information

Article

A Video Question Answering Model Based on Knowledge Distillation

Zhuang Shao [1], Jiahui Wan [2] and Linlin Zong [2,3,*]

[1] China Academy of Space Technology, Beijing 100094, China
[2] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China; 1393428962@mail.dlut.edu.cn
[3] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
* Correspondence: llzong@dlut.edu.com

Abstract: Video question answering (QA) is a cross-modal task that requires understanding the video content to answer questions. Current techniques address this challenge by employing stacked modules, such as attention mechanisms and graph convolutional networks. These methods reason about the semantics of video features and their interaction with text-based questions, yielding excellent results. However, these approaches often learn and fuse features representing different aspects of the video separately, neglecting the intra-interaction and overlooking the latent complex correlations between the extracted features. Additionally, the stacking of modules introduces a large number of parameters, making model training more challenging. To address these issues, we propose a novel multimodal knowledge distillation method that leverages the strengths of knowledge distillation for model compression and feature enhancement. Specifically, the fused features in the larger teacher model are distilled into knowledge, which guides the learning of appearance and motion features in the smaller student model. By incorporating cross-modal information in the early stages, the appearance and motion features can discover their related and complementary potential relationships, thus improving the overall model performance. Despite its simplicity, our extensive experiments on the widely used video QA datasets, MSVD-QA and MSRVTT-QA, demonstrate clear performance improvements over prior methods. These results validate the effectiveness of the proposed knowledge distillation approach.

Keywords: video question answering; multimodal fusion; knowledge distillation

Citation: Shao, Z.; Wan, J.; Zong, L. A Video Question Answering Model Based on Knowledge Distillation. *Information* **2023**, *14*, 328. https://doi.org/10.3390/info14060328

Academic Editor: Willy Susilo

Received: 20 March 2023
Revised: 2 June 2023
Accepted: 3 June 2023
Published: 12 June 2023

1. Introduction

Video question answering (QA) [1] is a research task that evaluates a computer's ability to efficiently process video information through question answering. It is similar to earlier tasks, such as visual question answering (VQA) [2] and text question answering (TQA) [3]. Video QA has gained significant attention from researchers since its proposal. In this task, as shown in Figure 1, a video and a set of questions related to the video are provided, and the machine is expected to analyze the video content and understand the question content in order to provide accurate answers.

The video QA task presents unique difficulties and challenges that are not encountered in general QA tasks, placing higher demands on the model. Firstly, the nature of the complex questions in video QA requires a comprehensive understanding of various aspects of the question, including the way it is framed, its purpose, and the specific focus of the video. This complexity necessitates a thorough understanding of all question elements. Secondly, video processing introduces temporal dynamics that are absent in static images. For example, in action-based questions such as "What are men doing?", understanding the actions often requires analyzing a sequence of frames rather than a single static image. The machine needs to observe information within each frame, identify targets, analyze

their relationships, recognize motion features of objects throughout the sequence, and exhibit reasoning capabilities [4–6]. Additionally, video QA is a cross-modal task that involves processing information from multiple modalities, including video and textual questions. Effectively integrating and leveraging information from different modalities to derive accurate answers poses a significant challenge.

Figure 1. This is an example of video QA. Given a video and a series of questions about the video, the video QA task requires the machine to give the correct answer after analyzing the video content and understanding the question semantics.

Although today's video QA task has received extensive attention from the academic community compared with the earlier visual QA and text QA, its research status still has many deficiencies. With respect to feature extraction, studies always used the pretraining model to extract the appearance and motion features to represent the video, while using the word vectors to represent the text [1,4,6]. These are then input into two separated streams to obtain the latent representation which is finally fused into a visual representation. In the interaction and fusion between video and problem text, recent works have mainly used attention [7,8], graph convolution for feature enhancement and object-relational reasoning [5,6,9–11]. These methods can effectively extract the important frames and objects of interest in the video, and reason with the guidance of the question. However, they lack interaction and ignore the latent complex correlation between the appearance and motion of the video.

This paper introduces a novel video QA model that addresses the limitations mentioned earlier. The proposed model enhances the fusion between the appearance features and motion features while also compressing the overall model size using knowledge distillation techniques [12]. The approach starts by training a teacher model, which serves as a reference model. Based on the knowledge learned by the teacher model, a relatively simpler student model is constructed and trained to improve the overall performance. This approach reduces the number of trainable parameters in the model, making its volume less while maintaining or improving its performance. Importantly, the proposed model focuses on capturing the latent complex correlations between the appearance and motion of the video, which strengthens the feature fusion process. The knowledge obtained from the multimodal fusion in the teacher model is distilled and used for uni-modal learning in the student model. As a result, the student model can leverage rich multimodal information during the process of uni-modal training. This early-stage multimodal interaction enables improved fusion effects between the appearance and motion modalities. In summary, the proposed model not only compresses the overall model size but also emphasizes the latent complex correlations between appearance and motion in videos, leading to enhanced feature fusion. By leveraging knowledge distillation and a student-teacher model framework, the proposed approach achieves improved performance while reducing model complexity. The main contributions of our work can be summarized as follows:

(1) Teacher-student framework: We introduce a teacher-student framework leveraging knowledge distillation techniques. This framework allows for the training of a simpler student model in a more convenient and efficient manner. By distilling the knowledge learned by the teacher model, the student model benefits from the expertise while maintaining a reduced model size.

(2) Multimodal knowledge distillation: We propose a novel approach to multimodal knowledge distillation. This technique enables the student model to acquire rich multimodal information during the training process of individual modalities. By incorporating multimodal interactions early on, the fusion of appearance and motion features is significantly enhanced.
(3) Competitive results on MSVD-QA and MSRVTT-QA: Through extensive experiments, we demonstrate the effectiveness of our proposed model on the popular MSVD-QA and MSRVTT-QA datasets. Our model achieves competitive performance compared to existing approaches, showcasing its capabilities in video question answering tasks.

