symmetry

MDPI

Article

Temporally Multi-Modal Semantic Reasoning with Spatial Language Constraints for Video Question Answering

Mingyang Liu, Ruomei Wang, Fan Zhou * and Ge Lin

National Engineering Research Center of Digital Life, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China; liumy77@mail2.sysu.edu.cn (M.L.); isswrm@mail.sysu.edu.cn (R.W.); linge3@mail.sysu.edu.cn (G.L.)
* Correspondence: isszf@mail.sysu.edu.cn

Abstract: Video question answering (QA) aims to understand the video scene and underlying plot by answering video questions. An algorithm that can competently cope with this task needs to be able to: (1) collect multi-modal information scattered in the video frame sequence while extracting, interpreting, and utilizing the potential semantic clues provided by each piece of modal information in the video, (2) integrate the multi-modal context of the above semantic clues and understand the cause and effect of the story as it evolves, and (3) identify and integrate those temporally adjacent or non-adjacent effective semantic clues implied in the above context information to provide reasonable and sufficient visual semantic information for the final question reasoning. In response to the above requirements, a novel temporally multi-modal semantic reasoning with spatial language constraints video QA solution is reported in this paper, which includes a significant feature extraction module used to extract multi-modal features according to a significant sampling strategy, a spatial language constraints module used to recognize and reason spatial dimensions in video frames under the guidance of questions, and a temporal language interaction module used to locate the temporal dimension semantic clues of the appearance features and motion features sequence. Specifically, for a question, the result processed by the spatial language constraints module is to obtain visual clues related to the question from a single image and filter out unwanted spatial information. Further, the temporal language interaction module symmetrically integrates visual clues of the appearance information and motion information scattered throughout the temporal dimensions, obtains the temporally adjacent or non-adjacent effective semantic clue, and filters out irrelevant or detrimental context information. The proposed video QA solution is validated on several video QA benchmarks. Comprehensive ablation experiments have confirmed that modeling the significant video information can improve QA ability. The spatial language constraints module and temporal language interaction module can better collect and summarize visual semantic clues.

Keywords: video question answering; visual language interaction; multi-modal semantic reasoning

Citation: Liu, M.; Wang, R.; Zhou, F.; Lin, G. Temporally Multi-Modal Semantic Reasoning with Spatial Language Constraints for Video Question Answering. *Symmetry* **2022**, *14*, 1133. https://doi.org/10.3390/sym14061133

Academic Editor: Changxin Gao

Received: 27 April 2022
Accepted: 27 May 2022
Published: 31 May 2022

1. Introduction

Video question answering (QA) involves theoretical approaches to computer vision and natural language processing, in which it is not only necessary to have the same ability as image retrieval or image caption [1,2] to understand the visual semantic information provided by a single image, but it is also necessary to have the ability of video caption or video moment retrieval [3,4] to understand the potential semantics in the rich visual, text, and audio clues scattered throughout the video. Most importantly, video QA also needs to calculate the above potential semantics to achieve a near-human QA capability [5] and find the most appropriate answers to the natural language questions in the video. Such intriguing video QA tasks have become an important research issue in artificial intelligence (AI) [6–8].

Compared with image QA, video QA is undoubtedly more challenging [9]. The input information is changed from a single image to a sequence with continuous images, and an entire storyline scatters across these frames, which requires an algorithm to collect multi-modal information scattered in the video frame sequence while extracting, interpreting, and utilizing the potential semantic clues provided by each modal information in the video, such as the text content provided by natural language questions, the appearance information, the spatial location information provided by the spatial dimension information of video frames, and the dynamic evolution of this visual information with the story's development. In order to achieve high-quality video QA capabilities, QA models need to perform further computations to integrate the multi-modal context of the above semantic clues and understand the cause and effect of the story as it evolves. For a given question, the QA model also needs to identify and integrate those temporally adjacent or non-adjacent effective semantic clues implied in the above context information and filter out irrelevant or harmful context information to provide reasonable and sufficient visual semantic information for the final question reasoning. The above requirements make deriving the most accurate and appropriate answer to a given question require more than image QA's next-generation AI capabilities.

In order to solve the challenging video QA tasks discussed above, this paper conducts research on feature extraction, spatial semantic cue reasoning, and temporal language interaction. The primary purpose is to locate the multi-modal semantic clues related to the question from the whole video efficiently to QA. Firstly, the previous video question answering algorithm mainly obtains multi-modal information of the video through uniform sampling [6,10]. However, there is no solid temporal regularity in the story's progression, and the multi-modal information may undergo rapid or slow semantic changes over a period of time. As a result, the visual semantic clues are likely to be non-uniformly distributed in the temporal space, making the uniform sampling liable to miss some critical visual semantic clues. Furthermore, increasing the sampling rate may lead to the redundancy of semantic clues due to repeated sampling. Secondly, the video consists of continuous frames and an entire storyline scattered in these frame sequences. Therefore, the video QA model is not only required to locate the semantic clues of the spatial dimension of a single image, but it is also necessary to locate the semantic clues of the temporal dimension of the whole video. Some existing methods focus on modeling the temporal context information in the video but insufficiently exploit the feature of the spatial dimension [7]. Other methods try to apply spatio-temporal attention to videos [10]. However, they show even worse performance than temporal-only attention, possibly due to the lack of spatial guidance for temporal dimensions.

As for these problems, a novel temporally multi-modal semantic reasoning with spatial language constraints video QA solution is reported in this paper, which includes a **significant feature extraction module** used to extract multi-modal features according to the characteristic that the spatial difference of pixels between video frames can reflect the development and change of temporal information, a **spatial language constraints module** (SLC) used to recognize and reason spatial dimensions in video frames under the guidance of questions, and a **temporal language interaction module** (TLI) used to locate the temporal dimension semantic clues of the appearance features and motion features sequence.

Figure 1 shows the overall architecture of our video QA solution. Features extracted from the significant feature extraction module are used to establish condensed semantic clues of video appearance and motion features for the answer decoder through the multi-modal data spatio-temporal reasoning (MSTR) network composed of the spatial language constraint module and the temporal language interaction module.

In detail, for a given question, the results processed by the spatial language constraints module of appearance features are to obtain visual clues related to the question and filter out unwanted spatial information. Further, the temporal language interaction module symmetrically integrates visual clues of the appearance information and motion infor-

mation scattered throughout the temporal dimensions, obtains the temporally adjacent or non-adjacent effective semantic clue, and filters out irrelevant or detrimental context information to obtain video appearance condensed features and video motion condensed features. Finally, we use different answer decoders to answer video QA questions based on different types (open-ended, multi-choice, and repeat count tasks).

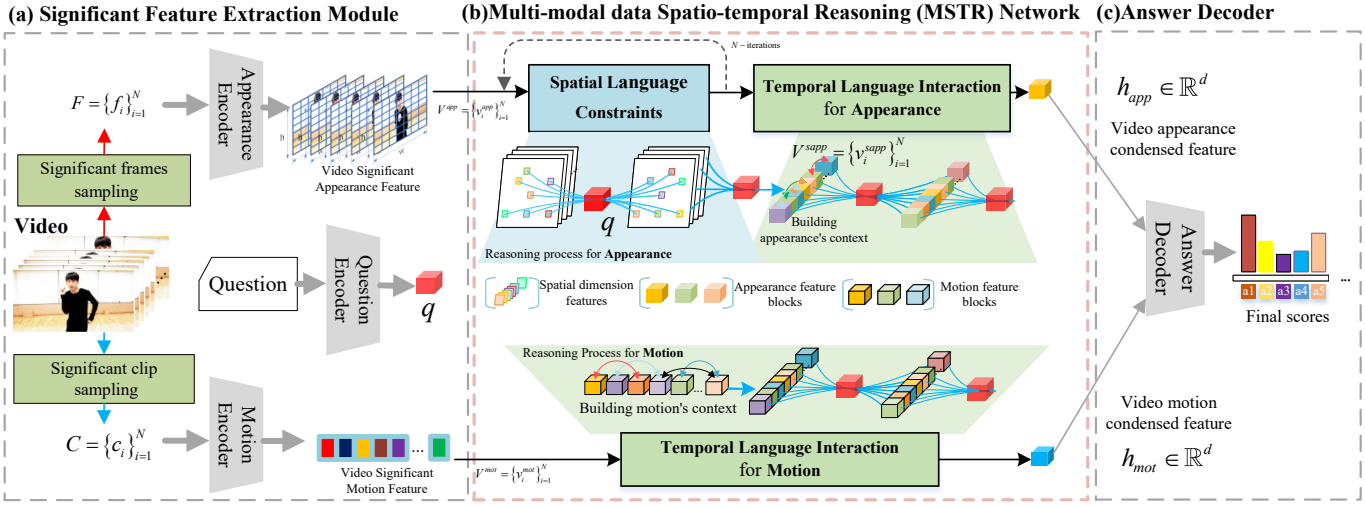

Figure 1. Overview architecture of our video QA solution. It consists of three modules: (**a**) significant feature extraction module, (**b**) multi-modal data spatio-temporal reasoning (MSTR) network with spatial language constraint module (SLC) and temporal language interaction module (TLI), and (**c**) answer decoder module. Specifically, the significant feature extraction module can take raw videos as input and accurately obtain the video's significant information through a series of deep networks. The MSTR networks take multi-modal visual features V^{app} and V^{mot} as inputs. It acquires visual semantic clues provided by different video frames and video clips in a bottom-up manner, providing more effective spatio-temporal visual semantic clues for the further understanding and reasoning of questions. Finally, our video QA solution uses different answer decoders to answer questions based on the different types of video QA.

