This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Perceive, Query & Reason: Enhancing Video QA with Question-Guided Temporal Queries

Roberto Amoroso[3,*] Gengyuan Zhang[1,2,*] Rajat Koner[1]
Lorenzo Baraldi[3] Rita Cucchiara[3,4] Volker Tresp[1,2]
[1]LMU Munich [2]MCML [3]University of Modena and Reggio Emilia [4]IIT-CNR
roberto.amoroso@unimore.it zhang@dbs.ifi.lmu.de

Abstract

Video Question Answering (Video QA) is a challenging video understanding task that requires models to comprehend entire videos, identify the most relevant information based on contextual cues from a given question, and reason accurately to provide answers. Recent advancements in Multimodal Large Language Models (MLLMs) have transformed video QA by leveraging their exceptional commonsense reasoning capabilities. This progress is largely driven by the effective alignment between visual data and the language space of MLLMs. However, for video QA, an additional space-time alignment poses a considerable challenge for extracting question-relevant information across frames. In this work, we investigate diverse temporal modeling techniques to integrate with MLLMs, aiming to achieve question-guided temporal modeling that leverages pre-trained visual and textual alignment in MLLMs. We propose T-Former, a novel temporal modeling method that creates a question-guided temporal bridge between frame-wise visual perception and the reasoning capabilities of LLMs. Our evaluation across multiple video QA benchmarks demonstrates that T-Former competes favorably with existing temporal modeling approaches and aligns with recent advancements in video QA.

1. Introduction

In recent years, video understanding tasks have gained increasing attention, driven by advancements in multimodal learning and vision-language models [4, 23, 41, 50]. Videos are ubiquitous on the Internet and their inherently dynamic and complex nature poses a significant challenge for video understanding [27, 30, 49, 53].

Video Question Answering (Video QA) emerges as a particularly fundamental and challenging task within this domain, requiring models to not only perceive videos for recognition-level tasks like classification or segmentation but also to comprehend and rationalize the video content in response to specific questions. The challenge is twofold: models must process the temporal and dynamic nature of videos with their sparse and heterogeneous event distributions, while simultaneously developing a deep understanding that connects these events through commonsense and temporal reasoning. This necessitates a blend of model abilities including visual perception, question contextualization, and answer reasoning.

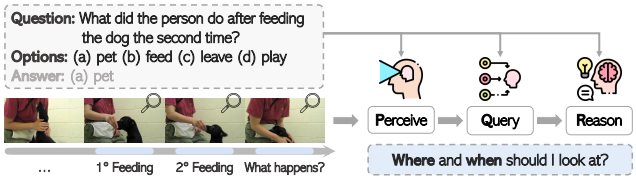


Figure 1. Adapting LLMs' visual reasoning capabilities to video QA requires extracting the most relevant video features based on the input question. Our approach addresses this challenge by extracting question-guided temporal features from dynamic frame sequences, enabling accurate and context-aware reasoning.

The advent of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) has marked significant progress in visual reasoning, including image QA [5, 19, 24] and video QA [46, 50] tasks. MLLMs exhibit strong general knowledge and reasoning capabilities, excelling in interpreting visual content for video QA applications. A common paradigm in MLLMs involves integrating a visual encoder for embedding video frames and a projection network that compresses visual features and aligns them with LLMs. Conventional compression strategies, such as sparse sampling and pooling-based methods [11, 15, 26, 54], face limitations when dealing with lengthy and dynamic videos. Recent studies [1, 50, 51, 55] have achieved more promising results by adopting attention-based models like Q-Former [7, 19]. The Q-Former extracts prominent visual features from the visual encoder into a small set

*Equal contribution.

of learnable queries, which are then aligned with the textual embeddings from a pre-trained LLM via a linear layer. This paradigm has proven effective across various vision-language tasks [47,50,51,55], demonstrating that visual features can be compressed without compromising the overall representation while also reducing computational costs.

However, directly adapting image-based Q-Former approaches to video data often results in suboptimal performance due to the spatio-temporal complexities and varying information density inherent in videos [46]. We hypothesize that pre-trained image-level models lack the necessary spatio-temporal inductive bias for video understanding. This limitation hinders the seamless integration of LLMs with video content, especially in scenarios involving complex dynamics and diverse information distributions.

In this work, we introduce T-Former, a question-guided temporal querying Transformer designed to efficiently sample and learn video-specific features tailored to a given question. Unlike existing methods, T-Former leverages knowledge from image-based pre-training without requiring additional video pre-training. With T-Former, we propose a new LLM-based model, PQR, based on three core principles: ***P**erceive* to extract framewise question-guided spatial features from input videos, ***Q**uery* to extract a fixed-size set of question-guided temporal visual features from the whole sequence, and ***R**eason* to integrate these extracted features with LLMs for generating answers, as illustrated in Fig. 1. T-Former employs an efficient temporal sampling strategy that incorporates an inductive bias for video queries by initializing learnable input queries with discrete key video event features. This design bridges the gap between pre-trained image-level knowledge and video-specific representations while focusing on relevant question-guided information. By doing so, T-Former ensures a seamless transition from generic to question-tailored video features.

Our evaluations demonstrate that T-Former outperforms alternative temporal modeling techniques, achieving peak performance at a sampling ratio equivalent to four frames, thereby effectively harnessing the potential of LLMs. Notably, PQR sets a new state-of-the-art benchmark in video QA, surpassing existing methods.