2. Related Work

The video QA task was introduced later than the general QA task and its development has been relatively slow due to the challenges in collecting video QA data and the complexity of video semantic analysis. However, with the continuous construction and improvement of datasets and the advancements in deep learning technology, research on video QA tasks has made significant progress. Various modeling methods have emerged, attracting significant attention from the academic community. The video QA datasets primarily fall into three categories: film and television, real-life, and generated datasets. Film and television datasets comprise video clips from movies and TV shows, including datasets such as MovieQA [13] and TVQA [14]. Real-life datasets consist of videos that capture daily life scenes, making them more applicable in practical scenarios. An example of such a dataset is LifeQA [15]. Generated datasets involve automatically generating videos with different virtual geometric objects. For instance, the SVQA dataset [16] contains videos generated using the Unity3D tool. Video QA datasets take various forms, including video retrieval, selection, filling in the blanks, and other methods. The answers in these datasets are generally predicted through classification techniques.

In recent years, research on the video QA task has been advancing steadily, and a common solution can be abstracted into a basic video QA framework. This framework comprises video feature extraction, question feature extraction, multimodal fusion, and final answer generation. Video feature extraction typically involves extracting static appearance features and dynamic motion features. For static appearance feature extraction, a common approach is to utilize pretrained models on ImageNet [17]. Network models such as VGG and ResNet [18] are commonly employed for this purpose. Dynamic motion features are typically extracted using pretrained models trained on the Kinetics dataset [19]. The C3D model [20] is a popular choice for extracting dynamic motion features. Subsequent research has sought to enhance the performance of the model by refining each module. Various improvements and optimizations have been explored by modifying the details of each component in the video QA framework.

For the extraction of textual features, pretrained word vectors are primarily utilized to encode each word, representing them as fixed-length vectors. Common techniques include Word2Vec, Glove [21], and BiLSTM. In the context of video QA tasks, the research has focused on the interaction and fusion of video appearance features, video motion features, and question text features. Various implementation methods have emerged, such as attention mechanisms, graph convolution networks, and more. Jang et al. [22] employed attention in both temporal and spatial dimensions to fuse video and question features, identifying crucial areas in key video frames and classifying their resulting features. Kim et al. [23] introduced memory mechanisms to enable the model to learn deeper representations and the meanings of features. Xu et al. [1] proposed an attention memory unit (AMU) based on dynamic memory network (DMN) principles, continually improving video feature attention through text-based cues. Gao et al. [4] considered the correlation between appearance and motion features, proposing the co-memory network method and utilizing dynamic attention to learn video features. Zhang et al. [7] explored convolutional approaches instead of recurrent neural networks and proposed hierarchical convolutional self-attention networks (HCSA), incorporating attention mechanisms at each

stage to continuously focus on problem-related information. While these methods primarily employ attention and memory mechanisms to represent learning, they often overlook object relationships and may have limitations in reasoning. Le et al. [24] introduced the conditional relationship network (CRN) to model relationships between visual objects, but it may be less efficient in dealing with multiple object relationships. With the emergence of graph convolution networks, Wang [5] and others leveraged this approach to perform object-relationship reasoning in behavior recognition tasks, effectively improving the learning effects between objects and relationships. Consequently, it has gained widespread use in video QA tasks. Jiang et al. [9] proposed heterogeneous graph alignment (HGA), employing fusion alignment features of the problem and video as graph nodes to perform graph convolution operations and infer relationship representations within and between modalities using an undirected heterogeneous graph. Huang et al. [10] proposed the location-aware graph convolution network (LGCN), which combines time embedding and position embedding and considers the interaction between objects in each frame.

Finally, when generating answers in video QA tasks, the common approach is to employ classification. Based on the probability distribution of each candidate answer calculated within a predefined set of possible answers, the cross-entropy loss function is utilized to compute the loss during training. During validation, the predicted answer is determined as the one with the highest probability.

3. Materials and Methods

In the video QA task, the goal is to classify the correct answer to a given question by comprehending both the video content and the question itself. The answer choices are predefined and form a fixed set of possible answers. To tackle this task, our proposed approach utilizes a knowledge-distillation-based video QA model. This model acts as an answer classification model, taking multimodal features, including videos and questions, as its input.

This paper models the overall framework of the teacher model according to DualVGR [6], which is shown in Figure 2. Based on this foundation, to compress the model and leverage the abundant multimodal knowledge of a larger model to enhance the feature learning process of a smaller model, this paper introduces a multimodal knowledge distillation approach to further enhance model performance. The teacher model and the student model share the same model structure, with only a slight difference in the number of graph layers. The teacher-student training structure of this approach is illustrated in Figure 3. The teacher model consists of two separate stacking modules for the appearance and motion modalities, which are trained individually. Through experimental adjustments of parameters, an optimal teacher model is obtained. Subsequently, the student model is constructed with fewer stacking modules, resulting in a simplified model. The knowledge distillation process involves transferring the fused visual features from the teacher model as "soft labels." These soft labels serve as guidance for the student model's learning of the appearance and motion features, respectively.

3.1. Teacher Model

3.1.1. Encoder

In the visual encoder module, capturing both the static spatial features and dynamic temporal features of the video is essential due to their spatio-temporal nature. To extract the appearance features, we employ the pretrained model ResNet-101 and process each video clip using BiLSTM. The resulting static appearance features are denoted as V_a. For the motion features, we utilize the pretrained model ResNeXt-101. The extracted dynamic motion features are denoted as V_m.

In the text encoder module, we aim to capture the word feature representation of the question sentence. To achieve this, we utilize the pretrained GloVe word vectors to encode and represent each word in the question. This process results in obtaining the word feature of the question, denoted as Q_w. Furthermore, to extract both the contextual features of

words and the semantic features of the entire sentence, we employ two BiLSTMs. One LSTM focuses on extracting the embedded features of each sentence, denoted as Q_e, while the other LSTM focuses on capturing the semantic features of the sentence, denoted as q.