The proposed video QA solution was validated on three vital video QA benchmarks: TGIF-QA [11], MSVD-QA [12], and MSRVTT-QA [13]. The experimental results show that our method outperforms state-of-the-art video QA algorithms across a range of question types. Moreover, comprehensive ablation experiments and visual results have also confirmed the necessity of video QA modeling based on significant information from video and the rationality of multi-modal data spatio-temporal reasoning networks.

The main contributions of this paper are as follows:

(1) A practical and straightforward significant feature extraction module is proposed, which can capture the significant visual semantic clues that change with the story progression and reduce the redundant expression of the video.

(2) A novel multi-modal data spatio-temporal reasoning network is proposed, which can establish the cross-modal interaction between vision and question along with the story of the video and provide the most accurate and reasonable visual semantic information for video QA tasks. Moreover, spatial reasoning can guide the temporal language interaction to focus on the appearance features sequence associated with the question, rather than the whole frame sequence.

(3) A novel video QA solution is reported, building upon the aforementioned two new algorithmic modules, which can collect the visual semantic clues provided by significant video frames and significant video clips in a bottom-up manner. This solution outperforms all state-of-the-art video QA algorithms across a range of question types,

as consistently demonstrated through our comprehensive comparison experiments executed on publicly available benchmark datasets.

2. Related Works

To achieve satisfactory video QA capabilities, QA models need to capture the rich visual semantic clues scattered throughout the video. Furthermore, for a given question, the QA model also needs to identify and integrate those temporally adjacent or non-adjacent effective semantic clues implied in the above context information and filter out irrelevant or even harmful context information. Visual semantic extraction and visual language interaction have become two key challenges of video QA at the present stage. We present a brief review of the related work on the two challenges.

2.1. Visual Semantic Extraction

Visual semantic extraction aims to obtain the rich visual semantics contained in the video and provide reasonable visual semantic clues for the following video QA reasoning. Video QA researchers extract visual semantic features from original videos through a series of deep networks [14,15], aiming at digesting various visual semantic information provided by videos.

Zhao et al. [16] used VGGnet [17] to extract the appearance features of video frames as the visual semantic expression. Yu et al. [18] first equidistantly sampled one per ten frames from a video and extracted the appearance features of the above frames from the res5c layer of ResNet [14]. Then, they used the pooling and convolution layers to reduce visual semantic clues requiring reasoning. In order to obtain visual semantic clues with more visual spatial semantic dimensions, Jiang et al. [10] uniformly sampled 10 frames of images from the video as video expressions firstly. Then, they used the res4c layer of ResNet to extract the corresponding appearance features.

In order to enable the video question answering model to acquire spatio-temporal visual semantic clues, Jang et al. [11] proposed a dual-stream architecture using appearance and motion features, which uniformly sampled video frames and video clips to reduce video redundancy, and used ResNet-152 [14] and C3D [15] networks to obtain visual semantic clues. Xu et al. [12] uniformly sampled 20 video frames and clips independently first. Then, they used VGG and C3D networks to obtain the videos' appearance and motion features. Le et al. [7] evenly divided the video into eight video clips, and then selected 16 consecutive frames from each video clip as the image sequence of each video. Finally, ResNet and ResNexT [19,20] were used to extract the visual semantics of the video.

The above works promote the development of visual semantic expression and provide effective visual semantic clues for video QA. However, the visual semantic expression provided by the video is non-uniformly distributed in the temporal space. As a result, the uniform sampling can miss some critical visual semantic clues, while increasing the sampling rate also leads to the redundancy of semantic clues due to repeated sampling.

In order to capture the complete visual semantic clues scattered throughout the video, Zhu et al. [21] proposed a multirate visual recurrent model based on a multirate gated recurrent network. These recurrent models [22] can capture significant visual information in the video but have difficulties in training. Feichtenhofer et al. [23] obtained visual semantic expressions of different distributions provided by videos through double-channel timing sampling. In the video QA, Wang et al. [24] presented a question-related video content localization module to learn the video's significant visual semantic distribution. Kim et al. [25] proposed a temporal localization module to learn the distribution of significant features of videos. However, this kind of training relies on the labeling of the dataset [26] and is challenging to promote.

In order to cope with this challenge, a significant feature extraction module is proposed in this paper. This module extracts multi-modal features according to the characteristic that the spatial difference of pixels between video frames can reflect the development and change of temporal information, and the video content can be accurately sampled. As

a result, the video QA model can obtain the complete visual semantic clues scattered in the video.

2.2. Visual Language Interaction

Visual language interaction is the task of projecting multi-modal information into a common space for interactive retrieval [27]. Several attention models have shown good visual language interaction ability in temporal visual language interaction [12,28], spatial visual language interaction [29,30], and spatio-temporal visual language interaction [11,31].

The spatial attention model is designed to focus on significant spatial dimensions rather than the whole image to obtain more valid semantic clues [32]. A large number of variations of spatial attention models have been proposed, including fine-grained attention and co-attention [33,34]. To obtain spatial semantic cues, Wan et al. [5] proposed a hierarchical fusion network capable of the nonlinear transformation and iterative fusion of input information, which effectively perceives the spatial semantics related to the question and provides effective visual clues for the final question prediction. These attention-based models significantly improve the performance of QA reasoning.

Compared with the spatial attention model, the temporal attention model transforms the processing objects into the visual semantic appearance of the video frame sequence. Moreover, it pays more attention to the temporal modeling of context between video frames, aiming to find the sequence of visual features relevant to the question [26,35]. Li et al. [36,37] proposed a co-attention attention model between learning video appearance sequences and questions and enhanced co-attention by using the self-attention [38] mechanism. To obtain fine-grained visual semantic language interaction information, some researchers have also proposed object temporal attention based on object–language interaction [8].

In order to be able to unify visual semantic clues between the appearance, motion, and the question, Jang et al. [11] first combined the spatial attention model with the temporal attention model and proposed a spatio-temporal attention model for the video QA model. Due to its superior performance, attention models such as multi-modal fusion memory [39], co-memory attention [31], and multi-modal attention [40] were soon derived Moreover, hierarchical attention [41,42] and multi-step progressive attention [43,44] are proposed to obtain a broader visual language interaction capability. Researchers also rely on the structure of the video to achieve multi-level visual language interactions from video frames to video clips [7,45]. The question-aware and reasoning features penetrate the visual language interaction of each video level, thus achieving good visual semantic interaction performance, which suggests that building the semantic interaction relying on the structure of the video is essential for improving the performance of video QA. Our work is related to the above visual semantic interaction but provides a different perspective.

In our visual language interaction framework, the low-level visual language interaction no longer models the temporal expression between adjacent frames and integrates essential representations in the spatial dimensions of the video. It can realize the constraint of the spatial semantic dimension under the guidance of the question and make the video QA algorithm focus on the appearance features related to the question. At the higher level of visual language interaction modeling, the video QA model can obtain the guidance of the question and enable the temporal language interaction modeling to focus on the spatial dimension related to the question with the constraint of the semantic space dimension. As a result, the model can better capture the visual semantics provided by the video and obtain complete spatio-temporal visual language interaction information.

3. Materials and Methods

This section introduces the problem formulation of video QA in Section 3.1, followed by the significant feature extraction module in Section 3.2, the multi-modal data spatio-temporal reasoning network in Section 3.3, and the answer decoder in Section 3.4.

