Foundation Models and Adaptive Feature Selection: A Synergistic Approach to Video Question Answering

Sai Bhargav Rongali
Indian Institute of Technology Bombay, India
rongalisaibhargav002@gmail.com

Mohamad Hassan N C
Indian Institute of Technology Bombay, India
mohdhassannc@gmail.com

Ankit Jha
LNMIIT, Jaipur, India
ankitjha16@gmail.com

Neha Bhargava
Fractal AI Research, India
neha.bhargava@fractal.ai

Saurabh Prasad
Univesity of Houston
saurabh.prasad@ieee.org

Biplab Banerjee
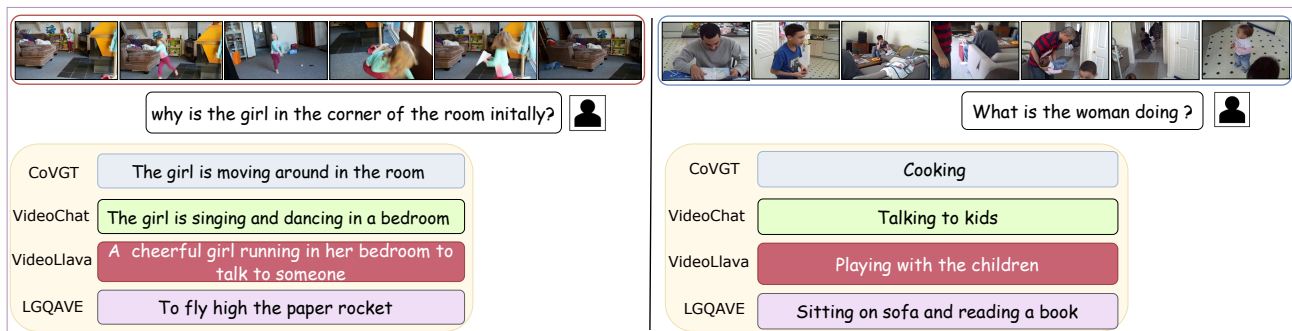Indian Institute of Technology Bombay, India
getbiplab@gmail.com

Figure 1. **Qualitative analysis of LGQAVE**. We present the answers produced by various state-of-the-art VideoQA models in response to a specific question paired with a sequence of frames from a given video in the NextQA [40] dataset. Our findings indicate that the answers generated by our LGQAVE model are notably more direct and precise in their semantic content.

Abstract

*This paper tackles the intricate challenge of video question-answering (VideoQA). Despite notable progress, current methods fall short of effectively integrating questions with video frames and semantic object-level abstractions to create question-aware video representations. We introduce **L**ocal - **G**lobal **Q**uestion **A**ware **V**ideo **E**mbedding (LGQAVE), which incorporates three major innovations to integrate multi-modal knowledge better and emphasize semantic visual concepts relevant to specific questions. LGQAVE moves beyond traditional ad-hoc frame sampling by utilizing a cross-attention mechanism that precisely identifies the most relevant frames concerning the questions. It captures the dynamics of objects within these frames using distinct graphs, grounding them in question semantics with the miniGPT model. These graphs are processed by a question-aware dynamic graph transformer (Q-DGT), which refines the outputs to develop nuanced global and local video representations. An additional cross-attention module integrates these local and global embeddings to generate the final video embeddings, which a language model uses to generate answers. Extensive evaluations across multiple benchmarks demonstrate that LGQAVE significantly outperforms existing models in delivering accurate multi-choice and open-ended answers.*

1. Introduction

Over the last decade, Video Question Answering (VideoQA) has evolved into a vital multidisciplinary field combining computer vision and natural language processing [50,55]. Despite advances, accurately interpreting video semantics relative to queries remains challenging, primarily due to the complex interplay between video content and questions, keeping VideoQA at the forefront of research demands. Current models focus on capturing spatiotemporal dynamics and aligning them with questions to derive answers [7,53], yet often require extensive dataset training and are prone to dataset biases, as frame selection is not guided

by language. Recent advancements have moved beyond simple feature summarization, constructing scene and temporal graphs to depict object-level interactions [44]. However, while adept at handling broad context queries, these approaches often miss the finer details necessary to analyze specific object interactions at the frame level.

Recently, the adoption of foundation models [31] (*e.g.* LlamaVid [22]) has significantly improved performance in several video comprehension tasks. While these models demonstrate superior performance, they face a specific challenge: they analyze all video frames indiscriminately, regardless of their relevance to the posed questions. Some studies have explored the paradigm of language-driven frame selection in contexts other than VideoQA [27, 49]. However, these approaches typically involve a complicated multi-stage pipeline, rely on secondary information sources such as image-based foundation models, or pose a multi-objective optimization framework, thus overburdening the entire process. Despite advancements in the frame selection stage, recent findings [41, 44] indicate that the outcomes of modern multi-modal foundation models for VideoQA are heavily biased towards the language cues, emphasizing the importance of the fundamental question: *To what extent are VideoQA outcomes relevant to the video contents?* Our research aims to precisely identify relevant video contents, both at the coarse and fine scales, guided by the given question semantics, for VideoQA.