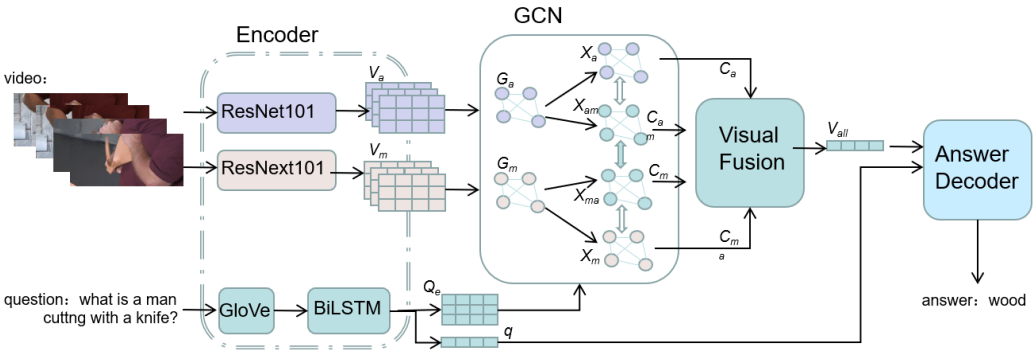

Figure 2. This is the overall framework of the teacher model, which includes four modules. First, an encoder extracts and represents the video and question. Then, visual–text interaction achieves the reasoning between the visual and textual features. The visual fusion is used to fuse the appearance and motion feature to the visual representation. Finally, answer generation predicts the answer using a decoder.

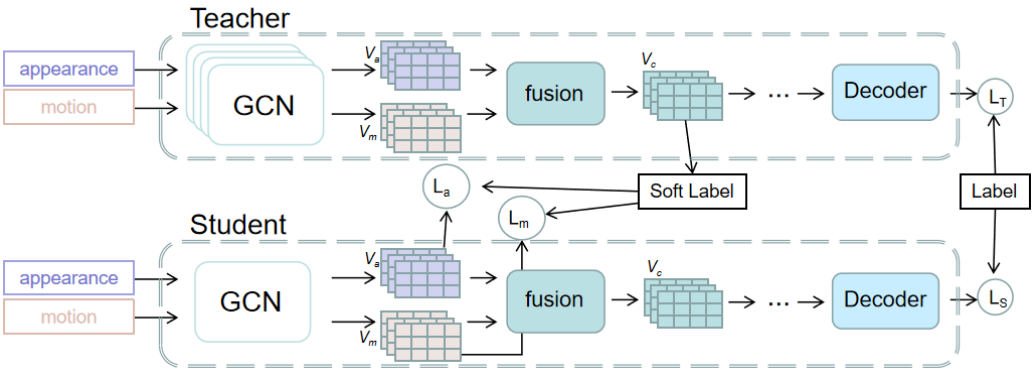

Figure 3. Multimodal knowledge distillation architecture. The teacher model with multiple GCN layers is first trained. Then, its fusion visual feature is used to guide the training of visual features in the student model, which has just one GCN layer.

3.1.2. Visual–Text Interaction

To emphasize the importance of certain words and downplay the significance of others in the word embedding of the text, the model incorporates the attention mechanism. This allows for the optimization of features by assigning attention scores to each word. These scores are calculated using the embedded feature Q_e as weights, and the word features Q_w are then weighted and summed to obtain the overall representation of the question, denoted as Q_{att}. The calculation formula is as follows:

$$\alpha = Softmax(L2Norm(Q_e W_1)W_2) \quad (1)$$

$$Q_{att} = \alpha^T Q_w \quad (2)$$

where W_1 and W_2 are learnable parameters.

In a video, there are multiple clips, and, when answering a question, the question may only pertain to a subset of these clips. The model needs to focus on understanding these specific clips in order to answer the question accurately, without considering the entire video. To achieve this, question-guided attention is applied to each clip of the video, determining the attention score. This allows the model to prioritize certain clips while appropriately disregarding others. This module follows the same processing steps for both

the appearance features (a) and motion features (m) of the video, which are uniformly represented as a/m. The attention score $S_{a/m}$ for each clip can be calculated as follows:

$$S_{a/m} = Sigmoid(V_{a/m} Q_{att} W_{a/m}) \tag{3}$$

where $W_{a/m}$ are learnable parameters of the appearance and motion features.

To capture the relationships between clips in the video and achieve a deeper feature representation, this approach draws inspiration from DualVGR [6], which combines the graph convolution network (GCN) model of GAT (graph attention network) and the concept of AM-GCN (attribute-matching graph convolutional network) for relational reasoning. GAT utilizes a multi-head graph convolution (multi-head GCN) strategy and incorporates an attention mechanism to effectively represent the relationships between nodes. In the teacher model, multiple layers of graph convolution are used to fully leverage the information obtained through graph convolution. In contrast, the student model aims to compress the model and reduce trainable parameters, so it only employs a single layer of graph convolution during training. Additionally, following the idea of AM-GCN, this method not only extracts the independent features of appearance and motion, but also obtains the joint features that fuse motion information into appearance and vice versa. By applying a loss function constraint, the approach emphasizes the distinction between independent features and joint features while encouraging similarity among joint features. This module processes appearance and motion features in a similar manner, which is why they are collectively represented as a/m.

Firstly, the visual feature $V_{a/m}$, serving as the input to the graph convolution network, undergoes multiple iterations of the graph convolution operation to obtain independent and joint features. The formulas for calculating these features are as follows:

$$G_0^{a/m} = V^{a/m} \tag{4}$$

$$X_i^{a/am} = GCN_i^{a/am}(G_{i-1}^a, S^a) \tag{5}$$

$$X_i^{m/ma} = GCN_i^{m/ma}(G_{i-1}^m, S^m) \tag{6}$$

$$G_i^{a/m} = X_i^{a/m} \tag{7}$$

where G_{i-1} represents the input features of the i-th layer graph convolution, and X_i represents the output features of the i-th layer graph convolution. The subscript am represents the joint feature obtained by fusing the appearance information in the appearance, while the subscript ma represents the joint feature obtained by fusing the appearance information in the motion. GCN_i represents the graph convolution operation at the i-th layer, where i is less than or equal to g.