To sum up, the contributions of this paper are as follows:

- We propose T-Former as a simple yet effective temporal modeling method that can effectively capture question-guided information from video frames;
- We show that T-Former achieves superior performance on video QA while maintaining a good balance of computation complexity;
- We present PQR, a novel LLM-based model achieving competitive performance across multiple video QA benchmarks, thereby redefining the state-of-the-art in this field.

2. Related Works

2.1. Video Question Answering

Video Question Answering is a fundamental task of video-language understanding [53]. Early approaches to video language modeling primarily relied on sequential neural networks, such as Long Short-Term Memory networks [14], 3D Convolutional Neural Networks [21], and Graph Neural Networks [28, 39]. These models were effective in capturing temporal and spatial information but faced limitations in handling complex video data and contextual reasoning.

With the advent of Transformer-based visual encoders and novel paradigms for video-language pre-training [2, 3, 15, 29, 35, 36, 42, 43, 46], encoder-decoder models have achieved significant advancements in the video QA task. These models benefit from the Transformers' ability to capture long-range dependencies and contextual information, leading to improved performance in understanding and reasoning about video content. The development of video QA has been driven by diverse benchmark datasets spanning multiple domains, including movies and TV shows [17, 18, 32,40], as well as wild, online videos [10,37,38,40,48]. Notably, datasets like NeXT-QA [38] and STAR [37] have introduced more challenging aspects of video understanding, emphasizing temporal reasoning and causal relationships, which align closely with our work's focus on question-guided temporal modeling.

2.2. Temporal Modeling for Video QA

Temporal feature extraction plays a crucial role in video understanding due to the dense, dynamic, and redundant content of video data. [3, 15, 16] propose sparse sampling based on dataset biases. While these methods can be effective, they often struggle with complex videos containing multiple events or subtle temporal dependencies, limiting their applicability in real-world scenarios. Several works also adopt a "first temporal, then spatial" strategy to enhance video understanding by training a temporal selection module. TranSTR [22] uses Adaptive Temporal Rationalization to select critical frames, whereas MIST [9] leverages a temporal attention layer to aggregate frame-wise features. These approaches demonstrate the importance of selective temporal processing but often involve complex architectural designs.

Recent advances in LLMs have opened new possibilities for video QA. However, integrating video inputs with LLMs remains challenging, particularly due to the quadratic computational complexity associated with processing long input sequences. In this work, we propose T-Former, a question-guided temporal querying Transformer that, achieves competitive performance while mitigating the computational overhead.

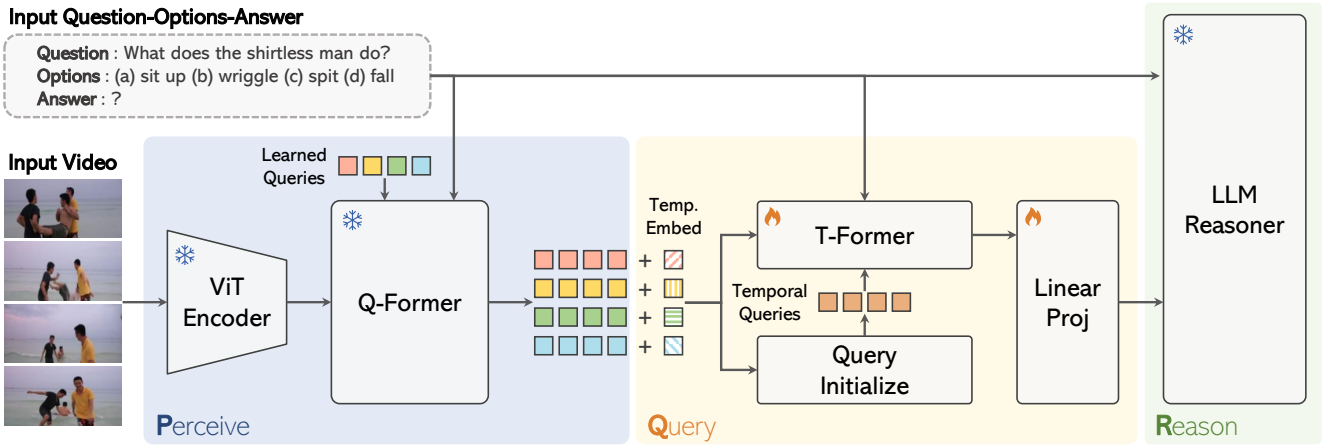


Figure 2. Our model consists of three primary stages: ***Perceive***, ***Query***, and ***Reason***. The *Perceive* stage employs a ViT encoder and a Q-Former to extract spatial visual features from each frame independently, conditioned on the input question. The *Query* stage introduces T-Former, our question-guided temporal querying Transformer that captures the most relevant temporal information conditioned on both the question and the visual context. Finally, in the *Reason* stage, the condensed visual-temporal features are integrated with the question and answer options before being fed into a frozen LLM as a reasoning agent to rationalize and answer the question.