Our solution: We introduce a unified solution, LGQAVE, for VideoQA with three principal novelties.

Our approach incorporates a learnable cross-attention module for **question-aware video frame selection**, which dynamically associates the question prompt with frame-level visual embeddings. This is achieved by applying a threshold to the cross-attention scores, enabling the precise isolation of video frames that are semantically aligned with the question. This method circumvents the complexities inherent in multi-stage frame selection pipelines and can be effortlessly integrated into any VideoQA system.

Looking forward, we propose the construction of spatial graphs for frames identified as most pertinent to the questions, termed as **question-aware local object selection and their interaction modeling**. This task is approached as a question-guided visual grounding, avoiding traditional object detection frameworks. To this end, we employ the miniGPT4 model [56] to process the questions alongside the selected frames, generating bounding box coordinates for the relevant objects.

Subsequently, frame-specific spatial graphs are constructed by considering the detected bounding boxes as the nodes and defining pairwise connections. When integrated with masked question embeddings, these graphs are fed into the Dynamic Graph Transformer (Q-DGT) model [44], which further refines the embeddings spatially and tempo-rally, enhancing the semantic coherence between the visual content and the question context.

In our final step, we aim to derive both **local and global video representations** from the outputs of Q-DGT to effectively address both long-video level and fine-grained frame-level questions. This is achieved by refining the global video representation through a query-key-value-based cross-attention mechanism, utilizing localized frame-level graph embeddings for answer generation (Fig. 1). Our significant contributions are summarized as,

[-] We introduce LGQAVE, an innovative model for Video QA that enhances the extraction of local and global video features, thoroughly guided by the question semantics.

[-] Our approach begins with cross-attention for question-aware frame selection, followed by using miniGPT4 for visual grounding to establish object interaction graphs based on the posed question. We then intuitively obtain the video representations through the Q-DGT module.

[-] We showcase the performance of LGQAVE on distinct VideoQA tasks and ablate the model rigorously. We observe steady improvements of 2-6% on average.

2. Related Works

Video question answering (VideoQA): Traditional VideoQA methods have primarily used video encoders on sparse frames [17, 36] or short segments [37], which struggle with spatiotemporal interactions and object compositionality [12], leading to suboptimal performance in reasoning tasks. Although cross-modal matching [3, 5, 6] and memory-based approaches [2, 28] have improved video content extraction, they rely heavily on frame-level or clip-level representations, which are often inadequate for detailed object relation reasoning. Advances in graph-based methods have facilitated object-level rationale; however, these methods tend to use either unified graphs that do not effectively differentiate spatial from temporal relations or static graphs that ignore temporal dynamics [44].

Transformers have significantly advanced the field of VideoQA. Models developed from datasets such as HowTo100M [24] employ proxy tasks like masked language modeling [13] and specific supervisions, such as future utterance prediction [46], to enhance performance. Despite outperforming traditional models [48, 57], transformer-based systems often focus on recognition or provide only shallow descriptions, struggling with visual relation reasoning due to noisy data and the limited scope of instructional videos [5]. Recent methods leveraging open-domain vision-text data face challenges with temporal relations and high operational costs. Despite their scalability, user-generated data can lead to overfitting. Large language models like BLIP-2 and MiniGPT-4 extend to video but encounter efficiency issues. Innovations such as Mo-bileVLM and LLaMA-Vid have improved feature repre-

sentation, and graphical models now effectively integrate both global and local features for enhanced dynamic reasoning [19, 25, 29, 51, 56]. *Our LGQAVE model advances beyond existing approaches by integrating visual and linguistic synergies at multiple scales. This integration enables the extraction of both global and local perceptions of video content, effectively addressing various queries.*

Graphs in VideoQA: Early VideoQA models such as TGIF-QA [14] and MSVD-QA [11] targeted specific actions and objects in video clips, leveraging spatio-temporal features to generate responses. Subsequent advancements led to more sophisticated models like HME-VideoQA [?] and Co-Mem [9], which utilize hierarchical memory networks and co-attentional frameworks to capture dynamic interactions within videos more effectively. HME-VideoQA builds hierarchical graphs by representing different levels of video granularity to capture temporal relationships. CoMem creates graphs through collaborative memory, linking video frames and question embeddings. Additionally, graph-based methods have proven effective for detailed visual understanding by representing video objects as graph structures. For example, LLaVA [25] enhances VLM performance by identifying objects pertinent to specific questions and constructing corresponding graphs. Conversely, the Contrastive Video Graph Transformer (CoVGT) [44] excels in providing global representations by focusing on the overall video content. They take all the objects that are present in a frame and form a graph, yet it falls short in local representations and lacks question-specific conditioning.

