# VideoQA-TA: Temporal-Aware Multi-Modal Video Question Answering

**Zhixuan Wu**[1], **Bo Cheng**[1*], **Jiale Han**[2], **Jiabao Ma**[1], **Shuhao Zhang**[1], **Yuli Chen**[1], **and Changbao Li**[3]
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2]Hong Kong University of Science and Technology
[3]North China Institute of Computing Technology
{wzxmogu, chengbo, jiabao.m, 2020111429, chenyuli}@bupt.edu.cn, jialehan@ust.hk, lichangbao_1@163.com
* is corresponding author

## Abstract

Video question answering (VideoQA) has recently gained considerable attention in the field of computer vision, aiming to generate answers rely on both linguistic and visual reasoning. However, existing methods often align visual or textual features directly with large language models, which limits the deep semantic association between modalities and hinders a comprehensive understanding of the interactions within spatial and temporal contexts, ultimately leading to sub-optimal reasoning performance. To address this issue, we propose a novel temporal-aware framework for multi-modal video question answering, dubbed ***VideoQA-TA***, which enhances reasoning ability and accuracy of VideoQA by aligning videos and questions at fine-grained levels. Specifically, an effective **S**patial-**T**emporal **A**ttention mechanism (STA) is designed for video aggregation, transforming video features into spatial and temporal representations while attending to information at different levels. Furthermore, a **T**emporal **O**bject **I**njection strategy (TOI) is proposed to align object-level and frame-level information within videos, which further improves the accuracy by injecting explicit temporal information. Experimental results on MSVD-QA, MSRVTT-QA, and ActivityNet-QA datasets demonstrate the superior performance of our proposed method compared with the current SOTAs, meanwhile, visualization analysis further verifies the effectiveness of incorporating temporal information to videos [1].

## 1 Introduction

Visual question answering (VQA) poses a meaningful task and has drawn increasing interest in recent years (Li et al., 2024; Hong et al., 2024) due to its potential for advancing the integration of visual and linguistic understanding. Thanks to the development of language modeling and multimodal learning (Han et al., 2023; Zha et al., 2024), video question answering (VideoQA) has achieved remarkable progress in understanding and reasoning over both visual and linguistic information simultaneously, enabling more accurate and context-aware responses to complex questions based on video contents, e.g., human-machine interaction (Yu et al., 2024; Zou et al., 2024; Ma et al., 2021), intelligent driving (Park et al., 2024; Xu et al., 2024), and intelligent interaction (Huang et al.). Despite the achievements, traditional multimodal models lack deep semantic alignment between modalities, limiting their ability to understand multimodal contexts. Moreover, these models exhibit weak capabilities in modeling spatial-temporal information and long-sequence dependencies, failing to effectively capture interaction relationships in complex scenarios.

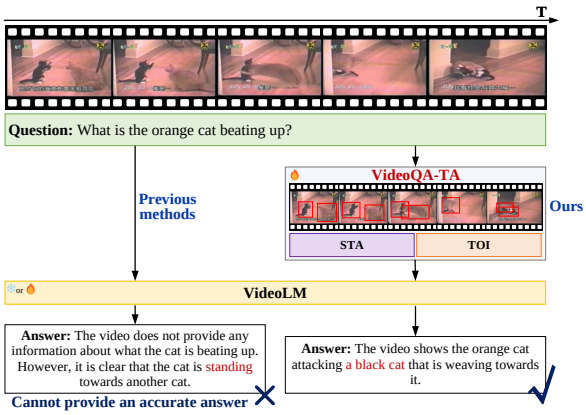



Figure 1: Comparison between our proposed method VideoQA-TA and the recent VideoQA method like Video-LLaVA. Like most prior methods, Video-LLaVA is a one-stage suited for VideoQA. Thus, as illustrated in the figure, it fails to answer a question that requires reasoning about temporal in a video. In comparison, our method can make temporal reasoning efficiently and produce the correct answer.

To this end, Large Language Models (LLMs) (Ataallah et al., 2024; Liu et al., 2024; Islam and

[1]https://github.com/YALYAshley/VideoQA-TA



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 7239–7252
January 19–24, 2025. 

Moushi, 2024) have been recently proposed and shown remarkable improvements in accuracy reasoning, enabling their widespread application in VideoQA tasks. Thanks to the well-designed visual encoder followed by LLMs, some attempts (Zhang et al., 2023b,a; Li et al., 2023b; Song et al., 2024b; Cheng et al., 2024a) strive to aggregate the entire video into a coarse-grained global representation, i.e., complex long-range temporal modeling, while struggling to accurately capture and preserve the temporal coherence and inter-frame dependencies throughout video sequences. As shown in Figure 1, Video-LLaVA (Lin et al., 2023) fails to identify what the orange cat is attacking: *"The video does not provide any information about what the cat is beating up"*. This lack of temporal understanding may prevent the model from correctly identifying and describing the development of events in the video, resulting in inaccurate or incomplete descriptions.

With these in mind, we propose a temporal-aware framework for multi-modal video question answering, named ***VideoQA-TA***, which ensures that the model can understand videos from both spatial and temporal perspectives, achieving video question answering (VideoQA). Specifically, VideoQA-TA first adopts a structure similar to the existing multi-modal models, which includes a visual encoder to extract visual features, a text encoder to obtain text features, and a fine-tuned large language model for answer prediction. Then, we use a **S**patial-**T**emporal **A**ttention mechanism (STA) for video aggregation to obtain the spatial and temporal salient regions in the video to improve the accuracy of VideoQA. Meanwhile, **T**emporal **O**bject **I**njection strategy (TOI) for video alignment is proposed to prevent the direct embedding of object information from making it incomprehensible to large models. We use the visual detection ability of RAM++ model to mine informative object cues for video understanding and a prompt template is sophisticatedly designed to facilitate better integration of video, object, and question features. Eventually, the answer is obtained through the fine-tuned Vicuna (Zheng et al., 2024). Our main contributions are summarized as follows:

- We devise a spatial-temporal attention mechanism for video aggregation, which utilizes spatial and temporal attention to focus on relevant video contents at different levels, thus removing irrelevant information.
- We propose a temporal object injection strategy for video alignment, this strategy injects explicit information into the model, aligns spatial and temporal data from object-level and frame-level to improve object relationship clarity and compositional reasoning.
- Experimental results show that proposed method performs well on MSVD-QA, MSVTT-QA, and ActivityNet-QA datasets, which are evaluated by Top-1 accuracy and GPT-3.5 [2].