2.3. Video QA with LLMs

Pre-trained LLMs have drawn significant attention in recent years owing to their emergent capabilities of instruction following [6,31,52] and human-like commonsense reasoning [13, 25, 44]. In the field of visual understanding, Multimodal Large Language Models (MLLMs) attempt to merge the capabilities of LLMs with visual data to address complex vision tasks, such as video QA [19]. These models have demonstrated superior visual reasoning abilities compared to traditional models, establishing a new baseline in visual comprehension. The application of LLMs to video QA presents both promising opportunities and unique challenges. The inherently multi-event nature of video content introduces complexities not present in static images. Videos require models to navigate and reason across temporal sequences, a task that demands more than just the integration of visual and textual modalities. To address these challenges, the Q-Former emerges as a powerful component to compress spatial visual features [7, 19, 50], thus facilitating the alignment of these features with the reasoning capabilities of LLMs. By adopting and further developing the Q-Former architecture, our work follows this direction of utilizing an LLM as a powerful agent for reasoning on video data and explores how to extract temporal visual features from videos and align them with MLLMs.

3. Proposed Method

In this section, we introduce PQR, our LLM-based video QA model, alongside T-Former, a novel question-guided temporal querying Transformer. Our model, illustrated in Fig. 2, consists of three main modules: (1) ***Perceive***: framewise visual encoding of input videos conditioned on given questions; (2) ***Query***: T-Former for selecting question-relevant temporal information across frames; and (3) ***Reason***: an adapted LLM as a reasoning agent for question answering.

We start by formulating the video QA task. Given a video v as a sequence of n image frames $I^{1:n}$, where I^i represents the i-th frame of the video, and a question q, a model $f(\cdot)$ assigns a probability to an answer a from the answer space $\mathcal{A}$. This can be formulated as:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} f(a|v, q, \mathcal{A}), \tag{1}$$

where $\hat{a}$ is the predicted answer given the video and questions. The answer space $\mathcal{A}$ can be large for open-end QA or preset for multiple-choice QA.

3.1. Perceive: Visual Encoding

State-of-the-art visual encoders such as CLIP [29] and BLIP-2 [19] have demonstrated remarkable zero-shot capabilities in generating visual features, making it a common practice [46] to leverage these powerful features for video QA tasks. Despite their success, these models are restricted to extracting vision-only features that ignore other modalities, like textual questions in our task. This leads to question-agnostic visual feature extraction and a lack of contextual relevance to complex questions.

To address this limitation, our PQR employs a question-guided visual feature extraction mechanism. Specifically, we adopt the instruction-aware Q-Former proposed by [7] to extract question-guided features from each frame. In more detail, given a sequence $I^{1:n}$, we encode each frame *independently* to get a sequence $\mathbf{e}_f^{1:n}$ of framewise visual tokens:

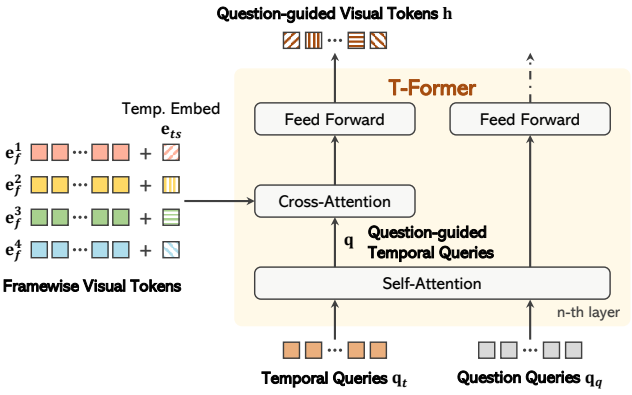


Figure 3. **Overview of T-Former,** a Temporal Querying Transformer. T-Former adopts 1) a self-attention layer between temporal queries and question queries and 2) cross-attention layers between query tokens and the full-length sequence of visual tokens.

$$\mathbf{e}_f^i = \text{VisualEncoder}(I^i, q, \mathcal{A}), \forall i \in \{1, 2, \ldots, n\}, \quad (2)$$

where $\mathbf{e}_f^i$ represents the visual features of i-th frame, and VisualEncoder refers to both the frame encoder and Q-Former for simplicity. This step aims to *perceive* and extract spatial visual information from raw frames.

3.2. Query: T-Former

Handling a full sequence of framewise visual tokens $\mathbf{e}_f^{i:n}$ leads to temporal information redundancy across frames and is also computationally demanding due to the complexity of LLMs. To mitigate these challenges, we propose T-Former, a question-guided ***t***emporal querying trans***former***, designed to select the most relevant visual features across time based on the input question, as presented in Fig. 3.

T-Former integrates three distinct types of input tokens: 1) the concatenation of framewise visual tokens $\mathbf{e}_f^{1:n}$ from the *perceive* step, 2) temporal queries $\mathbf{q}_t$ of a fixed length to *query* temporally relevant information from the whole framewise visual token sequence, and 3) question queries $\mathbf{q}_q$ to condition the output tokens on the questions. T-Former first contextualizes query tokens with self-attention between $\mathbf{q}_t$ and $\mathbf{q}_q$; and then performs cross-attention between $\mathbf{q}_t$ and $\mathbf{e}_f$ to select question-relevant visual information over time. We elucidate each module as follows.