Despite these advancements, processing all video frames remains computationally expensive. Current methods focus on spatio-temporal dynamics and semantic alignment, yet they often manage vast amounts of data, leading to overlooking redundant content. Essential visual cues may be neglected, diminishing the accuracy of video interpretation.

Vision-language models (VLMs): Multimodal learning outperforms unimodal methods in tasks that require visual-semantic integration, such as image and text bridging. Recent developments have introduced foundation models like CLIP [35], FLORENCE [39], and ALIGN [15], which are particularly effective in these multimodal contexts. These models harness large-scale image-text pairs to tackle a variety of tasks in the CV/NLP domains, including zero-shot classification, object detection, image captioning, and VQA, to name a few. Despite their efficacy with still images, these models face challenges with long video sequences, primarily due to the extensive number of tokens required to represent each frame.

Models such as CLIP and ALIGN have proven effective in video recognition [18, 24, 32, 34] and video-text retrieval [8, 28]. However, they encounter difficulties in accurately capturing interactions between video content and

labels. Innovative models like Flamingo [1] and BLIP-2 [19] utilize web-scale image-text pairs, while Instruct-BLIP [52] and MiniGPT-4 [56] leverage high-quality instructional data sources. Methods such as Video-LLaMA [51] and VideoGPT [30] incorporate spatial and temporal pooling to overcome computational hurdles associated with long videos. LLaMA-VID [22] adopts a dual-token strategy to enhance the processing of long sequences. *In contrast, LGQAVE is designed to systematically utilize question guidance for frame selection and the modeling of relevant objects within and across frames. This approach aims to minimize redundancy and irrelevance in video features, thereby enhancing the efficiency and accuracy of VideoQA.*

3. Proposed Methodology

In this section, we define the problem and outline the objectives for the LGQAVE framework. We consider a dataset $\mathcal{D}$ that consists of video sequences V, questions Q, and corresponding labeled answers A. The primary objective is to learn a mapping function $\phi : (V, Q) \rightarrow A$ that accurately predicts the correct answer A_i for each given question Q_i associated with a video V_i.

To accomplish this, LGQAVE is structured into four key components (Fig. 2): **a. Question-driven frame selection module**—identifies the most relevant video frames, thereby reducing redundancy at the frame level. **b. Frame-centric object graph construction**—emphasizes the important objects and their interactions within the selected frames through visual grounding, minimizing redundancy at a finer level. **c. Question-aware dynamic graph transformer (Q-DGT)**—facilitates effective selection and fusion of local and global video features. **d. Answer prediction module**—generates accurate answers based on the enriched video and question representations. These components, detailed below, collaboratively leverage the semantics of the questions to ensure a discriminative and contextually rich embedding space. The variables are summarized in the supplementary materials.

3.1. Question-aware video frame selection

In VideoQA, processing every video frame in a sequence is both computationally intensive and time-consuming, often leading to redundancy when dealing with frame splits. To tackle these issues, our LGQAVE framework incorporates a novel frame selection module designed to sample question-aware, key video frames from video-question pairs. This is achieved by utilizing a cross-attention mechanism to calculate relevance scores between the question tokens and video frames, ensuring that only the most pertinent frames are selected.