In each layer of GCN, the input feature G is passed through k graph convolution heads for processing, and the results of each head are concatenated. Firstly, the input feature G is mapped to a dimension of d_k using a linear layer, and this mapped feature is denoted as the visual feature g for further processing by the graph convolution heads. Each g_j is then multiplied by the visual attention score S to obtain the visual feature under attention, denoted as h_j. Additionally, in order to effectively represent the relationships between video clips, g is transformed into an undirected and fully connected graph g'. The attention mechanism is utilized to obtain the weights of the relationships between each pair of nodes, denoted as β_j. The formula for obtaining β_j is as follows:

$$g_j' = Graph(g_j) \tag{8}$$

$$\beta_j = Softmax(ReLU(g_j W)) \tag{9}$$

where $Graph$ represents the operation of constructing a fully connected undirected graph, and W is a learnable parameter. Then, the feature x_j of each segment is obtained by weighting the relationships between each clip. Finally, the features x_j obtained from the convolution of multiple graph convolution heads are concatenated, resulting in the output

feature of the layer graph denoted as X. With the above process, we obtain the independent feature C_a and joint feature C_{am} of appearance, and the independent feature C_m and joint feature C_{ma} of motion.

3.1.3. Visual Fusion

In the process of fusing visual features, an attention mechanism is used to emphasize the importance of independent features and joint features. This results in the fused visual features F_a, which can be represented by the following formula:

$$[\alpha_a, \alpha_{am}] = Softmax(tanh([C_a, C_{am}]W_1)W_2) \tag{10}$$

$$F_a = \alpha_a C_a + \alpha_{am} C_{am} \tag{11}$$

where W_1 and W_2 are learnable parameters, resulting in the final fused appearance feature F_a. Similarly, the fused motion features F_m are obtained using the same process.

At the same time, to address the issue of vanishing gradients during back-propagation, this method employs the residual connection strategy [18] to obtain the final visual feature $V_{a/m}$ that incorporates the relationship information. The formula for obtaining $V_{a/m}$ with residual connection is as follows:

$$V_{a/m} = V_{a/m} + F_{a/m} \tag{12}$$

Then, the visual appearance features and visual motion features are fused. This method utilizes the multimodal factored bilinear pooling [25] approach to obtain the fused visual features for each video clip, denoted as V_c. Furthermore, drawing inspiration from the graph readout operation [26], the fusion feature representation V_{all} for the entire video is obtained.

3.1.4. Answer Generation

This module integrates the visual features with the semantic features of the question, decodes these features, and generates the final answer. The model employs a standard decoder to decode the answer, resulting in the final feature p used for answer generation. The formula for obtaining p is as follows:

$$y = [V_{all}, qW_1] \tag{13}$$

$$p = ELU(yW_2)W_3 \tag{14}$$

where W_1, W_2 and W_3 are learnable parameters, and n is the number of answer categories.

3.2. Student Model

This method employs a multimodal knowledge distillation approach. It distills the knowledge obtained from multimodal fusion in the teacher model and utilizes it for unimodal learning in the student model. By doing so, the student model can leverage the rich multimodal knowledge of the teacher model during unimodal training, allowing for early-stage interaction and fusion between multimodal models. This approach aims to enhance the effectiveness of multimodal fusion in the later stages, leading to improved model performance. Moreover, this method not only reduces the model's complexity but also enhances the learning capabilities of multiple modal features.

As previously mentioned, the teacher model utilizes multi-layer graph convolution to iteratively process visual features. This approach enables the model to perform multi-step reasoning on video relationships and extract relationship features effectively through multiple graph convolution layers. On the other hand, the student model, designed for compression purposes, employs only one graph convolution layer to process visual features, resulting in a smaller model size. During the training stage, this method leverages the knowledge obtained from multimodal fusion in the teacher model to guide the unimodal learning of the student model. This guidance aims to optimize the learning process and

improve the overall performance of the student model. The training details of the student model can be found in Figure 3.

The method of knowledge distillation is based on the teacher-student framework. In this work, the teacher model trains its appearance and motion modal features separately using several stacking modules. Through experimentation and parameter adjustments, the optimal teacher model is obtained. Subsequently, the student model is constructed with fewer stacking modules, making it a simpler model. The knowledge distillation process involves transferring the fused visual features from the teacher model, referred to as "soft labeling", to guide the learning of the appearance and motion features in the student model. More specifically, the fused feature V_c from the teacher model is distilled into knowledge and used to guide the learning of the appearance feature V_a and the motion feature V_m in the student model. The details of the loss function calculation will be described in the next section. This approach allows the student model to learn from the knowledge of the other modality during the unimodal processing of appearance or motion, enhancing the interaction and fusion between modalities. Additionally, this compensates for the reduced number of graph convolution iterations, ensuring sufficient information extraction.

3.3. Loss Function

In the vision–text interaction module of the model, the objective is to learn both the independent knowledge of each visual modality (appearance and motion) and the knowledge associated with the other modality. To achieve this, four features are generated in the graph convolution of each layer for appearance and motion: the independent features G_a and joint feature C_{am} for appearance, and the independent feature G_m and joint feature G_{ma} for motion. The model aims to have a significant difference between the independent and joint features, while keeping the difference between the joint features small. To achieve this, the method draws inspiration from DualVGR and employs the Hilbert–Schmidt independence criterion (HSIC) [27] to constrain the differences between features at each layer, resulting in the matrix distance L_1. Additionally, a similarity constraint mechanism using a matrix distance method is applied to ensure similarity between the joint features, resulting in the matrix distance L_2. By incorporating these constraints, the model can effectively extract both the information specific to each visual modality and the information associated with other modalities, enhancing the representation of multimodal interactions.

The cross-entropy loss function is used to calculate the loss of the predicted probability distribution and the real label. The formula is as follows:

$$L_T = -\sum z \ln y \tag{15}$$

where y represents the predicted probability distribution, and z represents the real probability distribution.

To sum up, the loss function of the teacher model can be finally denoted as follows:

$$L_{total}^{T} = L_T + \gamma L_1 + \eta L_2 \tag{16}$$

Among them, the coefficients γ and η represent superparameters, which can be adjusted to optimize the model.