Temporal Query Initialization. The core of our approach lies in the initialization of a small number of fixed-length temporal queries, $\mathbf{q}_t$, which are specific to the input video and the question. Existing methods [7, 19, 50, 55] typically employ learnable queries with random initialization to compress features. However, these randomly initialized static queries require extensive video-based pre-training, as they lack any inherent alignment with the spatio-temporal dynamics of video events. Moreover, even after pre-training, these static queries remain identical across all videos, failing to capture video-specific contextual information. To overcome these limitations, we propose a novel strategy that replaces static, pre-trained queries with dynamic queries sampled directly from the input video features. By grounding the query tokens in the unique characteristics of the input video, our approach introduces a more effective inductive bias for T-Former that aligns with Q-Former's image understanding capabilities. This dynamic grounding not only enables a deeper understanding of the video but also facilitates more precise responses to video-related questions.

We investigate four distinct sampling strategies for temporal query initialization: uniform sampling, random sampling, k-means sampling, and k-medoids sampling:

$$\mathbf{q}_t = \text{sampler}(\mathbf{e}_f^1, \mathbf{e}_f^2, ..., \mathbf{e}_f^n) \quad (3)$$

The size of temporal queries $\mathbf{q}_t$ is considered a crucial model hyperparameter, governing two essential aspects. Firstly, it defines the sampling ratio of T-Former, thereby controlling how much the input sequence is compressed. Secondly, it influences the length of visual tokens fed into LLMs. Hence, a higher number of temporal queries increases computational overheads within LLMs. In Section 4.5, we investigate the effects of various sampling methods and sampling ratios.

Question-guided Temporal Query Tokens. To condition the temporal queries $\mathbf{q}_t$ based on the questions, we concatenate $\mathbf{q}_t$ with the question queries $\mathbf{q}_q$. These question queries are generated through the tokenization process applied to the question q and the answer options $\mathcal{A}$. To facilitate interaction between different modalities, we employ self-attention layers to contextualize both $\mathbf{q}_t$ and $\mathbf{q}_q$. This process is formulated as follows:

$$\mathbf{Q} = \mathbf{K} = \mathbf{V} = \phi(\text{concat}(\mathbf{q}_t, \mathbf{q}_q)), \quad (4)$$

$$\mathbf{q} = \text{self-attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \quad (5)$$

where ϕ represents the linear projection function of the attention [34] mechanism and $\mathbf{q}$ denotes the resulting question-guided temporal query tokens.

Query-Frame Cross-attention. The sampled temporal queries still lack spatio-temporal information from videos. Therefore, revisiting the video to enhance the queries with more detailed and fine-grained video-specific information becomes imperative.

To this aim, given the contextualized queries $\mathbf{q}$, we employ cross-attention between $\mathbf{q}$ and the entire sequence of video features. The process also incorporates learnable timestamp embeddings $\mathbf{e}_{ts}$ to capture temporal dependencies, as formulated in the following:

$$\hat{\mathbf{e}}_f^i = \mathbf{e}_f^i + \mathbf{e}_{ts}, \quad (6)$$

Sampling Strategy	#Input Frames	NExT-QA			
		Tem.	Cau.	Des.	Avg.
Single Frame	1	67.7	72.3	79.0	71.9
Concatenate	4	66.1	72.3	80.6	71.6
Concatenate	16	69.6	74.0	81.1	73.2
Mean-pooling	4	64.5	69.6	77.2	69.1
Mean-pooling	16	64.6	70.5	79.6	70.0
Spatio-Temporal	4	64.5	70.7	77.5	69.8
Spatio-Temporal	16	67.7	72.6	79.5	72.1
T-Former (Ours)	16	**72.8**	**76.9**	**84.7**	**76.6**

Table 1. **Comparing T-Former with other temporal modeling methods.** T-Former outperforms other sampling methods.

$$\mathbf{Q} = \psi_q(\mathbf{q}), \quad \mathbf{K} = \mathbf{V} = \psi_{kv}(\text{concat}(\hat{\mathbf{e}}_f^1, ..., \hat{\mathbf{e}}_f^n)), \quad (7)$$

$$\mathbf{h} = \text{cross-attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \quad (8)$$

Finally, we pass the final T-Former hidden state $\mathbf{h}$ through a feed-forward layer to derive a representation $\mathbf{t}_v$ that integrates video content, question context, and required information for a precise answer. Importantly, our cross-attention mechanism between the temporal queries and the full-length sequence of video tokens enables the T-Former to perform question-guided spatio-temporal feature extraction. This mechanism enriches temporal query tokens by integrating additional relevant information from the full sequence of video tokens, compensating for the partial context captured during initial sampling and ensuring a comprehensive, question-specific understanding of the video.

3.3. Reason: LLMs as reasoning agents

The reasoning component of our framework leverages an LLM to reason and extract meaningful insights from the rich visual-temporal representations produced by our T-Former. Aggregated visual features from T-Former, question tokens, and answer option tokens from the LLM tokenizer are fed as inputs to the LLM reasoner. We incorporate a linear projection layer to reformat the visual features and ensure alignment with the LLM's feature space:

$$\mathbf{h}_{vq\mathcal{A}} = \text{LLM}(\text{linear-proj}(\mathbf{t}_v), q, \mathcal{A}). \quad (9)$$

The LLM's hidden state $\mathbf{h}_{vqA}$ represents a unified embedding that captures the complex interactions between the contextualized representations of video, question, and answer space $(v, q, \mathcal{A})$. For multiple-choice video QA tasks, we append a linear projection layer, as a common practice, to generate logits for cross-entropy loss computation.