3.1. Problem Formulation

The goal of video QA is to deduce the answer a^* to question q from a video $\mathcal{V}$. This process can be defined as follows:

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \mathscr{F}(a|q, V), \tag{1}$$

where V represents multi-modal features in the video, such as the appearance feature [14] and motion feature [20] of the video; q is a sequence of text composed of natural language and represents a query to the content contained in the video; a is part of the candidate answer space $\mathcal{A}$, and $\mathscr{F}$ is the mapping function that maps a pair of videos and questions to an answer. For the open-ended question [12], the answer space contains all possible alternative answers, and for the multiple-choice question [11], this space contains a set of candidate answers.

3.2. Significant Feature Extraction Module

Multi-modal information (motion and appearance as the description of the video) is not evenly distributed in the video frame sequence. Different from multi-modal video information by uniform sampling, in this paper, we propose a significant feature extraction module to accurately sample the significant frame/clips of the video by the characteristic that the spatial difference of pixels between video frames can reflect the development and change of temporal information. Moreover, a series of deep networks are used to capture significant features in the above frames and clips.

Formally, we consider a video $\mathcal{V}$ of length L, which contains a set of frames. At first, we use the frame difference method to obtain the sequence $\{D_i\}_{i=1}^{L-1}$ of the pixel spatial difference between adjacent frames. Moreover, the following formula is used to weaken the difference between frames and obtain the sequence $\{\hat{D}_i\}_{i=1}^{L-1}$ of spatial differences of pixels with context information:

$$\hat{D}_i = \sum_{j=i-s}^{i+s} \alpha_j \times D_j, \tag{2}$$

where $\alpha_j = \sin \frac{j-(i-s)}{2s} \pi$ is the weight parameter, and s is the sliding window length. If the sliding window is outside of the video sequence, we fill it with the D_1 or D_{L-1}. Then, according to the above distribution, more sampling points are arranged in the region with large spatial pixel variation and N sampling points are used to obtain the **significant frames** $F = \{f_i\}_{i=1}^{N}$.

Finally, we use the res5c layer (i.e., $\mathbb{R}^{7\times7\times2048}$) of ResNet [14], and apply an average pooling with a linear projection matrices $W_{app} \in \mathbb{R}^{2048\times d}$ to extract the appearance feature $V^{app} = \left\{ v_i^{app} \middle| v_i^{app} \in \mathbb{R}^{w*h*d} \right\}_{i=1}^{N}$ provided by the frame corresponding to each sampling point, where w, h, and d are the height, width, and feature dimension of V^{app} feature map.

Moreover, for each sampling point f_i, we take adjacent frames $t_i - T$ to $t_i + T$ to capture the context information near the sampling point, where t_i is the position of the sampling point in the video, and we can obtain the **significant clips** $C = \{c_i\}_{i=1}^{N}$. Finally, the ResNeXt-101 [19,20] with linear projection matrices $W_{mot} \in \mathbb{R}^{2048\times d}$ is used to extract the motion feature $V^{mot} = \left\{ v_i^{mot} \middle| v_i^{mot} \in \mathbb{R}^{d} \right\}_{i=1}^{N}$.

For **linguistic representation**, we first embed all words in the question and all words in candidate answers in the case of multi-choice questions into vectors of 300 dimensions with pre-trained GloVe Word Embeddings [46]. Then, we use biLSTM to obtain the context information. Finally, we concatenate the output hidden states of the forward and backward LSTM passes as a linguistic representation. Through the above process, we can obtain the semantic information $q \in \mathbb{R}^d$ of the question and the semantic information $\{a_i \in \mathbb{R}^d\}_{i=1}^{N_{ca}}$

of the candidate answers, where N_{ca} represents the number of candidate answers for multi-choice tasks.

3.3. Multi-Modal Data Spatio-Temporal Reasoning Network

A video QA solution needs to understand the above multi-modal information. It also needs the ability of visual language interaction to provide effective visual semantic clues for the final question reasoning. In order to satisfy the above requirements, we propose a multi-modal data spatio-temporal reasoning network composed of a spatial language constraint (SLC) module and a temporal language interaction (TLI) module. With the above module, the visual language interaction between vision and question can be established, along with the story of the video, and provide the most accurate and reasonable visual semantic information for video QA tasks. Specifically, we use multi-modal visual features V^{app} and V^{mot} as the visual information input of the MSTR network. Under the guidance of question q, we use TLI and SLC modules to deduce and integrate effective semantic cues existing in video appearance and motion features.

The **spatial language constraint module (SLC)** aims to obtain the spatial visual clues related to the question. Figure 2 shows the module structure in detail. The question q is used to perceive the spatial dimensions related to the question in the appearance feature v_i^{app} of a significant frame and obtain the question label $Mask_i$, which marks the correlation between each spatial dimension and the question. This process can be defined as follows:

$$\begin{aligned} Mask_i &= g(v_i^{app}, q) \\ &= \text{softmax}(W_2(\text{ELU}(W_1[v_i^{app}; q]))), \end{aligned} \tag{3}$$

$$\hat{v}_i^{app} = Mask_i \odot v_i^{app}, \tag{4}$$

where $W_1 \in \mathbb{R}^{2d \times d}$ and $W_2 \in \mathbb{R}^{d \times 1}$ are different linear projection matrices (in this paper, W always is a linear projection matrix), ELU is the Exponential Linear Unit [47], $[.;.]$ represents the tensor concatenation, and $\odot$ represents the element-wise multiplication. Through the above formula, we can preliminarily obtain spatial features $\hat{v}_i^{app} \in \mathbb{R}^{w \times h \times d}$ related to the question.

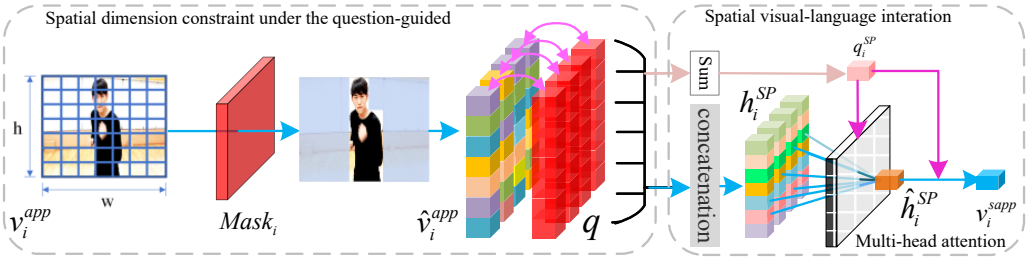

Figure 2. An illustration of spatial language constraint module. This module allows a frame to interact with question q in spatial dimensions, which obtains semantic clues provided by the spatial dimensions related to the question and summarizes them toward a higher-level condensed representation with the guidance of the question.

Then, the following formula $SP(.)$ is used further to obtain the semantic interaction feature h_i^{SP} between the spatial appearance feature $\hat{v}_i^{app}$ and the question q, which is the spatial semantic clue hidden state of the significant frame.

$$\begin{aligned} h_i^{SP} \in \mathbb{R}^{w \times h \times d} &= SP(\hat{v}_i^{app}, q) \\ &= \text{ELU}(W_3[\hat{v}_i^{app}; q]), \end{aligned} \tag{5}$$

where $W_3 \in \mathbb{R}^{2d \times d}$. At the same time, we notice that the question can perceive the video content related to the question, and video content can also complete the supplementary

understanding of the question. Therefore, we sum the spatial semantic information h_i^{SP} to obtain the video question $q_i^{SP} \in \mathbb{R}^d$ after the spatial semantic supplement.

Then, in order to further use question q_i^{SP} with the spatial semantic supplement to analyze and summarize the hidden states of spatial semantic clues, we use a multi-head attention model [38] to explore the semantic relationship between visual semantics h_i^{SP} and question q_i^{SP}. This model is shown in the following formula:

$$\begin{aligned} \hat{h}_i^{SP} &= \text{MultiHead}(q_i^{SP}, h_i^{SP}, h_i^{SP}) \\ &= W^O(head_1, head_2, ..., head_H), \end{aligned} \tag{6}$$

$$head_j = \text{softmax}(\frac{W_j^Q q_i^{SP} \times W_j^K h_i^{SP}}{\sqrt{d/H}}) W_j^V h_i^{SP}, \tag{7}$$

where $W_j^Q \in \mathbb{R}^{d \times \frac{d}{H}}, W_j^K \in \mathbb{R}^{d \times \frac{d}{H}}, W_j^V \in \mathbb{R}^{d \times \frac{d}{H}}$ and $W^O \in \mathbb{R}^{d \times d}$ are different linear projection matrices, $j \in 1, 2, ..., H$. Through this formula, we can realize the question's analysis of spatial information and obtain the condensed spatial semantic expression $\hat{h}_i^{SP} \in \mathbb{R}^d$ that the question q_i^{SP} pays attention to h_i^{SP} from different linear spaces.