## 2 Related Work

### 2.1 Video Question Answering

The main challenge in VideoQA is the semantic gap between visual understanding and natural language, and many studies (Xiao et al., 2024; Liao et al., 2024) have been proposed to address this issue. MoReVQA (Min et al., 2024) adopts a decomposed multi-stage reasoning framework, which, in contrast to earlier approaches with a single planning stage, significantly enhances robustness and accuracy in complex VideoQA scenarios. This improvement is achieved through event parsing, video content grounding, and a final reasoning stage. Video-LLaMA (Zhang et al., 2023b), LLoVi (Zhang et al., 2023a) integrate visual encoders, text encoders, and the alignment of embedding spaces with LLMs to utilize the reasoning capabilities of these models to answer questions. MovieChat (Song et al., 2024a), Glance and Focus (Bai et al., 2024a) combine different levels of semantic information by pre-training two-way dynamic memory networks and using memory cues to achieve VideoQA. Video-LLaVA (Lin et al., 2023) combines text encoder in Llama, allowing the model to learn interactions between modalities from a unified visual representation. Using LLM is efficiently improve the accuracy of VideoQA (Wu et al., 2024; Liu and Wan, 2024). In this paper, we use pre-trained Vicuna for VideoQA.

### 2.2 Learning with Temporal Relation

Temporal information injection (Zhou and Wu, 2023; Ma et al., 2024) has greatly succeeded on many vision and language tasks. Some methods (Jiang et al., 2020; Li et al., 2023c) incorporate attention mechanism modules within the feature ex-

[2] https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

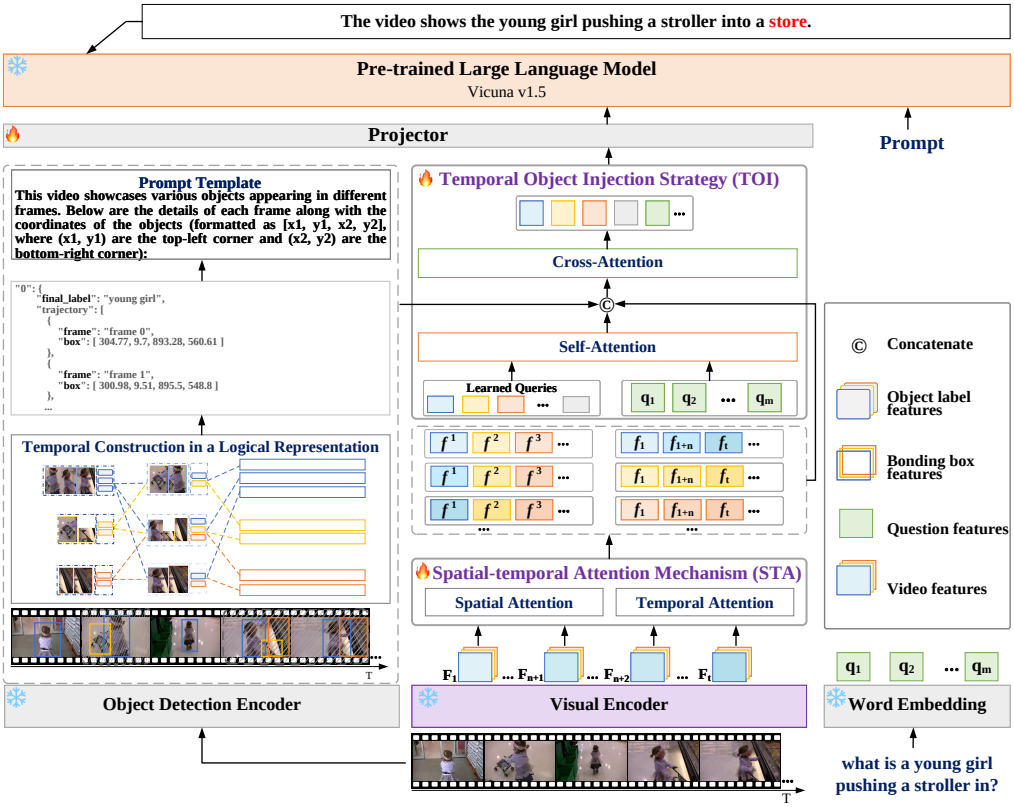



Figure 2: Our proposed method VideoQA-TA framework. $F_t$ represents the video features, $f^1, f^2, f^3$ represents the different patches in the same frames, $f_1, f_{1+n}, f_t$ represents the same patches in different frames. $q_1, q_2, q_m$ represents the questions.

traction components for temporal, thereby identifying temporal data corresponding to questions. Gao et al. (Gao et al., 2023) introduces a multi-scale temporal attention framework that refines temporal focus at varying granularities, enhancing the identification of key temporal segments. Bai et al. (Bai et al., 2024b) incorporates event-based attention to detect key events, ensuring precise alignment of temporal data with the question. Furthermore, other methods (Zhang et al., 2021; Ahmad et al., 2023) propose a temporal module to capture local and global contexts, establishing inter- and intra-modal correlations. Nie et al. (Nie et al., 2024) introduces a dynamic graph convolution, adapting graph structures in real-time to better capture complex temporal relationships. These correlations facilitate the semantic alignment of visual and textual data, thereby enhancing the understanding of the underlying relationships within and across the different modalities. Given this, it is crucial to construct different strategies performing the temporal correlation information among videos. In this paper, we use a spatial-temporal attention mechanism to aggregate spatial and temporal associations and propose a temporal object injection strategy to align fine-grained levels information in VideoQA tasks, which can enhance the performance of temporal reasoning and improve answer accuracy.

## 3 Methodology

We propose a temporal-aware framework for multi-modal video question answering, named ***VideoQA-TA***, as shown in Figure 2. In this section, we will elaborate the each component of our VideoQA-TA. Specifically, section 3.1 introduces feature extraction, section 3.2 presents **S**patial-**T**emporal **A**ttention mechanism (STA) for video aggregation, section 3.3 details **T**emporal **O**bject **I**njection strategy (TOI) for video alignment, and section 3.4 is answer prediction.