The student model includes the loss mentioned in the teacher model, L_T, L_1 and L_2. Additionally, it also includes the loss of distillation knowledge to learn from teachers. In order to learn the knowledge of other modals in the early stage of uni-modal learning, this method combines the teacher model with the feature V_c. Its knowledge is distilled to guide the appearance feature of the student model V_a and the motion feature V_m to improve the interaction and integration between the various modals. First, the Softmax activation function is normalized at the appropriate temperature T, and then the cross-entropy is used to calculate the loss. The loss function is as follows:

$$L_{a/m}^{T} = L_0(Softmax(V_c^t/T_{a/m}), Softmax(V_{a/m}^s/T_{a/m})) \tag{17}$$

where L_0 represents the cross-entropy loss calculation operation, V_c^t is the fusion feature of the teacher model, $V_{a/m}^s$ is the appearance and motion feature of the student model, $T_{a/m}$ is the temperature of knowledge distillation for appearance and motion. The temperature is used to justify the soft label's distribution. Its distribution is smoother while the temperature is higher. The weight is used to justify the knowledge distillation's influence on the total loss.

Therefore, the total loss of the student model can be denoted as follows:

$$L_{total}^S = L_T + \gamma L_1 + \eta L_2 + \lambda_a L_a + \lambda_m L_m \tag{18}$$

where the coefficients γ and η directly use the coefficients of the teacher model, λ_a and λ_m is a superparameter, and the model is optimized by adjusting them.

4. Experimental Results Analysis

4.1. Dataset

This method selects the widely used video QA dataset, MSVD-QA [28] and MSRVTT-QA [29] for the experiment. The MSVD-QA dataset consists of 1970 videos and 50,505 question-answer pairs. It is a popular choice for evaluating Video QA tasks due to its rich and diverse instances. The video content in this dataset is extracted from the Microsoft Research Video Description (MSVD) dataset. These videos mainly depict short daily life scenes, and the questions are generated programmatically based on the video description. On average, the videos have a duration of around 10 s, and the questions have an average length of approximately 6 words. MSRVTT-QA is a larger dataset comprising 10,000 videos and 243,000 question-answer pairs. The videos in this dataset have an average duration of approximately 15 s, and the questions have an average length of around 7 words.

In both the MSVD-QA and MSRVTT-QA datasets, each video has an average of about 25 questions. The questions in these datasets are open-ended and can be categorized into five different types: what, where, how, who, and when. However, there is a significant imbalance in the distribution of the question types. Taking MSVD-QA as an example, the majority of the questions belong to the 'what' and 'who' types, accounting for more than 96% of the total questions. Among these, the 'what' type questions are the most prevalent, making up more than 62% of the questions. On average, each video has around 16 'what' type questions. On the other hand, the 'how', 'when', and 'where' types are relatively rare, collectively accounting for about 3.3% of the total questions. Many videos do not have any questions belonging to these three types.

4.2. Implement Details

4.2.1. Teacher Model

In the video preprocessing stage, each video is divided into a fixed number of equally spaced clips. For the MSVD-QA dataset, the number of clips is set to 8, while for the MSRVTT-QA dataset, it is set to 16. Each clip consists of a certain number of image frames, where the number of frames per clip is 16. If a clip contains fewer than 8 frames at either the beginning or the end, it is padded with the first frame or the last frame to reach the required length. The dimension size of the model, denoted as d, is set to 768. This represents the size of the visual and textual feature vectors used in the model. In the encoder module, a bidirectional LSTM (BiLSTM) is employed for both visual coding and text coding. The BiLSTM is configured with a single layer. In the vision–text interaction module, the number of layers of graph convolution, denoted as g, is set to 4 for the MSVD-QA dataset and 6 for the MSRVTT-QA dataset. Additionally, the number of graph convolution heads, denoted as k, is set to 4. In the visual fusion module, the number of factors, denoted as f, is set to 4 for the multimodal factorized bilinear pooling.

In the loss function, the constraint loss coefficients γ and η of the independent and joint features are set to 100 and 1×10^{-6}, respectively. The optimizer used in the training

process is the Adam optimizer, the learning rate of training is set to 1×10^{-4}, the batch size of training data is 256, and the number of training iterations is 25.

4.2.2. Student Model

In the student model, the parameter configuration is kept the same as the teacher model, except for the number of layers in the graph convolution. Specifically, the student model utilizes a single-layer graph convolution operation to process the visual features, while maintaining the same parameter settings as the teacher model. By using only one layer of graph convolution, the student model achieves model compression by reducing the network size and complexity compared to the teacher model.

In the knowledge distillation loss function, the parameters for distilling the visual appearance and motion features are determined through a grid search experiment. The search is conducted to find the relatively appropriate configuration for these parameters. For the distillation of visual appearance features, the final temperature parameter T_a is set to 1, indicating a standard distribution. The coefficient parameter λ_a is set to 1. For the distillation of visual motion features, the temperature parameter T_m is set to 0.7 for MSVD-QA and 1 for MSRVTT-QA. The coefficient parameter λ_m is set to 100 for MSVD-QA and 1 for MSRVTT-QA.

4.3. Results Analysis

4.3.1. Visual Analysis

Analyzing the changes in accuracy and loss during the training process can provide insights into the model's learning dynamics and performance. By visualizing these metrics, we can observe how the model progresses over time and identify any potential issues or improvements.

The loss changes of the training set and the validation set during the training process in MSVD-QA are shown in Figures 4 and 5, which, respectively, represent the real label loss, the appearance feature distillation loss and the motion feature distillation loss. The continuous decrease and convergence of the loss values for both the training set and the validation set indicate that the training process is effective. This means that the model is learning and making progress in fitting the data. Furthermore, the decrease in the distillation loss of the appearance and motion features indicates that the knowledge distillation process is effective in transferring the knowledge from the teacher model to the student model.