Finally, we fuse the spatial semantic supplement question q_i^{SP} with the condensed spatial semantic expression $\hat{h}_i^{SP}$ as the spatial semantic clue v_i^{sapp} of the frame f_i. The formula is as follows:

$$v_i^{sapp} \in \mathbb{R}^d = \text{ELU}(W_4[\hat{h}_i^{SP}; q_i^{SP}]), \tag{8}$$

where $W_4 \in \mathbb{R}^{2d \times d}$. With the above process, we can complete the spatial semantic interaction between appearance features V^{app} of significant frames and question q, and obtain a higher level of appearance condensed representation $V^{sapp} = \left\{ v_i^{sapp} \middle| v_i^{sapp} \in \mathbb{R}^d \right\}_{i=1}^N$.

In the **temporal language interaction module (TLI)**, we symmetrically conduct temporal language interaction modeling for the higher level of appearance features $V^{sapp} = \left\{ v_i^{sapp} \middle| v_i^{sapp} \in \mathbb{R}^d \right\}_{i=1}^N$ and motion features $V^{mot} = \left\{ v_i^{mot} \middle| v_i^{mot} \in \mathbb{R}^d \right\}_{i=1}^N$, respectively, and obtain condensed visual clues related to the question. Figure 3 shows this module's details. It takes the question q and the video feature V^S as input, and outputs the higher-level visual semantic feature h_S . When V^S is defined as V^{sapp} and V^{mot}, respectively, the condensed visual clues are h_{app} and h_{mot}.

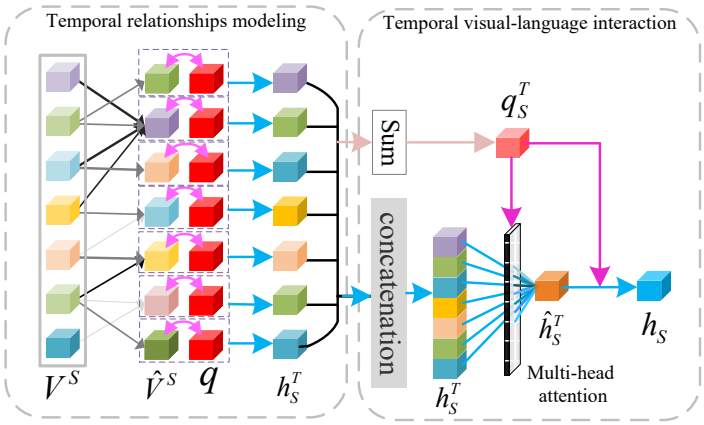

Figure 3. An illustration of the temporal language interaction module. This module allows a video feature sequence to interact with question q in temporal dimensions, which can recognize, infer, and integrate multi-modal context information under the question's guidance and provide the most accurate and reasonable visual semantic information for QA reasoning.

Firstly, we take V^S as the input of a multi-head attention model to explore the temporal correlation of V^S itself. Ignoring the multi-head process of Equation (7), the final output is given by:

$$\hat{V}^S \in \mathbb{R}^{N\times d} = \text{MultiHead}(V^S, V^S, V^S). \tag{9}$$

Then, in order to analyze the hidden temporal semantic clues h_S^T existing in the context, the following formula is used to obtain the semantic interaction feature of $\hat{V}^S = \{\hat{v}_i^S\}_{i=1}^N$ and question feature q:

$$\begin{cases} h_{S,i}^T = \text{ELU}(W^i[\hat{v}_i^S; q]) \\ h_S^T = [h_{S,1}^T; ...; h_{S,N}^T] \end{cases}, \tag{10}$$

where $\{W^i \in \mathbb{R}^{2d\times d}\}_{i=1}^N$ are different linear projection matrices. Identical to the SLC module, we sum the temporal semantic information h_S^T to obtain the video question $q_S^T \in \mathbb{R}^d$ supplemented by the temporal semantic information.

Moreover, the question q_S^T and hidden temporal semantic information h_S^T are taken as the input of multi-head attention, which serves to realize the question analysis of the temporal information and obtain the condensed temporal semantic expression $\hat{h}_S^T$ related to the question q_S^T. Ignoring the multi-head process of Equation (7), the process is given by:

$$\hat{h}_S^T \in \mathbb{R}^d = \text{MultiHead}(q_S^T, h_S^T, h_S^T). \tag{11}$$

Finally, we fuse the temporal semantic supplement question q_S^T with the condensed temporal semantic expression $\hat{h}_S^T$ as the semantic clue h_S of the video. The formula is as follows:

$$h_S \in \mathbb{R}^d = \text{ELU}(W_5[\hat{h}_S^T; q_S^T]), \tag{12}$$

where $W_5 \in \mathbb{R}^{2d\times d}$. With the above process, we can symmetrically realize the temporal visual language reasoning for the appearance feature and motion features and obtain higher-level visual semantic clues $h_{app} \in \mathbb{R}^d$ and $h_{mot} \in \mathbb{R}^d$.

3.4. Answer Decoder

To deal with video questions that need to understand both appearance and motion information, the video QA model needs to have the ability to integrate this multi-modal information. In this paper, different answer decoders with loss functions are adopted for the final QA reasoning to fulfil the answer prediction of different question types.

Open-ended questions. We treat the open-ended questions as multi-label classification problems. Firstly, we take the features h_{app}, h_{mot} and the question q as input, and use the following formula to decode the open-ended questions:

$$\delta'_{open} \in \mathbb{R}^{N_o} = W_{open}(\text{ELU}(W_o[h_{app}; h_{mot}; W_q q])), \tag{13}$$

where $W_o \in \mathbb{R}^{3d\times d}$, $W_{open} \in \mathbb{R}^{d\times N_o}$, $W_q \in \mathbb{R}^{d\times d}$ are linear projection matrices, and N_o represents the length of the answer space $|\mathcal{A}|$. Then, we obtain the final score $\delta_{open} \in \mathbb{R}^{N_o}$ of each candidate answer:

$$\delta_{open} \in \mathbb{R}^{N_o} = soft\max(\delta'_{open}). \tag{14}$$

Finally, the highest score is selected as the prediction answer:

$$a^* = \arg\max_{a^o}(\delta_{open}). \tag{15}$$

The cross-entropy loss function is used for the open-ended questions.

Repetition count task. We obtain the number $\delta_{count} \in \mathbb{R}^1$ of repetitions of the action by the following formula:

$$\delta'_{count} \in \mathbb{R}^1 = W_{count}(\text{ELU}(W_c[h_{app}; h_{mot}; W_q q])), \tag{16}$$

where $W_c \in \mathbb{R}^{3d \times d}$, $W_{count} \in \mathbb{R}^{d \times 1}$ are linear projection matrices. Moreover, we use a rounding function for integer count results. Finally, mean squared error (MSE) is used as the loss function.

Multi-choice question. The multi-choice question is treated as a multi-label classification problem. Firstly, we take the features h_{app}, h_{mot}, q and $\{a_i\}_{i=1}^{N_{ca}}$ as input, and the following formula is used to decode the multi-choice question:

$$\delta_i = W_{multi}(\text{ELU}(W_m[h_{app}, h_{mot}, W_q q, W_a a_i]), \tag{17}$$

where $W_m \in \mathbb{R}^{4d \times d}$, $W_{multi} \in \mathbb{R}^{d \times 1}$, and $W_a \in \mathbb{R}^{d \times d}$ are linear projection matrices. Finally, the candidate with the largest δ value is selected as the answer such that:

$$a^* = \arg\max_i(\delta_i). \tag{18}$$

In this type of video QA, the cross-entropy loss function is used for training.

In order to ensure that the multi-modal information of our video QA solution is not lost, we use multi-task loss to train our network:

$$loss = 0.8 \times loss_{all} + 0.1 \times loss_{app} + 0.1 \times loss_{mot}, \tag{19}$$

where $loss_{app}$ and $loss_{mot}$ represent the loss value of predicted results in the appearance and motion branches of the MSTR network, respectively. In addition, $loss_{all}$ represents the loss value of the last predicted results of our video QA solution.