### 3.1 Feature Extraction

For videos, like most multi-modal VideoQA tasks, the pretrained CLIP ViT-G serves as the frozen visual encoder to extract embeddings of each frame individually, and obtains video features $F_t = \{f_t\}_{t=1}^{N_t} \in \mathbb{R}^{N_t \times N_p \times C}$, $f_t = \{f_t^p\}_{p=1}^{N_p} \in \mathbb{R}^{N_p \times C}$, where $f_t^p$ denotes the $t$-th frame containing $N_p$

patches, $N_t$ is the number of frames, and $C$ denotes the embedding dimension. Concurrently, the pre-trained RAM++ serves as the frozen object detection encoder to extract object labels $L_t = \{l_t^n\}_{n=1}^{N_n}$ and bounding boxes $B_t = \{b_t^n\}_{n=1}^{N_n}$, where $l_t^n$ and $b_t^n$ denote the $t$-th frame containing $N_n$ objects, $b_t^n$ is composed of $[x_1, y_1, x_2, y_2]$ , $x_1$ denotes the horizontal coordinate of the top-left corner, $y_1$ denotes the vertical coordinate of the top-left corner, $x_2$ denotes the horizontal coordinate of the top-right corner, and $y_2$ denotes the vertical coordinate of the top-right corner. For language, we tokenize the question $Q = \{q_m\}_{m=1}^{N_m} \in \mathbb{R}^{N_m \times C}$ into a sequence of words, where $N_m$ denotes the sequence length of the question (i.e., the number of words), and $q_m$ represents the $m$-th word in the sequence. The tokenized sequence is then fed into BLIP-2 for further processing.

### 3.2 Spatial-Temporal Attention

Considering the spatial information within individual frames while also paying attention to the temporal changes between frames, we utilize STA, which builds upon the MA-TMM (He et al., 2024) architecture by incorporating spatial attention and temporal attention to improve feature extraction performance in VideoQA tasks.

**Spatial attention.** We represent as varying degrees of attention and aggregates with spatial features $F_{spa} = \{f^p\}_{p=1}^{N_p} \in \mathbb{R}^{N_p \times C'}$ to focus on salient spatial areas of videos. We aggregate the $i$-th vector of the spatial features in $t$-th frame, denoted as $F_{t,i}$, and average features across time dimension:

$$A_{spa,t} = \frac{\exp\left(W_{spa,i} \cdot F_{t,i}\right)}{\sum_j \exp\left(W_{spa,j} \cdot F_{t,j}\right)} \tag{1}$$

$$F_{spa} = \frac{1}{T} \sum_{t=1}^{T} (A_{spa,t} \cdot F_t) \cdot W'_{spa} \tag{2}$$

where $W_{spa,i}$, $W_{spa,j}$ and $W'_{spa}$ denote learnable spatial attention weight, respectively.

**Temporal attention.** We use temporal features $F_{tem} = \{f_t\}_{t=1}^{N_t} \in \mathbb{R}^{\mathrm{N_t} \times C'}$ to represent frames that are more critical for understanding the video content. Similar to the spatial attention:

$$A_{tem,p} = \frac{\exp\left(W_{tem,i} \cdot F_{t,i}\right)}{\sum_j \exp\left(W_{tem,j} \cdot F_{t,j}\right)} \tag{3}$$

$$F_{tem} = \frac{1}{P} \sum_{p=1}^{P} (A_{tem,p} \cdot F_t) \cdot W'_{tem} \tag{4}$$

where $W_{tem,i}$, $W_{tem,j}$ and $W'_{tem}$ denote learnable temporal attention weight, respectively.

### 3.3 Temporal Object Injection Strategy

We focus on adaptively generating meaningful temporal information that can provide relevant information about what is happening in the video to the LLM, thus facilitating fine-grained levels alignment, as is shown in Appendix A. Given object label $L_t$ and bounding box $B_t$ from each frame, we use a unified sampling strategy, which can construct a temporal sequence, denoted as $E = Concat[Prompt, L_t, B_t]$. In details, we construct a logical representation $< object-frame-bounding\ box >$, connecting different information of the object using semantic relationships. For the same object, by retrieving the bounding boxes in different frames within video, we inject information that includes both spatial information and temporal sequence characteristics. During the experimental phase, such information may contain lengthy (>1K characters), noisy, and potentially redundant/irrelevant textual input sequences, which encounter difficulties. To address this problem, we design a more specialized LLM prompt template that incorporates object features, thereby enhancing the input of video temporal features. The specific *prompt* template as follows:

> "This video showcases various objects appearing in different frames. Below are the details of each frame along with the coordinates of the objects (formatted as $[x_1, y_1, x_2, y_2]$, where $(x_1, y_1)$ are the top-left corner and $(x_2, y_2)$ are the bottom-right corner): In the $\{frame\_id\}$, it includes: $\{object\_name\}$ with $\{object\_bonding\_box\}$, $\{object\_name\}$ with $\{object\_bonding\_box\}$.... "

where $\{frame\_id\}$ equals $t$, $\{object\_name\}$ equals $L_t$, $\{object\_bonding\_box\}$ equals $B_t$. $E$ is used as input to the Q-former, obtaining specific visual context-aware representations and frame descriptions. Through this design, not only the static attributes of the object in a frame is preserved, but the dynamic characteristics of the object over time are also captured through the temporally serialized object labels and bounding boxes.

For aligning the visual embedding to the textual and prompt inputs embedding, we maintain a dy-

namic memory via the Q-Former. The language memory bank stores the question features extracted from the frozen word embedding, which can be formulated as:

$$Q_q = q_t W_Q \tag{5}$$

$$K_q = [q_1||\dots||q_t]W_K \tag{6}$$

$$V_q = [q_1||\dots||q_t]W_V \tag{7}$$

where $q_t$ denotes the learned important information that has been learned and is specific to each video up to the current timestep $t$, $W_Q$, $W_K$, $W_V$ denote learnable weight matrix for transforming the input query $q_t$ into the query, key and value space, respectively. $Q_q$, $K_q$, $V_q$ represent the query, key and value vectors, respectively. $[\cdot \ || \ \cdot]$ is the concatenation operator. Then we apply the self-attention operation to obtain $Q' = \{q'_m\}_{m=1}^{N_m}$:

$$Q' = \sigma(\frac{Q_q K_q^T}{\sqrt{d_k}})V_q \tag{8}$$

where $d_k$ is a factor in self-attention.