This enhanced reasoning module enables sophisticated multi-modal understanding, leveraging the pre-trained knowledge of the LLM while maintaining efficiency through our selective temporal feature extraction approach.

Sampling Strategy	NExT-QA			
	Tem.	Cau.	Des.	Avg.
Uniform	70.8	74.9	83.8	75.0
Random	71.7	75.0	84.6	75.4
k-means	71.9	76.7	83.4	76.2
k-medoids	**72.8**	**76.9**	**84.7**	**76.7**

Table 2. **Ablation studies of different sampling methods for temporal query initialization**. The number of input frames is 16. The clustering-based methods outperform other techniques.

4. Experiments

4.1. Task and Datasets

We evaluate PQR in a multiple-choice video QA setting, where each question is accompanied by a fixed number of answer choices. Our primary evaluation is conducted on the NExT-QA dataset [38], which serves as our development benchmark for exploring various architectural decisions. To ensure comprehensive validation, we extend our investigation to three additional competitive video QA benchmarks: NExT-QA, STAR [37], and How2QA [20], as well as one video event prediction benchmark: VLEP [18]. Notably, **PQR is trained exclusively on target datasets** without requiring any video-based pre-training phase.

We follow established evaluation protocols [7, 50] for each benchmark to ensure comparability with prior works. Further details about the experimental setup are provided in the supplementary material.

4.2. Implementation Details

Our framework adopts the ViT-G/14 from [8] and the instruction-aware Q-Former from [7] as the pre-trained visual encoder for frame processing. The reasoning component in PQR is powered by a frozen Vicuna-7B model [52], an instruction-tuned LLM built on top of the LLaMA architecture [33]. Within PQR, we integrate two layers of our T-Former, which is based on a BERT-structured Transformer. To ensure a smooth integration between T-Former 's visual features and the LLM's feature space, we utilize a linear projection layer. Additionally, a trainable linear layer is appended to the LLM for fine-tuning in the multi-choice video QA setting. During training, both the visual encoder and the LLM remain frozen, and only T-Former and the linear projection layers are trained.

LLM Prompts. For all tasks, we adhere to the protocol of [7, 19] and adopt a canonical context template: "*Question:* [<*Question*>]. *Options:* [<*Option Lists*>]. *Answer:*" used for textual inputs of Q-Former, T-Former, and the LLM reasoner. For the future event prediction task on VLEP, we use the same question template, "*Which event is more likely to happen right after?*", for all samples.

Temporal Queries Initialization	NExT-QA			
	Tem.	Cau.	Des.	Avg.
Learnable	63.0	65.2	73.2	65.7
Sampled	**72.8**	**76.9**	**84.7**	**76.7**

Table 3. **Effect of using learnable temporal queries instead of sampled temporal queries**. We found that sampled temporal queries achieve better performance.

#Hidden Layers	NExT-QA			
	Tem.	Cau.	Des.	Avg.
12	67.1	72.1	80.1	71.7
8	69.1	74.8	83.4	74.3
4	71.1	75.4	83.9	75.3
2	**72.8**	**76.9**	**84.7**	**76.7**
1	70.1	75.3	84.3	75.1

Table 4. **Effect of the number of hidden layers**. Overall, T-Former with two hidden layers has achieved the best performance.

Training details. Our PQR is trained on a single NVIDIA A6000 GPU with 48GB of memory. We employ the AdamW optimizer with a linear warmup and cosine annealing scheduler. Detailed batch sizes and learning rates for each dataset are provided in the supplementary material.

4.3. Evaluation Metrics

We report the standard answer accuracy metric for the multiple-choice video QA tasks. Each dataset has a different number of candidate answer options: NExT-QA provides five options per question, STAR and How2QA provide four options, and VLEP offers two options. Additionally, NExT-QA categorizes questions into three types: temporal, causal, and descriptive, while STAR delineates four types: interaction, sequence, prediction, and feasibility.

4.4. Comparison to Other Temporal Modeling

We compare our T-Former with several temporal modeling baselines on the NExT-QA dataset, including:

- *Single Frame Sampling*: A single frame is randomly sampled from the whole video sequence.
- *Frame Concatenation*: Visual tokens from each frame are concatenated before being fed into the LLM.
- *Mean-Pooling*: Framewise outputs from the visual encoder are merged via mean pooling.
- *Spatio-Temporal Transformers*: The ViT features from all frames are concatenated in sequence and input into the Q-Former, which jointly extracts spatio-temporal features before passing them to the LLM.

Our results, summarized in Tab. 1, demonstrate that T-Former outperforms these baselines by a significant margin, with improvements ranging from 3.4% to 7.5%. Interestingly, simple baseline methods like single-frame sampling achieve decent performance, suggesting that the dataset may still exhibit temporal bias. However, T-Former consistently excels, particularly in temporal and causal reasoning, highlighting its superior temporal modeling capabilities.

#Attn. Heads	NExT-QA			
	Tem.	Cau.	Des.	Avg.
2	71.5	75.8	82.6	75.5
4	71.8	75.9	83.7	75.8
8	71.9	76.5	84.1	76.2
12	**72.8**	**76.9**	**84.7**	**76.7**
16	72.7	76.5	84.0	76.4

Table 5. **Effect of the number of attention heads**. Increasing the number of attention heads improves the model's performance.