4. Experimental Results Analysis

In this section, we evaluate the performance of our video QA solution proposed in this paper through several experiments. In order to verify the effectiveness of our video QA solution and its components, we compared our solution with the state-of-the-art methods, conducted a large number of ablation experiments, and further analyzed the impact of our solution on performance through a large number of qualitative analyses.

4.1. Dataset Details

In order to evaluate our methods objectively and fairly, we selected three widely used and challenging video QA datasets for subsequent experiments: TGIF-QA [11], MSVD-QA [12], and MSRVTT-QA [13]. The detailed statistics on the number of QA pairs and possible answers for each question in the three datasets are shown in Table 1.

Table 1. Statistics of TGIF-QA [11], MSVD-QA [12], and MSRVTT-QA [13]. "Possible answers" denotes the number of possible answers to each question.

QA Pairs.	TGIF-QA				MSVD	MSRVTT
	Action	Trans	Frame	Count		
Train	18,427	47,433	35,452	24,158	30,933	158,581
Val	2048	5271	3940	2685	6415	12,278
Test	2274	6232	13,691	3554	13,157	72,821
Total	22,749	58,936	53,083	30,397	50,505	243,680
Possible answers	5	5	1541	11	1852	4000

TGIF-QA: TGIF-QA uses TGIF [48] as the data source, contains 165K QA pairs, and is divided into four subtasks according to the unique attributes of the question. (1) Repeating action is a multi-choice question with five candidate answers, which requires the algorithm to determine the actions by spatio-temporal reasoning according to the given number of actions. For example, "What does the woman do 10 or more than 10 times?". (2) State transition is also a multi-choice question with five candidate answers. It requires the algorithm to understand the spatio-temporal information and determine the change of state transition. For example, "What does the man do after putting hand on hips?". (3) Repeating counting is an open-ended question requiring an algorithm to determine the number of action occurrences with spatio-temporal information. Specifically, each count question has 11 possible answers, ranging from zero to ten. For example, "How many times does the man raise his left eyebrow?". (4) Frame QA is similar to image QA and is an open-ended question, which requires an algorithm to find the spatial visual clues relevant to the question among all videos and use spatial reasoning ability to answer the question. As an open-ended question, each question has 1541 possible answers. For example, "What is the man talking on a headset using?".

MSVD-QA: In this dataset, 50K video QA pairs are marked in 1970 video clips. Each video clip is approximately 10 s. As an open-ended question, each question has 1852 possible answers.

MSRVTT-QA: This dataset marks 243K question pairs in 10K video clips. Compared with the first two video question answering datasets, this dataset has a 10–30 s video sequence, making the scene in the video more complex and information more redundant, posing a higher challenge to the video QA algorithms.

4.2. Implementation Details

(1) **Experimental Settings:** For the significant feature extraction, the length of the sliding window is set to $s = 16$ and we sample $N = 8$ significant frames from the video to represent the appearance information of the video and obtain the $T = 8$ adjacent frames to capture the motion information near the significant frames. The pre-trained ResNet [14] and ResNext-101 [19,20] are used to obtain the appearance feature with $w = 4$, $h = 4$ and motion feature with $w = 1$, $h = 1$, respectively. For parameter settings, the number of heads in the multi-head attention network is set to 8, and we set the feature dimension d to 512. For a fair comparison, the original Adam optimizer [49] is employed to optimize the model, which is widely used in video QA. In the train processing, the batch size is set to 32, and the learning rate is 1e-4 with a staircase learning rate schedule, where we multiply the learning rate by 0.5 every 5 epochs. In addition, we employ the dropout rate of 0.1 to prevent over-fitting, and all experiments are terminated after 25 epochs.

(2) **Evaluation Metrics:** For the count task in the TGIF-QA dataset, mean squared error (MSE) loss is used to evaluate the difference between the ground truth and the predicted answer. Compared with the accuracy, MSE can objectively and accurately reflect the difference between the predicted and current values. Therefore, the smaller the difference (MSE value), the better the performance. For other tasks of video QA datasets, accuracy is employed to evaluate the performance of the models: the better the performance, the higher the accuracy.

4.3. Ablation Studies

In order to verify our contribution, extensive ablation experiments are performed on our proposed solution to verify the effectiveness of spatial dimensions, the feasibility of accurate sampling, the feasibility of significant feature extraction, and the effectiveness of our MSTR networks.

The effectiveness of spatial dimensions. In this paper, our video QA solution divides visual language interaction into appearance features with the spatial dimension. Intuitively, our video QA solution can perceive the location information of the object in the video and

obtain fine-grained visual semantic clues related to the question. In order to verify the effect of spatial attributes on video QA performance, we compare the influences of different spatial dimensions $(w, h) = \{(1,1), (2,2), (4,4), (7,7)\}$ on performance in the Frame QA task of the TGIF-QA datasets and MSVD-QA datasets, and we leave the rest of the model unchanged. As seen in Table 2, with the increase in the number of spatial dimensions, the performance of video QA is improved, which verifies the contribution of spatial attributes to video QA. In addition, the improvement in Frame QA is more significant than MSVD-QA, which verifies that spatial attributes are more helpful for us in answering questions about spatial reasoning.

Table 2. Verifying the effectiveness of spatial dimensions. We compare the influences of different spatial dimensions $(w, h) = \{(1,1), (2,2), (4,4), (7,7)\}$ on performance.

$(w \times h)$	1×1	2×2	4×4	7×7
Frame QA	58.5	58.9	60.1	60.4
MSVD-QA	39.0	39.1	39.9	40.2

The feasibility of accurate sampling. Our video QA solution believes that accurate sampling can effectively reduce redundant visual information, which helps us to focus on modeling the significant content of the video better. In order to further verify the impact of sampling frequency on performance, we used the different sampling frequencies $N = \{1, 2, 4, 8, 16\}$ to obtain different visual expressions. As seen from Table 3, with the increase in the number of samples, the performance of video QA has not quickly improved. In particular, the sample numbers N = 8 and N = 16 have very similar performance for the video QA test. However, the model parameters increased by 4M due to the excess sampling, while the data size increased by two times. The above results verify the feasibility of significant feature extraction and the redundancy of the video content itself.

Table 3. Verifying the feasibility of significant feature extraction. We compare the influences of different sampling frequencies $N = \{1, 2, 4, 8, 16\}$ on performance. The data size is for the MSRVTT-QA datasets.

Sample Num.	Frame QA	MSVD-QA	MSRVTT-QA	Parameter	Data Size
1	57.1	36.0	33.3	24 M	1.4 GB
2	58.7	38.1	35.2	25 M	2.8 GB
4	59.9	39.3	36.2	26 M	5.6 GB
8	60.1	39.9	36.7	28 M	11.1 GB
16	60.1	40.1	36.8	32 M	22.2 GB

The effectiveness of MSTR networks. This paper designs a multi-modal data spatio-temporal reasoning network composed of the spatial language constraint (SLC) module and temporal language interaction (TLI) module. To better understand the contribution of this network, we performed ablation experiments on network inputs and the module composition of the MSTR network, shown in Table 4. We evaluate the following settings: ▲ w/o appearance—removing the branch of the appearance feature, we only use the branch of the motion feature to solve the video QA problem. ▲ w/o motion—the branch of the motion feature is removed and we only use the branch of the appearance feature for the video QA problem. ▲ w/o SLC—we remove the spatial language constraint module and use video features with no spatial dimension information input to the MSTR network. ▲ w/o S-V2Q—we remove the supplementary understanding of the question by spatial semantics. ▲ w/o TLI—we remove the temporal language interaction module and use sum operations to summarize the temporal information. ▲ w/o T-V2Q—we remove the supplementary understanding of the question by temporal semantics. As seen in Table 4, all of the modules presented in this paper are important, and removing any of them will degrade the corresponding performance. It is worth noting that when we only use the SLC

module, our algorithm in the Frame QA task also achieves excellent performance, which also verifies the superiority of the spatial visual language interaction module.

The necessity of significant feature extraction. This paper accurately sampled the significant video information by the characteristic that the spatial difference of pixels between video frames can reflect the development and change of temporal information. Moreover, it built the video significant visual features, which effectively reduced the cost of data storage and network model. In order to prove the necessity of this visual feature extraction method, we compare the performance of HCRN [7] and our video QA solution in original data (extracted by the HCRN), our features without a spatial dimension, and our features in Table 5.

Table 4. Verifying the effectiveness of our MSTR networks. We performed ablation experiments on network inputs and module composition of MSTR networks. The lower, the better for the count.