We represent the input $F_{spa}$ and $F_{tem}$ as the set of question-dependent video features, that is going to be useful for VideoQA task. Then, we fuse spatial features, temporal features, object features and question features through a cross-attention to achieve fine-grained level alignment, where the spatial features, temporal features and object features are regarded as key and value vectors while question features and object features serve as query vector. Formally, the updated aligned features $F_i$:

$$F_{1i} = \textit{MLP}([F_{spa}||F_{tem}||E]) \tag{9}$$

$$F_{2i} = \textit{MLP}([Q'||E]) \tag{10}$$

$$F_i = \sigma\left(\frac{P_Q(F_{2i})P_K(F_{1i})^T}{\sqrt{d_k}}\right)P_V(F_{1i}) \tag{11}$$

where $P_Q$, $P_K$, $P_V$ same as MeMViT (Wu et al., 2022), which pools spatial-temporal dimensions of $Q$, $K$, and $V$.

### 3.4 Answer Prediction

We learn a projection layer that correlates video features to text, with the language model being frozen, which can be written as:

$$z_i = F_i W_{proj} \tag{12}$$

where $z_i$ denotes the textual feature representation obtained through the projection layer, $W_{proj}$ denotes the learnable parameters with projection layer. Same as MA-TMM, we quantify the relevance between the question and the video content by calculating cosine similarity between the aligned features of question and video, thus providing accurate answers for VideoQA task. We utilize an annotated dataset that includes video and text pairs and perform supervised learning through with standard cross-entropy loss. This supervised approach enables the model to predict answer $A$ in an autoregressive manner, enhancing prediction accuracy. The formula is as follows:

$$\mathcal{L} = -\sum_i \log P(a_i \mid z_i, q_i) \tag{13}$$

where $q_i$ is the input question, and $a_i$ represents the corresponding answer. We adjust the parameters of the Q-Former while keeping the weights of the visual encoder and the language model fixed.

## 4 Experimental Setup

### 4.1 Datasets and Implementation Details

We conduct experiments on MSVD-QA, MSRVTT-QA, and ActivityNet-QA datasets. Detailed dataset information is provided in Appendix B.

We initialize the video encoder with CLIP (ViT-G) [3](Radford et al., 2021), the object detection with pre-trained RAM++ (Huang et al., 2023), and the word embedding with BLIP-2 (Li et al., 2023a). We use the pre-trained Q-Former weights from InstructBLIP [4] and adopt Vicuna-7B (Zheng et al., 2024) as the LLM. All the experiments are conducted on 4 A40 GPUs. We use the AdamW optimizer with a weight decay 0.05 and betas (0.9, 0.999). Cross-entropy loss is employed with an initial learning rate of $1 \times e^{-4}$, the batch size of 32 per GPU, and training for 5 epochs. We extract video frames from each video at 10 fps, based on the annotations of each dataset.

### 4.2 Evaluation Metrics

We use Top-1 accuracy (Top-1 acc.), accuracy (Acc.) and score by GPT-3.5 for test metrics to evaluate models. Top-1 acc. is determined by checking whether each predicted answer matches the true answer, which can be formulated as:

$$\text{Top-1 acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\text{pred}_{i,1} = \text{gt}_i\right) \tag{14}$$

[3]https://huggingface.co/openai/clip-vit-large-patch14

[4]https://github.com/salesforce/LAVIS/tree/main/projects/instructblip

| Method | LLM | MSVD-QA | MSRVTT-QA | ActivityNet-QA |
|---|---|---|---|---|
| ***Traditional methods without LLM*** | | | | |
| mPLUG-2(Xu et al., 2023) | - | 58.1 | 48.0 | - |
| ***Short VideoQA*** | | | | |
| Video-LLaMA(Zhang et al., 2023b) | Vicuna - 7B | 58.3 | 46.5 | 45.5 |
| Video-LLaMA2(Cheng et al., 2024b) | Vicuna - 7B | <u>60.6</u> | 46.9 | <u>51.8</u> |
| Video-ChatGPT(Maaz et al., 2024) | Vicuna - 7B | 59.5 | 46.8 | 46.1 |
| Video-LLaVA(Lin et al., 2023) | Vicuna - 7B | 60.1 | 47.3 | 48.4 |
| ***Long-term VideoQA*** | | | | |
| Chat-UniVi(Jin et al., 2024) | Vicuna - 7B | 59.8 | 47.2 | 50.3 |
| LLaMA-VID(Li et al., 2023b) | Vicuna - 7B | <u>60.6</u> | 47.1 | 50.7 |
| LLaMA-VID(Li et al., 2023b) | Vicuna-13B | 59.8 | 47.4 | 51.4 |
| MiniGPT4-Video(Ataallah et al., 2024) | Llama 2-7B | 59.6 | <u>48.9</u> | 49.8 |
| MiniGPT4-Video(Ataallah et al., 2024) | Mistral - 7B | 58.7 | 47.5 | 48.6 |
| MA-LMM(He et al., 2024) | Vicuna - 7B | <u>60.6</u> | 48.5 | 49.8 |
| **VideoQA-TA** | Vicuna - 7B | **66.5** | **51.3** | **52.2** |

Table 1: Comparison with state-of-the-art methods on the VideoQA task in Top-1 acc.. The best result is highlighted in bold, and the second best is underlined.

where $N$ is the total number of prediction samples, $pred_{(i,1)}$ is the first answer in the list of predicted answers for $i$-th samples, $gt_i$ is the ground-truth answer for $i$-th samples, $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 the condition is true and 0 otherwise. Acc. evaluates the correctness of predicted answers against correct answers using the GPT-3.5 model. It measures the proportion of correct predictions (labeled 'yes') out of the total number of predictions (both 'yes' and 'no'). Score, on the other hand, quantifies the degree of correctness of each prediction on a scale from 0 to 5, with higher scores indicating a closer match between the predicted and correct answers.

### 4.3 Baselines

In order to validate the effectiveness of the proposed method, we compare with three kinds of baselines. ***Traditional methods without LLM:*** 1) mPLUG-2 (Xu et al., 2023), a modular multimodal model for text, image, and video tasks.