Sampling Ratio	NExT-QA			
	Tem.	Cau.	Des.	Avg.
$8 \rightarrow 1$	70.9	75.9	82.4	74.8
$16 \rightarrow 1$	71.0	75.3	81.9	74.9
$32 \rightarrow 1$	71.4	75.4	83.2	75.3
$64 \rightarrow 1$	71.5	75.8	82.4	75.4
$128 \rightarrow 1$	71.4	75.3	83.3	75.3
$8 \rightarrow 4$	71.9	76.8	**84.8**	76.4
$16 \rightarrow 4$	72.8	76.9	84.7	**76.6**
$32 \rightarrow 4$	73.0	76.1	84.6	76.4
$64 \rightarrow 4$	72.4	**77.0**	84.3	**76.6**
$128 \rightarrow 4$	**73.2**	76.6	84.4	**76.6**
$8 \rightarrow 8$	71.6	75.8	83.5	75.7
$16 \rightarrow 8$	72.1	**77.0**	84.3	76.5
$32 \rightarrow 8$	72.8	76.7	84.2	**76.6**
$64 \rightarrow 8$	72.3	76.2	84.7	76.3

Table 6. **Impact of sampling ratios**. We conduct a comprehensive analysis to determine the optimal sampling ratio, defined as "#input frames $\rightarrow$ #equivalent output frames".

4.5. Exploring Design Choices

We perform a comprehensive analysis of different design choices for T-Former, examining the impact of sampling methods, Transformer settings, and sampling ratios.

Sampling methods. We compare four methods for initializing queries $\mathbf{q}_t$: uniform sampling, random sampling, k-means sampling, and k-medoids sampling. In our experiments, we fix the input frames at 16. The clustering-based methods, *i.e.*, k-means and k-medoids, outperform other techniques, as shown in Tab. 2. This aligns with our intuition that clustering selects more diverse and crucial frames, providing better initialization.

We also conducted an experiment involving learnable temporal queries akin to the learnable queries in Q-Former. The results, shown in Tab. 3, reveal that our sampling-based approach consistently outperforms learnable temporal queries by a significant margin. This suggests that the initial sampling strategy for temporal queries contributes substantially to the overall performance.

T-Former	Question-Guided	Temp. Embed.	NExT-QA				STAR					How2QA	VLEP
			Tem.	Cau.	Des.	Avg.	Int.	Seq.	Pre.	Fea.	Avg.		
✗	✗	✗	70.3	75.8	83.4	75.4	61.0	66.3	70.3	69.8	65.1	81.2	68.6
✓	✗	✗	72.2	76.0	84.6	76.1	60.8	67.2	71.6	70.4	65.4	84.4	69.0
✓	✓	✗	71.5	77.0	83.9	76.3	60.8	66.8	71.2	70.2	65.4	85.6	69.3
✓	✗	✓	**73.5**	75.1	**85.5**	76.2	63.0	68.2	70.2	72.6	66.6	84.5	69.2
✓	✓	✓	72.8	**77.0**	84.7	**76.7**	**62.8**	**69.6**	**72.8**	**70.0**	**67.6**	**85.9**	**69.6**

Table 7. **Ablation studies of each module of PQR.** Ablating different modules including T-Former, question-guided visual features, and temporal embeddings. The impact on the model's performance demonstrates each module's efficacy.

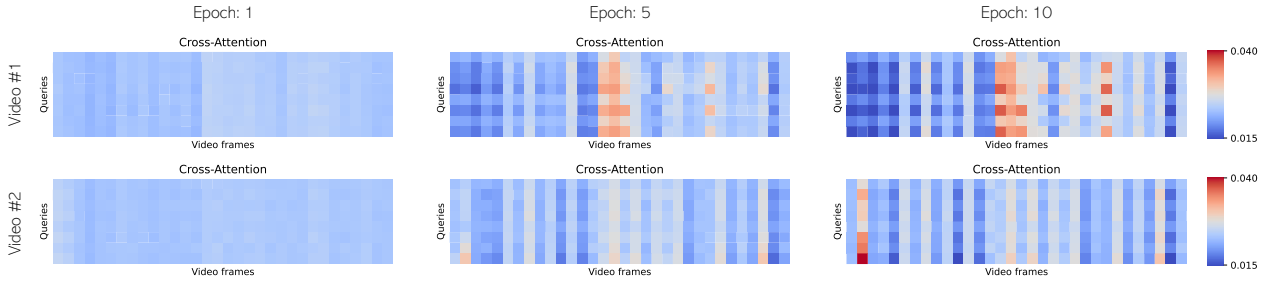


Figure 4. For different video samples (on each row), we visualize the attention map between question-guided temporal queries $\mathbf{q}$ and framewise visual tokens $\mathbf{e}_f^{1:n}$ in the last layer of T-Former during epochs 1, 5, and 10.

Investigating T-Former settings. We investigate the impact of varying our T-Former architecture settings, including the number of hidden layers and attention heads.

In Tab. 4, we show that T-Former with two hidden layers represents the most effective solution. We speculate that since Q-Former already captures intricate and abstract visual features, T-Former can easily grasp the temporal modeling of video frames with its lightweight structure.