Model	TGIF-QA		MSVD-QA
	Frame	Count	
Input conditioning			
w/o appearance	51.5	3.94	35.1
w/o motion	**59.2**	4.21	37.8
Components of MSTR			
w/o SLC	58.0	3.89	38.8
w/o S-V2Q	58.6	**3.86**	**39.2**
w/o TLI	59.0	3.97	38.4
w/o T-V2Q	59.1	3.95	38.9

Table 5. The necessity of significant feature extraction, where the data size is the sum of all datasets. w/o sp denotes the feature without spatial dimension. The original data are extracted by HCRN. Our feature denotes the significant feature proposed in this paper. The lower, the better for the count.

Model	TGIF-QA		MSVD-QA	Parameter	Data Size
	Frame	Count			
HCRN (*original*)	55.9	3.82	36.1	44 M	91.4 GB
HCRN (*our feature w/o sp*)	57.0	3.83	38.1	28 M	**10.76 GB**
HCRN (*our feature*)	57.0	3.84	37.8	44 M	91.4 GB
Ours (*original*)	58.2	**3.78**	38.4	**21 M**	91.4 GB
Ours (*our feature w/o sp*)	58.5	3.86	39.0	**21 M**	**10.76 GB**
Ours (*our feature*)	**60.1**	3.80	**39.9**	23M	91.4 GB

As shown in Table 5, the HCRN can achieve better performance than the original algorithm when using the significant features without the spatial dimension proposed in this paper. Moreover, the network parameters can be reduced by nearly **36%** and the visual features that we extracted are only **11.7%** of the original features [7]. The above results strongly prove the redundancy in feature extraction of the current video QA solution and verify the necessity of extracting the significant content of the video. Furthermore, when our features with spatial dimensions were used, HCRN's performance did not improve further. It even decreased in MSVD-QA, indicating that the video QA model needs the ability to process spatial dimension information to obtain better QA capability.

We also conducted the same experiment in our video QA solution. The results show that our solution achieves superior (or equal) performance to HCRN when using features extracted by HCRN, which verifies the superiority of the multi-modal data spatio-temporal reasoning network. It is worth noting that we can see from the counting tasks that the significant features are not superior to those extracted by HCRN. The above result is that the actions in the video always have a particular time rule, which also verifies that the significant feature proposed in this paper can handle these high-frequency repetition count tasks.

4.4. Comparison to State-of-the-Art

We compare our video QA solution against the following state-of-the-art video QA methods.

- Co-mem [31]: A motion-appearance co-memory network that uses appearance and motion semantic clues to compute video QA's attention distribution.
- HME [39]: A model that generates a global context-aware text representation and visual representation by first interacting current inputs with memory content. Then, it integrates visual and textual features for QA reasoning.
- L-GCN [50]: A model that builds the relationships between detected objects for a video QA task by a location-aware graph, which incorporates an object's location features into the graph construction.
- HGA [51]: A deep heterogeneous graph alignment network is designed for QA reasoning with four steps: representation, fusion, alignment, and reasoning.
- QueST [10]: An attention network for spatio-temporal context based on question guidance is proposed, which divides question information into spatial and temporal parts and interprets visual features better under the guidance of the corresponding dimensional question information.
- HCRN [7]: A conditional relationships network, and as the network block to build the video QA solution.
- HOSTR [8]: An object-oriented video QA method, which uses location information to model video entity relations, and obtains fine-grained spatio-temporal representation and QA reasoning.
- DualVGR [6]: A dual visual graph reasoning unit for video QA simulates rich spatio-temporal interactions between video clips related to the question through iterative stacking.
- ACRTransformer [52]: An action-centric relation transformer network, which emphasizes the frames with high actionness probabilities and exploits the interplays between temporal frames.
- HRNAT [53]: A hierarchical representation network with auxiliary tasks, which is used to learn the multi-level representations and obtain syntax-aware video captions.

Table 6 shows the experimental results of the TGIF-QA, MSVD-QA, and MSRVTT-QA datasets. Our model achieves state-of-the-art performance and outperforms existing methods in all tasks except the count task. In particular, we obtain the best performance in the MSRVTT-QA dataset, which shows that our method can handle large datasets well.

Table 6. Comparison of our solution with state-of-the-art methods on several video QA datasets. Mean square error is used as the evaluation metric for the count test, and accuracy for others.

Model	TGIF-QA				MSVD ↑	MSRVTT ↑
	Action ↑	Trans ↑	Frame ↑	Count ↓		
Co-mem [31]	68.2	74.3	51.5	4.10	31.7	32.0
HME [39]	73.9	77.8	53.8	4.02	33.7	33.0
L-GCN [50]	74.3	81.1	56.3	3.95	34.3	——
HGA [51]	75.4	81.0	55.1	4.09	34.7	35.5
QueST [10]	75.9	81.0	59.7	4.19	36.1	34.6
HCRN [7]	75.0	81.4	55.9	3.82	36.1	35.4
HOSTR [8]	75.6	83.0	58.2	**3.65**	39.4	35.9
DualVGR [6]	——	——	——	——	39.0	35.5
ACRTransformer [52]	75.8	81.6	57.7	4.08	——	——
HRNAT [53]	——	——	——	——	38.2	35.3
ours	**77.2**	**83.5**	**60.1**	3.80	**39.9**	**36.7**

The experimental results in Table 6 show that our video QA solution performs better than the most state-of-the-art methods on these three datasets. The improvement is that our video QA solution can accurately obtain significant video information, enabling the video

QA model to obtain the complete visual semantic clues provided by the video. Moreover, it can accurately calculate the spatial dimension information of video frames and the dynamic evolution of this visual information. As a result, it can obtain reasonable and sufficient visual semantic clues. The experiment proves the effectiveness and superiority of our video QA solution.

4.5. Qualitative Results

In order to verify the rationality of using the significant feature extraction module to extract information features, we prepared some visualizations of video significant information distribution. As shown in Figure 4a, there is a strict temporal rule in the count task, which explains why our performance in the count task is not improved significantly. However, in Figure 4b, although the focus is on counting actions, there is no strong regularity in repeating action videos. Therefore, we can obtain excellent performance by capturing such non-uniform distribution. In Figure 4c, Frame QA video has a relatively concentrated sequence of significant frames, which explains why we can still perform better in Frame QA tasks with only a single significant frame. In Figure 4d, the significant information in the state transition video is mainly concentrated in the process of state transformation, which demonstrates the reason that our video QA solution can achieve better performance in state transition tasks. Finally, in Figure 4e,f, video significant information distribution has no significant time rule, verifying that using uniform sampling may lose some important information. It also verifies the rationality of using the significant feature extraction module to extract significant visual features.

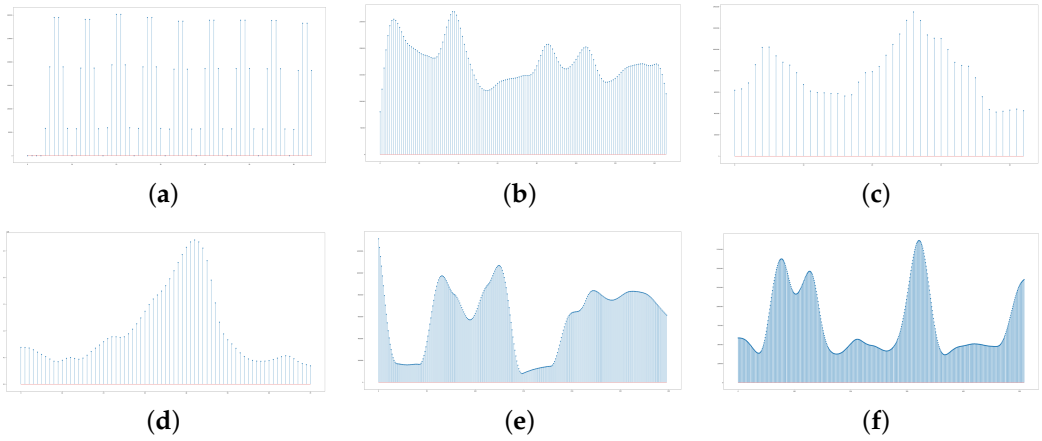

Figure 4. Visualizations of video significant information distribution. Different videos have different significant information distributions. By exploring the video information distribution, the video QA model can better sample video content and obtain the complete storyline. (**a**) Count Task; (**b**) Repeating Action; (**c**) Frame QA; (**d**) State Transition; (**e**) MSVD-QA; (**f**) MSRVTT-QA.