***Short VideoQA methods***: 2) Video-LLaMA (Zhang et al., 2023b), unifies visual representation with the linguistic feature space, enhancing video analysis with LLaMA foundation, 3) Video-LLaMA2 (Cheng et al., 2024b), integrates custom spatial-tiemporal convolution (STC) connectors and jointly trained audio branches, advancing video comprehension skills, 4) Video-ChatGPT (Maaz et al., 2024), integrates multimodal discussions with contextual awareness, and 5) Video-LLaVA (Lin et al., 2023), combines language and visual processing, optimizing for video-related tasks and interactions.

***Long-term VideoQA methods***: 6) Chat-UniVi (Jin et al., 2024), a unified visual representation, 7) LLaMA-VID (Li et al., 2023b), assigns the value of two tokens to an image within LLMs, 8) MiniGPT4-Video (Ataallah et al., 2024), uses interleaved visual-textual tokens, and 9) MA-LMM (He et al., 2024), a memory-augmented multi-modal model that specializes in long-term video understanding.

## 5 Results and Discussion

### 5.1 Comparison to Baselines

We provide the overall evaluation results of our method and baselines in Table 1.

1) **MSVD-QA**. Our proposed model VideoQA-TA achieves 66.5% (Top-1 accuracy). Particularly noteworthy is the improvement in Short VideoQA, where VideoQA-TA outperforms Video-LLaMA2 in Top-1 accuracy (66.5% $v.s.$ 60.6%). Video-LLaMA2 focuses on processing multiple frames simultaneously but is constrained by GPU memory limitations. In contrast, VideoQA-TA enhances performance gain mainly from learning spatial and temporal features, improving its ability to interpret short video dynamics and provide more accurate, relevant answers.

2) **MSRVTT-QA**. MSRVTT-QA is a large-scale dataset from web, which high demands on understanding capabilities. Our proposed model VideoQA-TA outperforms MA-LMM in Top-1 ac-

| STA | TOI | MSVD-QA | | MSRVTT-QA | | ActivityNet-QA | |
|---|---|---|---|---|---|---|---|
| | | Top-1 acc. | Acc. | Top-1 acc. | Acc. | Top-1 acc. | Acc. |
| ✗ | ✗ | 60.6 | 72.8 | 48.5 | 59.5 | 49.8 | 45.0 |
| ✓ | ✗ | 61.4 | 74.9 | 49.1 | 59.6 | 50.3 | 45.8 |
| ✗ | ✓ | 65.7 | 76.5 | 50.7 | 60.3 | 51.6 | 50.0 |
| ✓ | ✓ | **66.5** | **76.7** | **51.3** | **61.1** | **52.2** | **50.8** |

Table 2: Top-1 acc. and Acc. with different components. The best result is highlighted in bold.

curacy (51.3% $v.s.$ 48.5%). This result indicates that VideoQA-TA effectively enhances comprehension and answer accuracy on complex video-based questions.

3) **ActivityNet-QA**. Different from MSVD-QA and MSRVTT-QA datasets, there are much longer videos in ActivityNet-QA with questions that require complex inference to derive the answers. Notably, in Long-term VideoQA, our proposed model VideoQA-TA outperforms Chat-UniVi in Top-1 accuracy (52.2% $v.s.$ 50.3%). Chat-UniVi is a unified visual representation method that employs dynamic tokens to simultaneously represent images and videos, capturing both high-level concepts and low-level details through multi-scale representations. In contrast, VideoQA-TA fine-tunes weights from the pre-trained QFormer and incorporates temporal information using a long-term memory bank. This result indicates that VideoQA-TA's ability to analyze long-term temporal dependencies enables it to track and relate information over extended periods, yielding more precise and coherent answers.

### 5.2 Ablation Study

**Prompt analysis.** We provide the ablation study on the different prompt in VideoQA-TA. For details of each prompt, please refer to Appendix C. Our results show that different prompt will bring different results. For comparison fairness, comparison with our proposed method VideoQA-TA, we only replace different prompts and experiment on the MSVD-QA, MSRVTT-QA and ActivityNet-QA datasets to verify the superiority of $E$ in TOI, as shown in Figure 3. $E$ achieves 66.5% (Top-1 accuracy), the performance is higher at 4.6%, 2.8%, 1.7%, 5.9% than $E_1$, $E_2$, $E_3$, $E_4$ on MSVD-QA dataset, respectively. Although $E_4$ adds temporal objectives and object locations, it fails to understand the bounding box associated with the $\{frame\_id\}$, leading to worse performance. $E$ performances significantly exceeds that of other prompts, as it is capable of generating temporal information from videos.

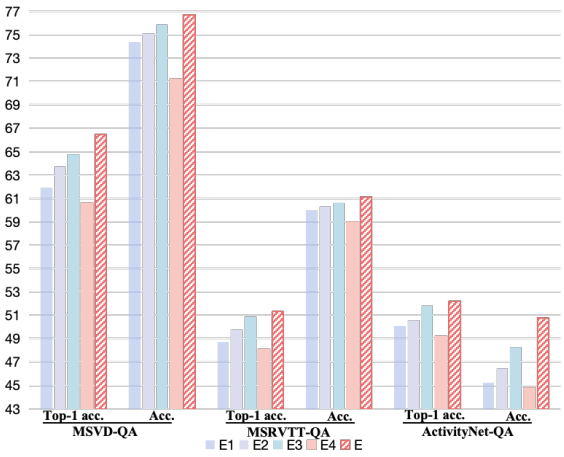



Figure 3: Ablation study showing Top-1 acc. and Acc. to illustrate the impact of prompt designs. $E_1$ is $Concat[Prompt_1]$, $E_2$ is $Concat[Prompt_2, L_t]$, $E_3$ is $Concat[Prompt_4, L_t, B_t]$, $E_4$ is $Concat[Prompt_4, L_t, B_t]$, $E$ is $Concat[Prompt, L_t, B_t]$.