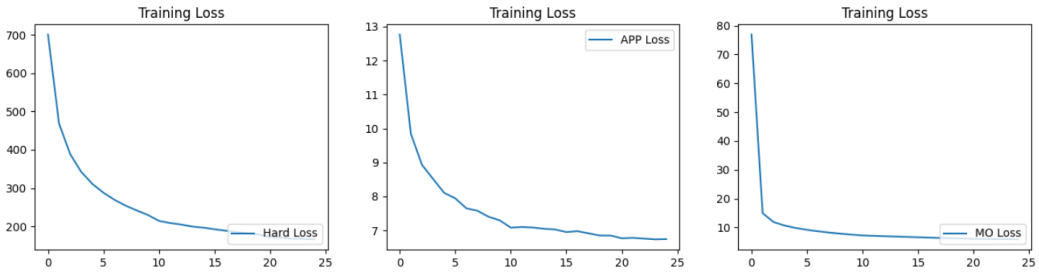

Figure 4. The loss changes of the training set during the training process represent the real label loss, the appearance feature distillation loss and the motion feature distillation loss, respectively. The hard loss is calculated by output and real labels; the APP and MO loss are calculated by features and soft labels.

Then, in order to further analyze the prediction performance of the model, some examples of incorrect prediction are selected from the prediction results of the model, as shown in Figure 6. The shortcomings and improvements of the model in the video QA task can be analyzed based on the errors in the model predictions. In question 1, due to the limited number of words in the word list, some words need to be represented by '<UNK>', which leads to the model being unable to understand the semantics of some words, and, thus, unable to accurately predict the answers. In question 2, the model incorrectly predicting "rabbit" instead of "bear" suggests a limitation in target recognition.

The similarity between the brown background and the color of the bear might have misled the model. In question 3, the model providing a verb "chase" instead of identifying the target object indicates a deficiency in sentence semantics understanding. Regarding the errors in question 4 and question 5, where the model's predicted answers may align with human understanding but do not match the fixed answers in the dataset, this could be attributed to the limitations of the dataset itself.

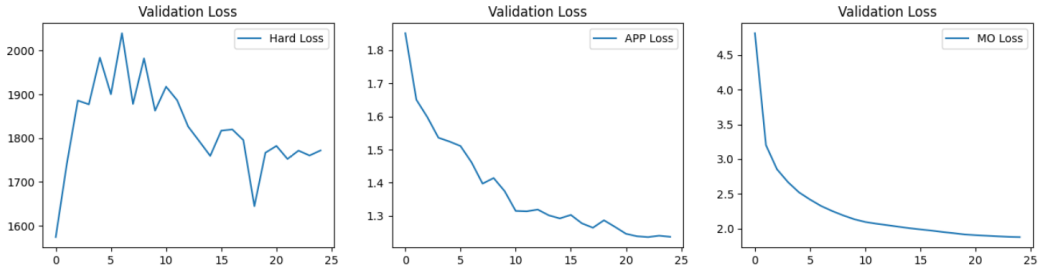

Figure 5. The loss changes of the validation set during the training process represent the real label loss, the appearance feature distillation loss and the motion feature distillation loss, respectively. The hard loss is calculated by output and real labels; the APP and MO loss are calculated by features and soft labels.

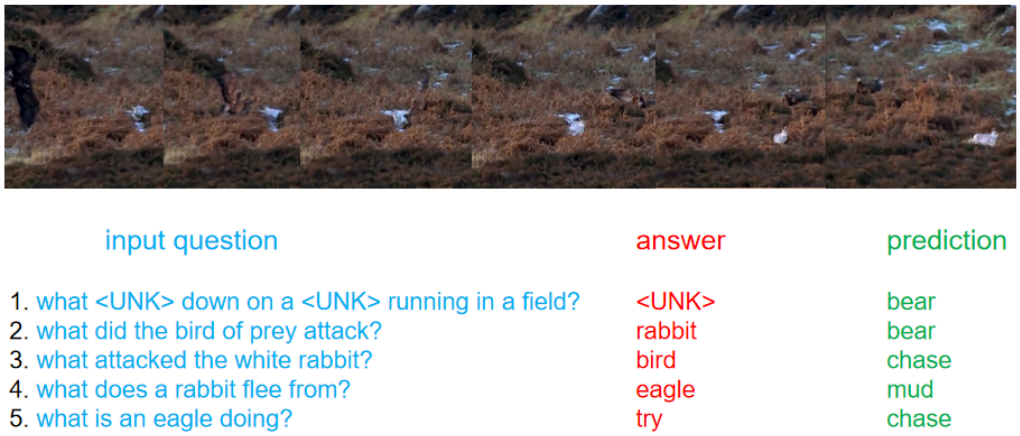

Figure 6. Examples of some false predictions.

4.3.2. Comparative Analysis

In order to analyze and verify the effectiveness of this method in video QA tasks, the most advanced model algorithms based on video QA tasks are compared. Next, these algorithms are briefly introduced, and then the experimental results are compared and analyzed.

- Co-Mem [4]. This method is developed from the dynamic memory network (DMN) in visual QA, and is improved based on video QA. In the context memory module, the attention mechanism of appearance-action collaborative memory is introduced, and the convolution-deconvolution network, based on time-series and the dynamic fact integration method, is used to mine video information deeply.
- AMU [1]. The algorithm is an end-to-end video QA model, which applies the fine-grained features of the question to video understanding. It reads the words in the question word-by-word, interacts with the appearance features and motion features through the attention mechanism, constantly refines the video attention features, and finally obtains the video understanding that integrates the different scale features of the problem.
- HGA [9]. The graph network is introduced into the model for reasoning learning. It constructs the video clip and the question word into the form of a graph, and carries out a cross-modal graph reasoning learning process.

- HCRN [24]. This is a stackable model of relational network modules based on clips. The relational network takes the input as a set of tensor objects and a conditional feature, outputs a set of relational information containing them, and then realizes multi-step reasoning of the relational information by hierarchically stacking the network modules.
- DSAVS [8]. The answer to the question may be deduced from a few frames or fragments in the video, and the appearance and motion information are generally complementary. To this end, the author proposes a visual synchronization dynamic self-attention network, which selects important video clips first and synchronizes various features in time.
- DualVGR [6]. This model is a stacked model of an attention graph inference network. In the attention graph inference network module, the query punish mechanism is used to strengthen the features of key video clips, and then the relationship is modeled by the multi-head graph network combined with attention. The model performs multi-step reasoning of the relationship information by stacking the network module.