Furthermore, in order to better understand our contribution to visual language interaction in video QA, we have prepared some **visual demonstration results** of visual language interaction. As shown in Figure 5, the visual attention is unencumbered by complex visual information and pays attention to the correct visual information, namely "dog", "man", "violin", "lady", and "onion." The above visual experiment proves that our video QA solution can find the location of the answer semantics in the visual–spatial dimension. Moreover, it demonstrates that our video QA solution can understand the question, deduce the correct answer, and verify our solution's superiority.

We also prepared some visualizations of our **unsuccessful predictions**, as shown in Figure 6, which included four video QA questions. In question (a), although our video QA solution can focus the visual semantics on the correct area, due to a lack of understanding of complex concepts, such as "treadmill," it makes the wrong prediction, "box". In question (b), although our video QA solution can focus the visual semantics on the correct area, some

critical physical features are obscured, leading to false predictions on "woman". In question (c), although our video QA solution can focus the visual semantics on the correct area, our video QA solution does not fully recognize the difference between "shoot" and "fire", thus making the wrong prediction. The most interesting question is question (d). It can be seen from the video that people are using ropes to climb, and our video QA solution accurately understands the semantic clues provided by the video and gives its prediction of "rope", but this is different from the correct answer, "gear". The above visual experiment proves that our video QA solution can accurately understand the semantic clues provided by the video but it failed to distinguish more complex semantic expressions, resulting in wrong predictions.

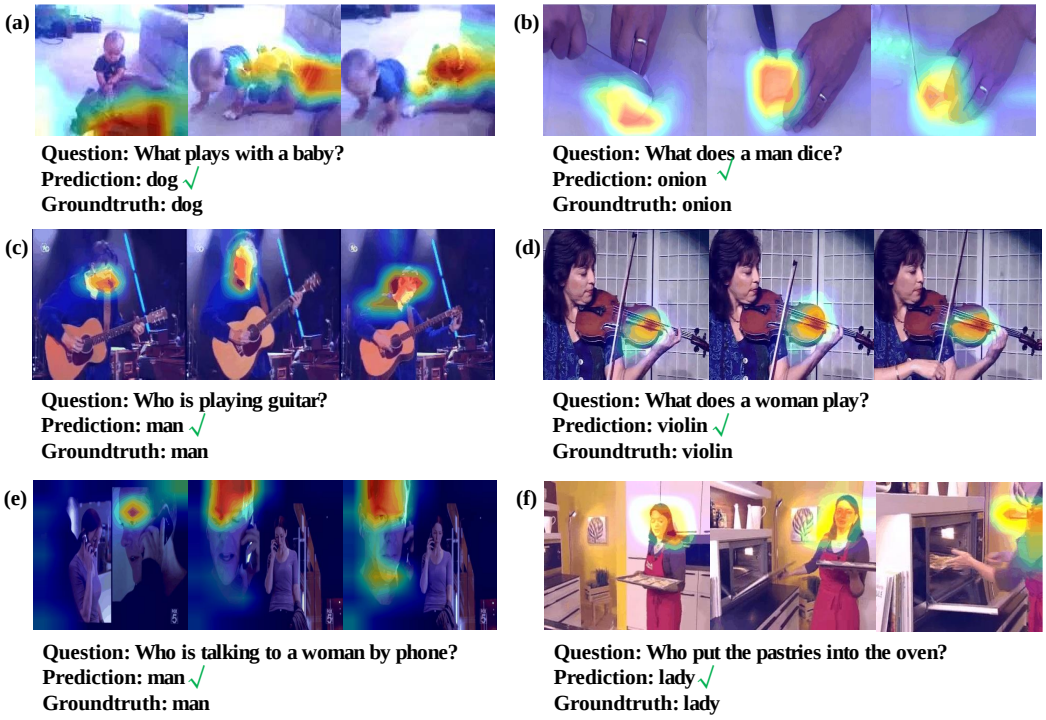

Figure 5. Visualization of visual language interaction. For each QA pair, we visualized the spatial information concerned by the MSTR network in video frames. Our model can accurately pay attention to the spatial semantic content related to the question and deduce the correct answer.

Finally, we prepared some visualization results of **feature distributions**, as shown in Figure 7. In Figure 7a, the original visual features have a broader semantic space than the question, proving that only a small part of the visual semantics within the visual features are associated with the question, which verifies the redundancy of visual information. Furthermore, after the interaction between visual and question through our MSTR network shown in Figure 7b, visual information and question have a similar distribution in the feature space. The above results prove that our video QA solution can effectively obtain visual clues related to the question from complex videos.

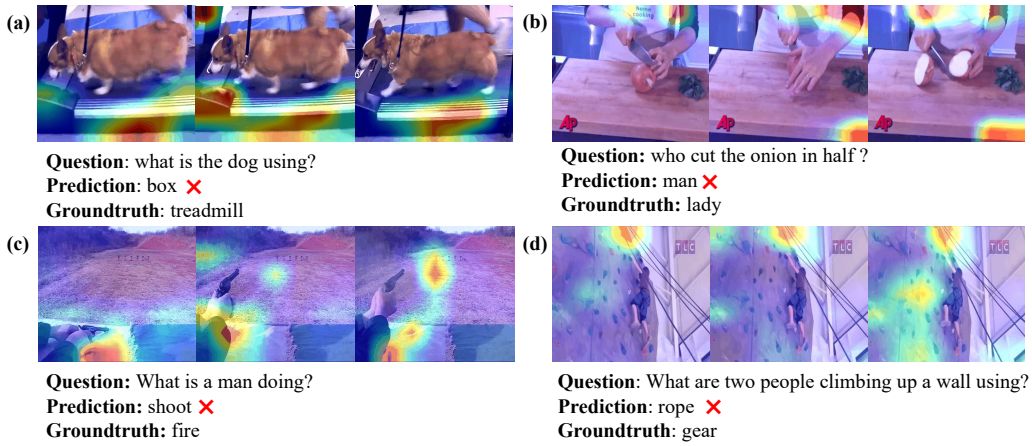

Figure 6. Unsuccessful predictions on the MSVD-QA dataset. For each QA pair, we visualized the spatial information concerned by the MSTR network in video frames. Taking question (**a**), our model can accurately pay attention to the spatial semantic content related to the question. However, due to a lack of understanding of complex concepts, such as "treadmill", it makes the wrong prediction, "box".

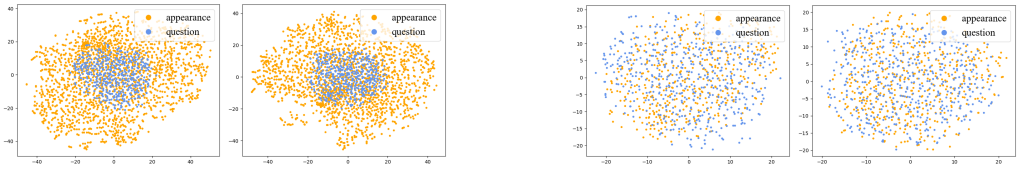

Figure 7. *t*-SNE plots for visualizing the embedding distribution of various features. Original features are derived directly from feature extractors, and MSTR descriptors are outputs of our MSTR network. (**a**) Original features; (**b**) MSTR descriptors.

5. Conclusions

This paper proposes a novel temporally multi-modal semantic reasoning with spatial language constraints video QA solution. Specifically, a significant feature extraction module is designed to capture the significant visual semantic clues that change with the story progression, providing more affluent visual information for the video question reasoning and reducing the redundancy of visual features. Furthermore, we design a new multi-modal data spatio-temporal reasoning network that can model the significant dependency relationship from the spatial dimension of a single frame image to the whole video. The qualitative and quantitative experimental results show that the performance of our method outbalances the state-of-the-art video QA algorithms. In the future, we will study the imbalanced dataset problem and multi-label feature selection problem of video QA and develop an end-to-end video QA network to contribute to artificial intelligence.

Author Contributions: Conceptualization, F.Z. and M.L.; methodology, R.W. and M.L.; software, M.L.; validation, M.L., R.W. and F.Z.; formal analysis, R.W.; investigation, R.W.; resources, R.W.; data curation, M.L. and R.W.; writing—original draft preparation, R.W. and M.L.; writing—review and editing, R.W.; visualization, M.L.; supervision, R.W. and G.L.; project administration, F.Z. and G.L.; funding acquisition, F.Z. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Key-Area Research and Development Program of Guangdong Province (No. 2020B010165001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within this article.