**Contribution of each component.** To verify the effectiveness of each component, we use MA-LMM as the baseline network in this paper and experiment with combinations of components on MSVD-QA, MSRVTT-QA and ActivityNet-QA datasets, as is shown in Table 2. Our proposed model VideoQA-TA achieves an improvement of approximately 5.9% in Acc.. Incorporating STA and TOI into the baseline results in improvements of about 5.1% and 0.8%, respectively. On MSRVTT-QA, it improves the baseline Acc. by 2.8%, with STA and TOI contributing individual improvements of 2.2% and 0.6%, respectively. For ActivityNet-QA, the improvement is 2.4%, with STA and TOI enhancing the baseline Acc. by 1.9% and 0.6%, respectively. We demonstrate our VideoQA-TA method not only affords the significantly of learning spatial and temporal information but also can achieve excellent scores in terms of standard metrics.

**Large language model.** We analyze the performance of our framework using different LLMs while predicting answer, as is shown in Table 3. The results indicate that Vicuna-13B achieves the best performance (66.7% in Top-1 accuracy), followed by Vicuna-7B (66.5% in Top-1 accuracy). Furthermore, Vicuna-13b also outperforms in Acc. (77.8%), surpassing Vicuna-7b (76.7%). Llama 2-7B and Mistral-7B perform slightly lower performance, with Llama 2-7B reaching 65.9% in Top-1

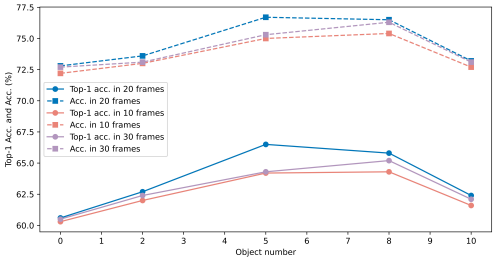



Figure 4: Top-1 acc. and Acc. with object numbers and video lengths.

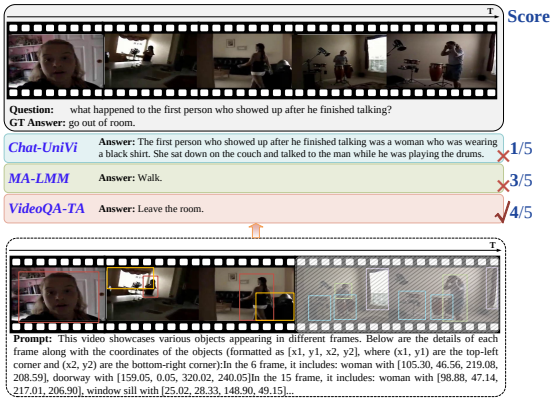



Figure 5: Visualization results on ActivityNet-QA dataset against MA-LMM and Chat-Univi.

| LLM | Model size | Top-1 acc. | Acc. |
|---|---|---|---|
| Vicuna - 7B | 7 B | 66.5 | 76.7 |
| Vicuna-13B | 13B | 66.7 | 77.8 |
| Llama 2-7B | 7 B | 65.9 | 74.9 |
| Mistral - 7B | 7 B | 65.1 | 73.0 |

Table 3: Top-1 acc. and Acc. with different large language model.

acc. and 74.9% in Acc., while Mistral-7B records 65.1% and 73.0%, respectively. Considering the trade-off between accuracy and computational cost, we primarily use Vicuna-7B for our experiments unless otherwise specified.

**Object numbers and video lengths.** We compare the results of object numbers and lengths of video on MSVD-QA dataset. From calculating the Top-1 acc. and Acc. shown in Figure 4, we notice that all methods, as well as the MA-LMM baseline, experience performance drops when video exceeds 20 frames (drop: 1.3%–6%) or contains more than 5 objects (drop: 0.7%–4.9%). Such issue persists even with the proposed method. However, VideoQA-TA mitigates it by filtering out redundant/irrelevant frames, improving performance on long VideoQA tasks.

### 5.3 Visualized Examples of Output

We evaluate our proposed model on MSVD-QA and ActivityNet-QA datasets against MA-LMM and Chat-Univi. Figure 5 shows VideoQA-TA higher accuracy in question answering. VideoQA-TA correctly answers 'Leave the room' for what happened to the first person who showed up after he finished talking. In contrast, Chat-UniVi says 'The first person who showed up after he finished talking was a woman who was wearing a black shirt. She sat down on the couch and talked to the man while he was playing the drums.', MA-LMM says 'Walk'. Chat-UniVi's response includes multiple actions that misalign with the temporal focus of the question, missing the key detail of 'the first person leaving the room.' MA-LMM's response of 'Walk' indicates an action but lacks clarity to confirm it as 'leaving the room'. Overall, our method outperforms in detail accuracy, effectively integrating temporal information, and provids more comprehensive understanding. More visualized examples of output, please refer to Appendix D.

## 6 Conclusion

This paper investigates the challenging problem of model temporal relations existed in the video, and introduces **VideoQA-TA**, a novel temporal-aware structure for multi-modal VideoQA. Specifically, *spatial-temporal attention* module is proposed to effectively aggregate the spatial and temporal features of videos by removing irrelevant information. Moreover, fine-grained levels alignments is constructed to align spatial, temporal, object and question features presented by the video with a novel *temporal object injection strategy*, further improving the accuracy of VideoQA. Experimental results on MSVD-QA, MSRVTT-QA and ActivityNet-QA datasets demonstrate that our method surpasses existing state-of-the-art methods, achieving impressive performance in VideoQA tasks.

## Limitations

We argue that the main limitation of our work lies in the generative answers, where the quality of the detected object on the performance of the model, even though our method is more efficient than other VideoQA models (section 5.1) for a detailed discussion. Despite the remarkable abilities of detection modules and multi-modal large language modal, it may still struggle to the limited types and diversity

of objects detected, which may cause the question answering system to fail in addressing certain questions or video content. LLMs still struggle to better understand the semantics of the 'bonding box' although we have made some improvements to the input to a certain extent (section 5.2). For example, the question 'What did a man keep on the tray?' is answered with 'shrimp in [26.92, 5.32, 254.74, 239.23]'.

## Acknowledgments

We would like to thank the anonymous reviewers, our meta-reviewer and senior area chairs for their thoughtful comments and support on this work. This work was supported by the National Key R&D Program of China (Grant No. 2022YFF0902701), National Natural Science Foundation of China (Grant Nos. U21A20468, 62372058, U22A2026), and BUPT Excellent Ph.D. Students Foundation (Grant No. CX20241018).

