

[42] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In *CVPR*. 3261–3269.

[43] Zheng-Jun Zha, Jiawei Liu, Tianhao Yang, and Yongdong Zhang. 2019. Spatiotemporal-Textual Co-Attention Network for Video Question Answering. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2s (2019), 1–18.

[44] J. Zhang and Y. Peng. 2020. Video Captioning With Object-Aware Spatio-Temporal Correlation and Aggregation. *IEEE Transactions on Image Processing* 29 (2020), 6209–6222.

[45] J. Zhang, J. Shao, R. Cao, L. Gao, X. Xu, and H. T. Shen. 2020. Action-Centric Relation Transformer Network for Video Question Answering. *IEEE Transactions on Circuits and Systems for Video Technology* (2020), 1–1.

[46] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. 2020. Frame Augmented Alternating Attention Network for Video Question Answering. *IEEE Transactions on Multimedia* 22, 4 (2020), 1032–1041.

[47] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks. In *IJCAI*. 3518–3524.