Tables 1 and 2 summarize the comparison of the experimental results of each model on the MSVD-QA and MSRVTT-QA datasets, which shows a competent result for our method. The observation that the proposed method surpasses the DualVGR model in terms of accuracy demonstrates the effectiveness of the approach. This method successfully improved the accuracy compared to other comparison models on the MSVD-QA and MSRVTT-QA datasets. This suggests that knowledge distillation serves as an effective technique for reducing model complexity while enhancing the performance of cross-modal information transmission and fusion, thereby improving the feature extraction capabilities of individual modalities. Overall, the results indicate that knowledge distillation not only enables model compression but also enhances the overall performance of the model. By distilling knowledge from the teacher model to guide the learning process of the student model, the proposed approach achieves improved accuracy in video question answering tasks, surpassing existing models.

Table 1. Comparison to other models on MSVD-QA.

Model	What	Who	How	When	Where	All
Co-Mem	19.6	48.7	81.6	74.1	31.7	31.7
AMU	20.6	47.5	83.5	72.4	53.6	32.0
HGA	23.5	50.4	83.0	72.4	46.4	34.7
HCRN	/	/	/	/	/	36.1
DSAVS	25.6	53.5	**85.1**	**75.9**	**53.6**	37.2
DualVGR	28.7	53.8	80.0	70.7	46.4	39.0
ours	**29.22**	**53.98**	80.81	74.14	53.57	**39.48**

Best result in bold.

Table 2. Comparison to other models on MSRVTT-QA.

Model	What	Who	How	When	Where	All
Co-Mem	23.9	42.5	74.1	69.0	**42.9**	32.0
AMU	26.2	43.0	80.2	72.5	30.0	32.5
HGA	29.2	45.7	83.5	75.2	34.0	35.5
HCRN	/	/	/	/	/	35.6
DSAVS	29.5	**46.1**	**84.3**	75.5	35.6	**35.8**
DualVGR	29.4	45.6	79.8	76.7	36.4	35.5
ours	**29.67**	45.51	80.91	**76.51**	35.20	35.71

Best result in bold.

4.3.3. Ablation Study

In this method, it is mainly proposed to compress the model through knowledge distillation and strengthen the cross-modal feature learning and fusion to achieve the purpose

of improving the model. In order to verify the effectiveness of knowledge distillation, an ablation study was conducted.

In MSVD-QA, the learnable parameters in the constructed 'Teacher' model include about 31.19 million parameters. This method trains the teacher model and adjusts the parameters to achieve the optimal parameter configuration of the teacher model. Finally, the accuracy achieved on the test set is 39.03%. Then, this method constructs a relatively simple student model by reducing the number of layers of graph convolution, and its learnable parameters are reduced to about 24.09 million. In the case that other parameter configurations are the same as the teacher model, this method first trains the student model separately, which is denoted as 'Student' here. Then, through the method of multi-modal knowledge distillation, the teacher model guides and trains the student model, and we optimize the knowledge distillation temperature and weight. The model with extra knowledge is denoted as 'Student-kd'. The results of the whole ablation study are shown in Table 3. Similar to MSVD-QA, the experimental results for the MSRVTT-QA dataset also demonstrate this phenomenon, as shown in Table 4. The student model has poor strength to represent the visual semantics due to the less learnable parameters. However, it demonstrates excellent accuracy when guided by the teacher model, even higher than that of the teacher model.

Table 3. Ablation Study on MSVD-QA.

Model	Accuracy	Number of Trainable Parameters
Teacher	39.03%	31.19 million
Student	38.85%	24.09 million
Student-kd	39.48%	24.09 million

Table 4. Ablation Study on MSRVTT-QA.

Model	Accuracy	Number of Trainable Parameters
Teacher	35.52%	41.29 million
Student	35.11%	29.09 million
Student-kd	35.71%	29.09 million

By comparing the test results of 'Student' and 'Student-kd', it can be observed that both models have the same architecture and the same number of trainable parameters. However, 'Student-kd', which benefits from knowledge distillation and the learned cross-modal features from the teacher model, exhibits higher accuracy. This suggests that knowledge distillation effectively improves the inter-modal fusion. Prior to the fusion module, the individual modalities, such as appearance, can acquire multimodal knowledge, enabling them to have a pre-inclination towards the fusion distribution. This approach helps avoid unstable fusion of the appearance and motion features.

Furthermore, comparing the test results of 'Teacher' and 'Student-kd', it becomes evident that knowledge distillation significantly reduces the model size and the number of trainable parameters, while maintaining or slightly increasing the accuracy. The prediction differences among these three models are illustrated in Figure 7. For easy questions, all three models achieve correct predictions. However, in more challenging scenarios, due to its limited parameter learning, the student model struggles to arrive at the correct answer. In such cases, the teacher's knowledge effectively guides the student in making the right choice. Notably, when faced with exceptionally difficult questions that even the teacher model struggles with, the student model, equipped with rich multimodal knowledge, surpasses the teacher and achieves accurate predictions.

These results highlight that, in this approach, knowledge distillation not only reduces the model size but also enhances the feature fusion between modalities, leading to improved performance and feature enhancement.

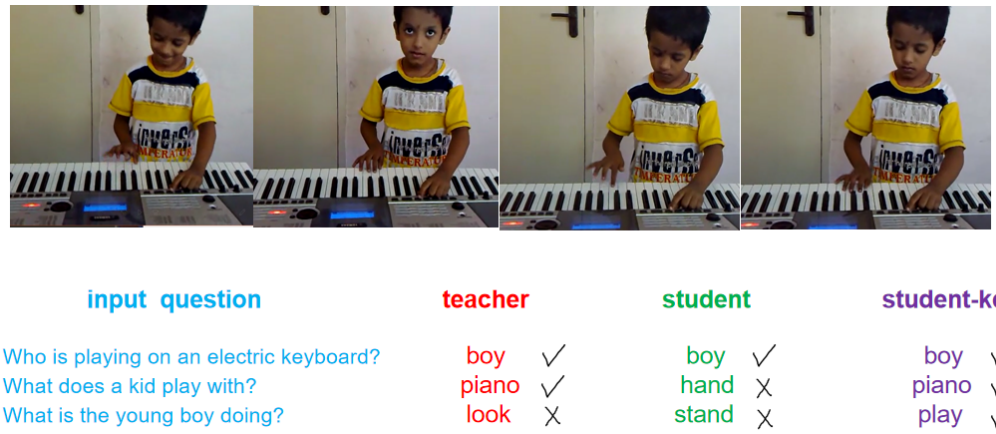

Figure 7. Examples of prediction difference of three models.