## References

Mobeen Ahmad, Geonwoo Park, Dongchan Park, and Sanguk Park. 2023. Mmtf: Multi-modal temporal fusion for commonsense video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4657–4662.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*.

Ziyi Bai, Ruiping Wang, and Xilin Chen. 2024a. Glance and focus: Memory prompting for multi-event video question answering. *Advances in Neural Information Processing Systems*, 36:34247–34259.

Ziyi Bai, Ruiping Wang, Difei Gao, and Xilin Chen. 2024b. Event graph guided compositional spatial-temporal reasoning for video question answering. *IEEE Transactions on Image Processing*, 33:1109–1121.

Keyuan Cheng, Gang Lin, Haoyang Fei, Lu Yu, Muhammad Asif Ali, Lijie Hu, Di Wang, et al. 2024a. Multi-hop question answering under temporal knowledge editing. In *Conference On Language Modeling*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14773–14783.

Jiale Han, Bo Cheng, Zhiguo Wan, and Wei Lu. 2023. Towards hard few-shot relation classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9476–9489.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Forty-first International Conference on Machine Learning*.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023. Open-set image tagging with multi-grained text supervision. *arXiv preprint arXiv:2310.15200*.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

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

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. 2024. How to configure good in-context sequence for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26710–26720.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023b. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023c. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878.

Zhaohe Liao, Jiangtong Li, Li Niu, and Liqing Zhang. 2024. Align and aggregate: Compositional reasoning with video alignment and answer aggregation for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13395–13404.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Hui Liu and Xiaojun Wan. 2024. Qavidcap: Enhancing video captioning through question answering techniques. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 155–164.

Nan Ma, Deyi Li, Wen He, Yue Deng, Jiahong Li, Yue Gao, Hong Bao, Huan Zhang, Xinkai Xu, Yuansheng Liu, et al. 2021. Future vehicles: interactive wheeled robots. *Science China Information Sciences*, 64:1–3.

Nan Ma, Zhixuan Wu, Yifan Feng, Cheng Wang, and Yue Gao. 2024. Multi-view time-series hypergraph neural network for action recognition. *IEEE Transactions on Image Processing*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245.

Jie Nie, Xin Wang, Runze Hou, Guohao Li, Hong Chen, and Wenwu Zhu. 2024. Dynamic spatio-temporal graph reasoning for videoqa with self-supervised event recognition. *IEEE Transactions on Image Processing*, 33:4145–4158.

SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. 2024. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 980–987.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024a. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. 2024b. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597.

Jinmeng Wu, Pengcheng Shu, Hanyu Hong, Lei Ma, Ying Zhu, and Lei Wang. 2024. Pre-trained bidirectional dynamic memory network for long video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5550–5557.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video.

In *International Conference on Machine Learning*, pages 38728–38748.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.

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

Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. 2024. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 852–861.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023a. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*.

Fuwei Zhang, Ruomei Wang, Songhua Xu, and Fan Zhou. 2021. Fusing temporally distributed multimodal semantic clues for video question answering. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jianxiong Zhou and Ying Wu. 2023. Temporal feature enhancement dilated convolution network for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6028–6037.

Bo Zou, Chao Yang, Yu Qiao, Chengbin Quan, and Youjian Zhao. 2024. Language-aware visual semantic distillation for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27113–27123.

# Appendix

Our appendix consists of Algorithm of TOI (Section A), Datasets (Section B), Prompt templates details (Section C) and Visual analysis of experimental results (Section D).

## A Algorithm of TOI

Algorithm 1 details how TOI to align object-level and frame-level features. We first transform the input query $q_t$ into query, key, and value vectors using the weight matrices $W_Q$, $W_K$, and $W_V$. Then, the query vector $Q_q$ is updated through a self-attention mechanism to obtain $Q'$. After that, spatial features $F_{spa}$ and temporal features $F_{tem}$ are combined with $E$ and processed through cross-attention to align the object-level features with the appearance features. Third, the query vector is aligned with the feature vector through the cross-attention to achieve the final feature representation $F_i$.

**Algorithm 1** TOI algorithm

**Input**: Weight matrix $W_Q$, $W_K$, $W_V$, respectively. Spatial features $F_{spa}$, temporal features $F_{tem}$, question $Q$. Object label $L_t$ and Bounding box $B_t$. $[\cdot \,||\, \cdot]$ denotes the concatenation operator.
**Output**: Aligned features $F_i$
// Initialization
// Construct a temporally sequenced
$E = Concat[Prompt, L_t, B_t]$
**for** $i = 1 \ldots I$ **do**
  **for** $t = 1 \ldots T$ **do**
// Convert input queries into query, key, and value vectors
    $Q_q \leftarrow q_t \, W_Q$
    $K_q \leftarrow [q_1 || \ldots || q_t] W_K$
    $V_q \leftarrow [q_1 || \ldots || q_t] W_V$
// Obtain the updated $Q'$
    $Q' \leftarrow \sigma(\frac{Q_q K_q^T}{\sqrt{d_k}} V_q)$
  **end for**
// Alignment
  $F_{1_i} \leftarrow MLP([F_{spa} || F_{tem} || E])$
  $F_{2_i} \leftarrow MLP([Q' || E])$
  $F_i \leftarrow \sigma(\frac{P_Q(F_{2_i}) P_K(F_{1_i})^T}{\sqrt{d_k}}) P_V(F_{1_i})$
**end for**
**return** $F_i$

| |
|---|
| Let's think step by step. Therefore, the answer (one sentence) is: |

Table 4: Prompt for Zero-shot Chain-of-Thought

| |
|---|
| This video features various objects appearing in different frames. Here are the details: The $\{object\_name\}$ appears in the following frames: $\{frame\_id\}$, $\{frame\_id\}$ ... |

Table 5: Prompt for ($L_t$, $t$)

| |
|---|
| This video features different frames appearing in various objects. Here are the details: In the $\{frame\_id\}$, it includes $\{object\_name\}$, $\{object\_name\}$... |

Table 6: Prompt for ($t$, $L_t$)

| |
|---|
| This video showcases different frames appearing in various objects. Below are the details of each frame along with the coordinates of the objects (formatted as $[x_1, y_1, x_2, y_2]$, where $(x_1, y_1)$ are the top-left corner and $(x_2, y_2)$ are the bottom-right corner): The $\{object\_name\}$ appears in the following frames: $\{frame\_id\}$ with $\{object\_bonding\_box\}$, $\{frame\_id\}$ with $\{object\_bonding\_box\}$... |

Table 7: Prompt for ($L_t$, $B_t$, $t$)

## B Datasets

**MSVD-QA**[5](Xu et al., 2017) is a specialized VQA corpus derived from the Microsoft Research Video Description (MSVD) dataset, which consists of over 120,000 descriptive sentences for more than 2,000 video clips. MSVD-QA extends this by generating approximately 50,500 Question-Answer (QA) pairs from these descriptions, covering 1,970 video clips, with the associated videos available in the foundational MSVD dataset.