As shown in Tab. 5, increasing the number of attention heads improves performance, with 12 heads yielding the best results. We believe that having more attention heads helps to encode diverse relationships between video frames, a critical capability for comprehensive temporal reasoning.

Impact of video-frame sampling ratio. We analyze how the sampling ratio, defined as "number of input frames $\rightarrow$ numbers of equivalent output frames", influences the performance of PQR. A higher sampling ratio compresses the input frame sequence into a smaller number of output tokens, referred to as the number of equivalent output frames. Conversely, a lower sampling ratio produces more fine-grained representations, but it also leads to longer input sequences for LLMs.

As shown in Tab. 6, for a fixed sampling ratio, such as $8 \rightarrow 1$ and $32 \rightarrow 4$, larger output sizes consistently yield better performance by accommodating more information. Conversely, compressing entire video frames into a small number of tokens while keeping the input frame size constant, as in $16 \rightarrow 1$ and $16 \rightarrow 4$, can hinder performance. Moreover, when the output frame size is fixed, as in $16 \rightarrow 4$, $32 \rightarrow 4$, and $64 \rightarrow 4$, incorporating more input frames enhances the performance by introducing additional information. However, there is an upper limit beyond which adding more frames no longer enhances performance.

Our findings confirm that increasing the number of input frames improves the performance, but only up to a certain limit. While additional frames provide more question-aware information, they may also introduce excessive redundancy. Additionally, the number of output frames is critical for achieving optimal results. We found that using four output frames consistently yields the best results, irrespective of the number of input frames.

We hypothesize that an insufficient number of output frames may fail to capture enough information, while a large number could introduce redundancy. Longer sequences may also divert the attention of LLMs, potentially compromising performance. Furthermore, the absolute number of output frames outweighs a fixed sampling ratio, highlighting the importance of carefully selecting the number of output frames to ensure optimal performance.

4.6. Ablation Study

To assess the contribution of various components in PQR, we conduct an ablation study, as detailed in Tab. 7. We ablate T-Former, question-guided queries, and temporal embeddings. Our results indicate that all key components contribute to improved performance across multiple datasets. Notably, the absence of T-Former results in significant performance degradation, highlighting its critical role. Additionally, leveraging question tokens to extract question-guided visual features also enhances performance. This collective improvement underscores the success and efficacy of our question-guided, localization-free spatio-temporal feature extraction approach.

Model	w/ LLM	#Params Train (Tot.)	NExT-QA				STAR					How2QA	VLEP
			Tem.	Cau.	Des.	Overall.	Int.	Seq.	Pre.	Fea.	Overall.		
Flamingo-9B 4-shot [1]	✓	- (9B)	-	-	-	-	-	-	-	-	42.8	-	-
Flamingo-9B 32-shot [1]	✓	- (9B)	-	-	-	-	-	-	-	-	41.2	-	-
Flamingo-80B 4-shot [1]	✓	- (80B)	-	-	-	-	-	-	-	-	42.4	-	-
Flamingo-80B 32-shot [1]	✓	- (80B)	-	-	-	-	-	-	-	-	42.2	-	-
FrozenBiLM [43]	✗	30M (30M)	-	-	-	-	-	-	-	-	-	81.5	-
All-in-One [35]	✗	110M (110M)	48.6	48.0	63.2	50.6	47.5	50.8	47.7	44.0	47.5	-	-
ATP [3]	✗	-	53.1	50.2	66.8	54.3	50.6	52.9	49.4	40.6	48.4	-	-
MIST [9]	✗	-	56.6	54.6	66.9	57.1	55.5	54.2	54.2	44.4	51.1	-	-
TranSTR [22]	✗	-	59.7	60.2	70.0	61.5	-	-	-	-	-	-	-
HiTeA [45]	✗	-	58.3	62.4	75.6	63.1	-	-	-	-	-	-	-
InternVideo [36]	✗	1.3B (1.3B)	58.5	62.5	75.8	63.2	62.7	65.6	54.9	51.9	58.7	79.0	63.9
BLIP-2^{voting} [19]	✓	188M (7.8B)	65.2	70.1	80.1	70.1	52.3	54.8	49.0	51.2	51.8	79.6	67.0
BLIP-2^{concat} [19]	✓	188M (7.8B)	68.1	72.9	81.2	72.6	65.4	69.0	59.7	54.2	62.0	82.2	68.6
SeViLa [46]	✓	106M (4.1B)	69.4	74.2	81.3	73.8	63.7	**70.4**	63.1	62.4	64.9	83.6	68.9
LLaMA-VQA [12]	✓	4.5M (7B)	72.7	69.2	75.8	72.0	**66.2**	67.9	57.2	52.7	65.4	-	**71.0**
PQR (Ours)	✓	50.1M (7.8B)	**72.8**	**76.9**	**84.7**	**76.7** (+2.9)	62.8	69.6	**72.8**	**70.0**	**67.6** (+2.2)	**85.9** (+2.3)	69.6

Table 8. **Comparison with SOTA video QA models.** NExT-QA involves *tem*poral, *cau*sal, and *des*criptive question types. STAR contains four question types: *int*eraction, *seq*uence, *pre*diction, and *fea*sibility. We emphasize the best performance in **bold**, and underline the second-best performance. PQR consistently outperforms SOTA models. Particularly, it manifests great temporal and causal reasoning capabilities on NeXT-QA and STAR. These results underscore the effectiveness of PQR's temporal modeling for VideoQA tasks.