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

References

1. Pham, N.T.; Lee, J.; Kwon, G.; Park, C. Hybrid Image-Retrieval Method for Image-Splicing Validation. *Symmetry* **2019**, *11*, 83. [CrossRef]
2. Tian, P.; Mo, H.; Jiang, L. Image Caption Generation Using Multi-Level Semantic Context Information. *Symmetry* **2021**, *13*, 1184. [CrossRef]
3. Aggarwal, A.; Chauhan, A.; Kumar, D.; Mittal, M.; Roy, S.; Kim, T. Video Caption Based Searching Using End-to-End Dense Captioning and Sentence Embeddings. *Symmetry* **2020**, *12*, 992. [CrossRef]
4. Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; Chua, T. Attentive moment retrieval in videos. In Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR 2018), Ann Arbor, MI, USA, 8–12 July 2018; Collins-Thompson, K., Mei, Q., Davison, B.D., Liu, Y.; Yilmaz, E., Eds.; ACM: New York, NY, USA, 2018; pp. 15–24.
5. Wan, Z.; He, H. AnswerNet: Learning to Answer Questions. *IEEE Trans. Big Data* **2019**, *5*, 540–549. [CrossRef]
6. Wang, J.; Bao, B.; Xu, C. DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering. *IEEE Trans. Multimed.* **2021**. [CrossRef]
7. Le, T.M.; Le, V.; Venkatesh, S.; Tran, T. Hierarchical conditional relation networks for video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9969–9978.
8. Dang, L.H.; Le, T.M.; Le, V.; Tran, T. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021), Montreal, QC, Canada, 19–27 August 2021; pp. 636–642.
9. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual question answering. In Proceedings of the ICCV IEEE Computer Society, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
10. Jiang, J.; Chen, Z.; Lin, H.; Zhao, X.; Gao, Y. Divide and Conquer: Question-Guided Spatio-Temporal Contextual Attention for Video Question Answering. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11101–11108.
11. Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; Kim, G. TGIF-QA: Toward spatio-temporal reasoning in visual question Answering. In Proceedings of the CVPR IEEE Computer Society, Honolulu, HI, USA, 21–26 July 2017; pp. 1359–1367.
12. 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 ACM Multimedia, Mountain View, CA USA, 23–27 October 2017; pp. 1645–1653.
13. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the CVPR IEEE Computer Society, Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the CVPR IEEE Computer Society, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Tran, D.; Bourdev, L.D.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
16. Zhao, Z.; Zhang, Z.; Xiao, S.; Yu, Z.; Yu, J.; Cai, D.; Wu, F.; Zhuang, Y. Open-Ended Long-form Video Question Answering via Adaptive Hierarchical Reinforced Networks. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018), Stockholm, Sweden, 13–19 July 2018; pp. 3683–3689.
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
18. Yu, Y.; Ko, H.; Choi, J.; Kim, G. End-to-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 3261–3269.
19. Xie, S.; Girshick, R.B.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.
20. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6546–6555.
21. Zhu, L.; Xu, Z.; Yang, Y. Bidirectional Multirate Reconstruction for Temporal Modeling in Videos. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 1339–1348.
22. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
23. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Korea, 27 October–2 November 2019; pp. 6201–6210.
24. Wang, W.; Huang, Y.; Wang, L. Long video question answering: A Matching-guided Attention Model. *Pattern Recognit.* **2020**, *102*, 107248. [CrossRef]

25. Kim, S.; Jeong, S.; Kim, E.; Kang, I.; Kwak, N. Self-supervised Pre-training and Contrastive Representation Learning for Multiple-choice Video QA. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual Event, 2–9 February 2021; pp. 13171–13179.
26. 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.
27. Zeng, K.; Chen, T.; Chuang, C.; Liao, Y.; Niebles, J.C.; Sun, M. *Leveraging Video Descriptions to Learn Video Question Answering*; AAAI Press: Palo Alto, CA, USA, 2017; pp. 4334–4340.
28. Chowdhury, M.I.H.; Nguyen, K.; Sridharan, S.; Fookes, C. Hierarchical Relational Attention for Video Question Answering. In Proceedings of the ICIP IEEE, Athens, Greece, 7–10 October 2018; pp. 599–603.
29. Kim, J.; Lee, S.; Kwak, D.; Heo, M.; Kim, J.; Ha, J.; Zhang, B. Multimodal Residual Learning for Visual QA. In Proceedings of the Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 361–369.
30. Xiong, C.; Merity, S.; Socher, R. Dynamic Memory Networks for Visual and Textual Question Answering. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA 19–24 June 2016; Volume 48, pp. 2397–2406.
31. Gao, J.; Ge, R.; Chen, K.; Nevatia, R. Motion-appearance co-memory networks for video question answering. In Proceedings of the CVPR Computer Vision Foundation/IEEE Computer Society, Salt Lake, UT, USA, 18–22 June 2018; pp. 6576–6585.
32. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A.J. Stacked Attention Networks for Image Question Answering. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
33. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the CVPR Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
34. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5-10 December 2016; pp. 289–297.
35. Zhu, L.; Xu, Z.; Yang, Y.; Hauptmann, A.G. Uncovering the Temporal Context for Video Question Answering. *Int. J. Comput. Vis.* **2017**, *124*, 409–421. [CrossRef]
36. Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; Gan, C. *Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering*; AAAI Press: Palo Alto, CA, USA, 2019; pp. 8658–8665.
37. Li, X.; Gao, L.; Wang, X.; Liu, W.; Xu, X.; Shen, H.T.; Song, J. Learnable Aggregating Net with Diversity Learning for Video Question Answering. In Proceedings of the 27th ACM International Conference on Multimedia (MM 2019), Nice, France, 21–25 October 2019; pp. 1166–1174.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
39. Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; Huang, H. Heterogeneous memory enhanced multimodal attention model for video question answering. In Proceedings of the CVPR Computer Vision Foundation/IEEE, Long Beach, CA, USA, 16–20 June 2019; pp. 1999–2007.
40. Kim, K.; Choi, S.; Kim, J.; Zhang, B. Multimodal dual attention memory for video story question answering. In Proceedings of the Computer Vision ECCV 2018—15th European Conference, Munich, Germany, 8–14 September 2018; Volume 11219; pp. 698–713.
41. Zhao, Z.; Yang, Q.; Cai, D.; He, X.; Zhuang, Y. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia 19–25 August 2017; pp. 3518–3524.
42. Zhao, Z.; Jiang, X.; Cai, D.; Xiao, J.; He, X.; Pu, S. Multi-Turn Video Question Answering via Multi-Stream Hierarchical Attention Context Network. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweeden, 13–19 July 2018; pp. 3690–3696.
43. Kim, J.; Ma, M.; Kim, K.; Kim, S.; Yoo, C.D. Progressive attention memory network for movie story question answering. In Proceedings of the CVPR Computer Vision Foundation/IEEE, Long Beach, CA, USA, 16–20 June 2019; pp. 8337–8346.
44. Song, X.; Shi, Y.; Chen, X.; Han, Y. Explore Multi-Step Reasoning in Video Question Answering. In Proceedings of the ACM Multimedia, Amsterdam, The Netherlands, 12–15 June 2018; pp. 239–247.
45. Zhao, Z.; Zhang, Z.; Xiao, S.; Xiao, Z.; Yan, X.; Yu, J.; Cai, D.; Wu, F. Long-Form Video Question Answering via Dynamic Hierarchical Reinforced Networks. *IEEE Trans. Image Process.* **2019**, *28*, 5939–5952. [CrossRef] [PubMed]
46. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
47. Clevert, D.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv* **2015**, arXiv:1511.07289.
48. Li, Y.; Song, Y.; Cao, L.; Tetreault, J.R.; Goldberg, L.; Jaimes, A.; Luo, J. TGIF: A New Dataset and Benchmark on Animated GIF Description. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 4641–4650.
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

50. 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 (AAAI2020), New York, NY, USA, 7–12 February 2020; pp. 11021–11028.
51. 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.
52. Zhang, J.; Shao, J.; Cao, R.; Gao, L.; Xu, X.; Shen, H.T. Action-Centric Relation Transformer Network for Video Question Answering. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 63–74. [CrossRef]
53. Gao, L.; Lei, Y.; Zeng, P.; Song, J.; Wang, M.; Shen, H.T. Hierarchical Representation Network With Auxiliary Tasks for Video Captioning and Video Question Answering. *IEEE Trans. Image Process.* **2022**, *31*, 202–215. [CrossRef] [PubMed]