**MSRVTT-QA**[6](Xu et al., 2017) serves as a prominent benchmark for VQA, which is built upon the foundation of the MSRVTT, a collection encompassing 10,000 videos. This extensive dataset contains 243,000 questions and offers 1.5 million potential answers.

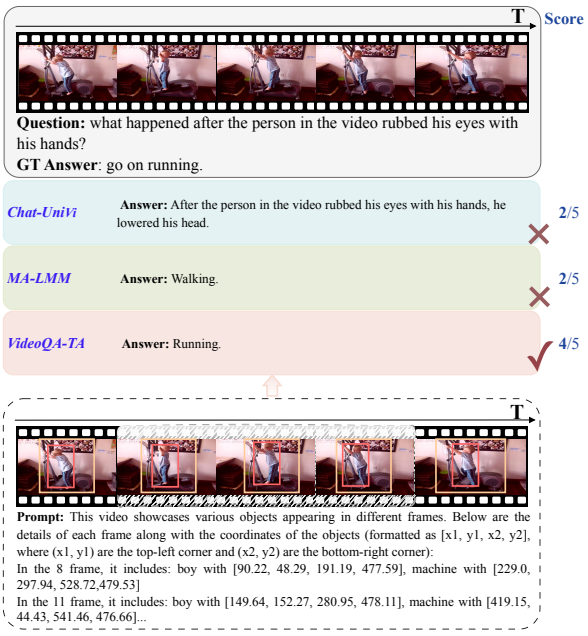
**ActivityNet-QA**[7](Yu et al., 2019) comprises 58,000 QA pairs, each annotated by humans, across 5,800 videos sourced from the well-known ActivityNet dataset. It serves as a standard for evaluating the capabilities of VQA models in long-term spatial-temporal reasoning.

## C Prompt Templates Details

In this section, we use several input combinations to obtain detected objects information and present a detailed analysis of these results. Given object label $L_t$, bounding box $B_t$, and frame $t$:

- Zero-shot Chain-of-Thought (Kojima et al., 2022) $\rightarrow$ $E_1 = Concat[Prompt_1]$: Think step by step, LLM can focus on each frame information.
- $(L_t, t) \rightarrow E_2 = Concat[Prompt_2, L_t]$: Framework uses object labels as input for obtaining detected objects information. Providing the presence of each object within frames is advantageous, as it allows the LLM to gather additional context information for answering.
- $(t, L_t) \rightarrow E_3 = Concat[Prompt_3, L_t]$: Providing the presence of each frame within objects is advantageous, as it allows the LLM to gather extra context and time-series information for answering.
- $(L_t, B_t, t) \rightarrow E_4 = Concat[Prompt_4, L_t, B_t]$: Framework uses object labels and bounding boxes as input to obtain detected objects information. Providing the presence of each object within frames along with corresponding bounding box is advantageous, as it allows the LLM to gather extra context and location information for answering.
- $(t, L_t, B_t) \rightarrow E = Concat[Prompt, L_t, B_t]$: Providing the presence of each frame within objects and corresponding bounding box is advantageous, as it allows the LLM to gather extra context, time-series and location information for answering.

[5]https://github.com/xudejing/video-question-answering
[6]https://github.com/xudejing/video-question-answering
[7]https://github.com/MILVLG/activitynet-qa




(a) Visualization results on MSVD-QA dataset against MA-LMM and Chat-Univi.


(b) Visualization results on Activity-QA dataset against MA-LMM and Chat-Univi.

Figure 6: Visualization results comparison on MSVD-QA and ActivityNet-QA datasets.

We provide detailed prompts for Zero-shot Chain-of-Thought prompt in Table 4, $(L_t, t)$ prompt in Table 5, $(t, L_t)$ in Table 6, and $(L_t, B_t, t)$ prompt in Table 7. To implement Self-Consistency, we run model with different prompts for 3 times, and calculate the average accuracy.

## D Visual Analysis of Experimental Results

In this section, we visualize our proposed model VideoQA-TA against MA-LMM and Chat-UniVi on MSVD-QA and ActivityNet-QA datasets. As is shown in Figure 6(a), when asked who the man is getting married to, VideoQA-TA correctly identifies 'Woman', while Chat-UniVi says 'The video does not provide any information about the identity of the man getting married. However, it is clear that he is the center of attention as he is being adorned with a traditional Indian headdress by a group of people', MA-LMM says 'someone'. Figure 6(b) shows VideoQA-TA answers 'Running' when asked what happened after the person in the video rubbed his eyes with his hands, while Chat-UniVi says 'After the person in the video rubbed his eyes with his hands, he lowered his head.', MA-LMM says 'Walking'. Chat-UniVi struggles with understanding complex temporal information, often overlooking significant dynamic changes in the video. Although MA-LMM can capture actions and events to some extent, it frequently fails to provide accurate answers when detailed temporal reasoning is required.

Additionally, we visualize the impact of each component on VideoQA-TA performance. As mentioned in Section 5.2, using both TOI and STA, improving scores in some extent. For example, Figure 7 shows that VideoQA-TA correctly answers 'Saying to someone' when asked what an old woman is doing. In contrast, while without TOI says 'Walking and saying', without STA says 'Saying'. The former appears to diminish the model's ability to focus on the most relevant actions, likely due to reduced video information acquisition capabilities. TOI enhances the contextual understanding necessary for a more precise and contextually appropriate response.


Figure 7: Visualization results on MSRVTT-QA dataset against without TOI and STA. "*w/o*" means without corresponding module from the VideoQA-TA.