4.7. Visualizing T-Former Attention Mechanisms

To highlight the learning behavior of T-Former, we attempt to visualize the query-video cross-attention maps. Fig. 4 showcases the attention map between question-guided temporal queries $\mathbf{q}$ and framewise visual tokens $\mathbf{e}_f^{1:n}$ in the last layer of T-Former across epochs 1, 5, and 10.

We observe that throughout the training, T-Former effectively learns the temporal dependencies between frames at different timestamps across the entire video sequence. The frames with the highest attention scores are distributed across the video, indicating T-Former's ability to discern the importance of different frames. This also demonstrates that T-Former can attend to and differentiate the significance of frames, enabling it to identify the most crucial frames without relying on explicit localization methods.

4.8. Comparison to State-of-the-art Models

Finally, we compare our PQR with state-of-the-art models across all four benchmarks, as presented in Tab. 8. Notably, PQR consistently outperforms existing models.

On NExT-QA, our model achieves the best performance in the overall category, with a 2.9% improvement. It also excels in the three subcategories that emphasize temporal modeling and causal reasoning, demonstrating the effectiveness of T-Former. Similarly, on the STAR dataset, PQR achieves a 2.2% performance boost over the current benchmark and attains the best performance in each subcategory. On How2QA, our model surpasses all others by 2.3%.

Our results reveal a significant performance boost compared to non-LLM baseline models, confirming the potency of LLMs as robust visual reasoners when provided with appropriate visual context. Moreover, PQR surpasses LLM-based competitors, particularly excelling in causal reasoning on NExT-QA and event prediction on STAR, further validating the efficacy of PQR.

Notably, PQR utilizes a smaller LLM as the reasoning backbone and a limited number of learnable parameters compared to other models. Furthermore, unlike most competitors that rely on extensive video-based pre-training, our model is trained exclusively on the target dataset. This highlights the efficiency and effectiveness of our approach.

5. Conclusion

In this work, we introduced T-Former, a novel question-guided temporal querying Transformer designed to address the challenges of dense video processing for multiple-choice video QA tasks. Our integrated framework, named PQR, combines T-Former with LLMs as reasoning agents, significantly outperforming other temporal modeling methods. Our approach also achieves substantial improvements over existing LLM-based models across four challenging video QA benchmarks while being trained solely on the target datasets without requiring any video-based pre-training phase. Through extensive experimental evaluations, we showcased the effectiveness of our approach. By leveraging compact, question-guided temporal tokens, we successfully bridge the gap between the complex temporal dynamics in videos and the reasoning capabilities of LLMs.

In future studies, we aim to extend the application of T-Former to a broader spectrum of video understanding tasks, including free-form video QA and more general video reasoning tasks. We believe that T-Former holds significant potential as a robust adaptation mechanism for MLLMs across diverse video understanding applications.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1, 8

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2

[3] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *CVPR*, 2022. 2, 8

[4] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. *ACL*, 2024. 1

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023. 1, 3, 4, 5

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 5

[9] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *CVPR*, 2023. 2, 8

[10] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *CVPR*, 2021. 2

[11] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Mingqian Tang, Ziwei Liu, and Marcelo H Ang Jr. Tada! temporally-adaptive convolutions for video understanding. *arXiv preprint arXiv:2110.06178*, 2021. 1

[12] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. In *EMNLP*, 2023. 8

[13] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 2022. 3

[14] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Neural reasoning, fast and slow, for video question answering. In *IJCNN*, 2020. 2

[15] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 1, 2

[16] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 2

[17] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 2

[18] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *EMNLP*, 2020. 2, 5

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. 1, 3, 4, 5, 8

[20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 5

[21] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, 2019. 2

[22] Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. Discovering spatio-temporal rationales for video question answering. In *ICCV*, 2023. 2, 8

[23] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 1

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[25] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023. 3

[26] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1

[27] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 1

[28] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *ACM Multimedia*, 2021. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[30] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang,

Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 1

[31] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 3

[32] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 2

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[35] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *CVPR*, 2023. 2, 8

[36] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 8

[37] Bo Wu, Shoubin Yu, Tenenbaum Joshua B Chen, Zhenfang, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021. 2, 5

[38] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 2, 5

[39] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, 2022. 2

[40] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 2

[41] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *TPAMI*, 2023. 1

[42] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 2

[43] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *NeurIPS*, 2022. 2, 8

[44] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. 3

[45] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *ICCV*, 2023. 8

[46] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *NeurIPS*, 2024. 1, 2, 3, 8

[47] Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Connecting speech encoder and large language model for asr. In *ICASSP*, 2024. 2

[48] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 2

[49] Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. Multi-event video-text retrieval. In *ICCV*, 2023. 1

[50] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *EMNLP*, 2023. 1, 2, 3, 4, 5

[51] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 1, 2

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 3, 5

[53] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *EMNLP*, 2022. 1, 2

[54] Yizhou Zhou, Xiaoyan Sun, Dong Liu, Zhengjun Zha, and Wenjun Zeng. Adaptive pooling in multi-instance learning for web video annotation. In *ICCVW*, 2017. 1

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 4