5. Conclusions

This paper introduces a novel video question answering model that utilizes knowledge distillation to address the challenge of capturing the latent complex correlation between appearance and motion in videos. While existing methods, such as attention mechanisms and graph convolution networks, enhance the attention of visual and text-specific features and reasoning about video relationships, they often overlook the interaction between appearance and motion. To overcome this limitation, our proposed approach leverages knowledge distillation to uncover the latent correlation between the static appearance and dynamic motion features in videos. By distilling the knowledge from a teacher model, our method strengthens the fusion of appearance and motion features while compressing the model. This enables the student model to learn from the rich multimodal knowledge of the teacher model, improving the interaction and fusion between appearance and motion features. The ablation experiments and comparisons conducted in our study validate the effectiveness of our approach in visual feature learning and its ability to enhance video QA performance. However, we acknowledge a limitation in our work, which is that the proposed knowledge distillation method is currently limited to the fusion of appearance and motion features in videos. As part of our future work, we plan to explore the application of knowledge distillation in other types of video features beyond appearance and motion.

Author Contributions: Conceptualization, Z.S. and L.Z.; methodology, Z.S. and J.W.; software, J.W.; validation, Z.S., J.W. and L.Z.; formal analysis, Z.S.; investigation, Z.S.; resources, L.Z.; writing—original draft preparation, Z.S. and J.W.; supervision, L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Social Science Planning Foundation of Liaoning Province under Grant L21CXW003; in part by the State Key Laboratory of Novel Software Technology, Nanjing University under Grant KFKT2022B41; and in part by the Dalian High-level Talent Innovation Support Plan under Grant 2021RQ056.

Data Availability Statement: Data available in a publicly accessible repository.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; Zhuang, Y. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In Proceedings of the 25th ACM International Conference on Multimedia, San Francisco, CA, USA, 23–27 October 2017; pp. 1645–1653.
2. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Parikh, D. Vqa: Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 2425–2433.
3. Gupta, P.; Gupta, V. A Survey of Text Question Answering Techniques. *J. Comput. Appl.* **2012**, *53*, 1–8. [CrossRef]
4. Gao, J.; Ge, R.; Chen, K.; Nevatia, R. Motion-appearance Co-memory Networks for Video Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6576–6585.
5. Wang, X.; Gupta, A. Videos as Space-time Region Graphs. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 399–417.
6. Wang, J.; Bao, B.; Xu, C. DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering. *IEEE Trans. Multimed.* **2022**, *24*, 3369–3380. [CrossRef]
7. Zhang, Z.; Zhao, Z.; Lin, Z.; Song, J.; He, X. Open-ended Long-form Video Question Answering via Hierarchical Convolutional Self-attention Networks. In Proceedings of the 28 International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 4383–4389.
8. Liu, Y.; Zhang, X.; Huang, F.; Shen, S.; Tian, P.; Li, L.; Li, Z. Dynamic Self-Attention with Vision Synchronization Networks for Video Question Answering. *Pattern Recognit.* **2022**, *132*, 108959. [CrossRef]
9. Jiang, P.; Han, Y. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11109–11116.
10. Huang, D.; Chen, P.; Zeng, R.; Du, Q.; Tan, M.; Gan, C. Location-Aware Graph Convolutional Networks for Video Question Answering. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11021–11028.
11. Wang, X.; Zhu, M.; Bo, D.; Cui, P.; Shi, C.; Pei, J. AM-GCN: Adaptive Multi-channel Graph Convolutional Networks. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 23–27 August 2020; pp. 1243–1253.
12. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
13. Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; Fidler, S. MovieQA: Understanding Stories in Movies through Question-Answering. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4631–4640.
14. Lei, J.; Yu, L.; Bansal, M.; Berg, T.L. Tvqa: Localized, Compositional Video Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1369–1379.
15. Castro, S.; Azab, M.; Stroud, J.; Noujaim, C.; Wang, R.; Deng, J.; Mihalcea, R. LifeQA: A Real-life Dataset for Video Question Answering. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 4352–4358.
16. Song, X.; Shi, Y.; Chen, X.; Han, Y. Explore Multi-step Reasoning in Video Question Answering. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Repubic of Korea, 22–26 October 2018; pp. 239–247.
17. Jia, D.; Wei, D.; Socher, R.; Li, L.J.; Kai, L.; Li, F.F. Imagenet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
19. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New model and the Kinetics Dataset. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
20. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4489–4497.
21. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
22. Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; Kim, G. Tgif-qa: Toward Spatio-temporal Reasoning in Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2758–2766.
23. Kim, K.M.; Heo, M.O.; Choi, S.H.; Zhang, B.T. Deepstory: Video Story QA by Deep Embedded Memory Networks. In Proceedings of the 26 International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2016–2022.
24. Le, V.M.; Le, V.; Venkatesh, S.; Tran, T. Hierarchical Conditional Relation Networks for Video Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9972–9981.
25. Yu, Z.; Yu, J.; Fan, J.; Tao, D. Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 24–27 October 2017; pp. 1821–1830.
26. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef] [PubMed]

27. Song, L.; Smola, A.; Gretton, A.; Borgwardt, K.; Bedo, J. Supervised Feature Selection via Dependence Estimation. In Proceedings of the 24th Annual International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 823–830.
28. Chen, D.; Dolan, W.B. Collecting Highly Parallel Data for Paraphrase Evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 190–200.
29. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5288–5296.