Mathematically, we denote the video frame at the tth time step for the ith instance as $V_i^t \in \mathbb{R}^{H \times W \times 3}$, where $t \in \{1, \ldots, \mathcal{T}_i\}$ and $\mathcal{T}_i$ represents the total number of video
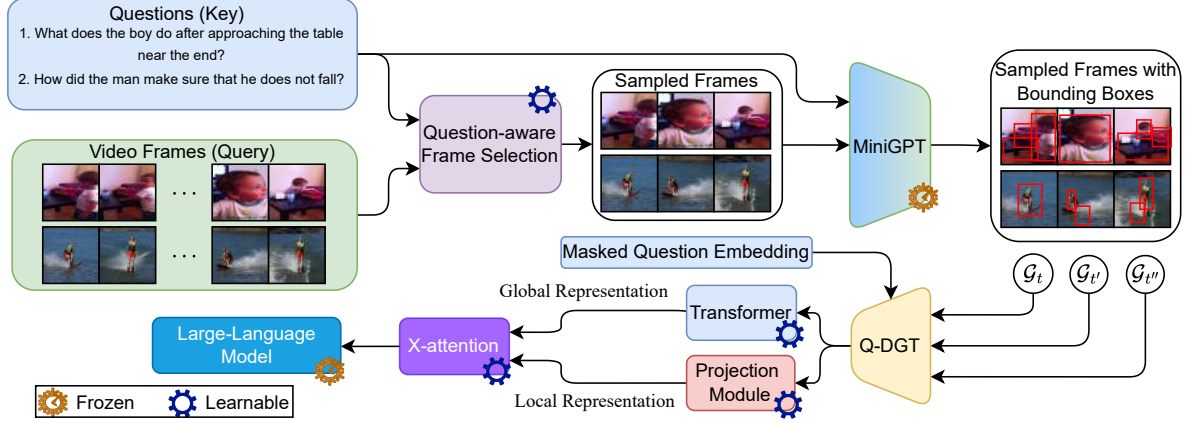
Figure 2. **Schematic of the model diagram for LGQAVE**. Given a question and its corresponding video, our process begins with a question-aware frame sampling module that identifies the pertinent frames from the video. Subsequently, a miniGPT4-based visual grounding module constructs object relation graphs from these selected frames. The Q-DGT module then processes these graphs along with masked question embeddings to produce local and global video representations. A cross-attention module further refines the global features by incorporating contextual knowledge from the local features. Finally, a language model-based answer generator utilizes these refined features to predict the answers.

frames for the ith instance. Here, H and W denote the height and width of the extracted frames, respectively. We use a frozen CLIP image encoder f_v to extract visual features $\mathbf{E_i^t} \in \mathbb{R}^{\mathcal{N} \times \mathcal{C}}$ from V_i^t, where $\mathcal{N} = \frac{H}{p} \times \frac{W}{p}$, with p representing the patch size, and $\mathcal{C}$ being the embedding dimension. Additionally, We extract the text-guided query $\mathbf{Q}_i \in \mathbb{R}^{\mathcal{M} \times \mathcal{C}}$ using a pre-trained RoBERTa [26] model for the question Q_i, where $\mathcal{M}$ denotes the number of queries. The visual features $\mathbf{E}_i^t$ and the text-guided query features $\mathbf{Q}_i$ are then passed through learnable projection ϕ_e and ϕ_q layers to obtain the projected features $\tilde{\mathbf{E}}_i^t$ and $\tilde{\mathbf{Q}}_i$, respectively. These projected features are subsequently fed into the cross-attention module defined in Eq. 2, where a cross-attention score between the question and the t-th frame, s_t, is computed as follows,

$$\tilde{\mathbf{E}}_i^t = \phi_e(\mathbf{E_i^t}), \tilde{\mathbf{Q}}_i = \phi_q(\mathbf{Q_i}) \qquad (1)$$

$$s_t = \text{Mean}\left(\text{Softmax}\left(\tilde{\mathbf{E}}_i^t \cdot \tilde{\mathbf{Q}}_i^\top\right) \cdot \tilde{\mathbf{Q}}_i\right) \qquad (2)$$

Finally, we select the frame V_i^t based on the cross-attention score s_t, provided it surpasses a predefined threshold β. The subset of selected frames from V_i is denoted as $\mathcal{V}_i$. These selected frames are then processed further to construct spatial object graphs for each selected frame.

3.2. Obtaining graph-based frame representation

We utilize the MiniGPT-4 architecture to construct question-aware object graphs from the selected frames in $\mathcal{V}_i$, contrasting with traditional models that perform object detection across all frames without considering the question context. MiniGPT-4's efficiency lies in requiring only

a linear layer to align visual features with the Vicuna model [54]. Additionally, we redefine the object detection task in LGQAVE as a visual grounding task, using the frames from $\mathcal{V}_i$ and the question Q_i, where MiniGPT-4 excels.

For each selected frame $V_i^{t'} \in \mathcal{V}_i$, we also include the two preceding and two subsequent frames: $V_i^{t'-2}, V_i^{t'-1}$ through $V_i^{t'+1}, V_i^{t'+2}$—to ensure temporal continuity and minimize the risk of missing critical sequential information, which we fixed through empirical validation. These frames, along with the question prompt $\mathbf{Q_i}$ from the frozen RoBERTa model, are fed into MiniGPT-4, which processes them to generate m bounding boxes $\mathcal{B}_i^{t'}$ around the objects pertinent to the question in $V_i^{t'}$. Four coordinates define each bounding box, and the total number of bounding boxes per frame is limited to $m \leq 10$.

We enhance the graph representation methodology by utilizing detected objects, advancing beyond the approach in [44]. For each highlighted object instance in a video frame $V_i^{t'}$, we extract Region of Interest (RoI)-aligned features as object appearance representations $F_o^{t'}$ which also contains spatial locations $F_s^{t'}$ of the respective objects. Additionally, we capture a frame-level feature $F_I^{t'}$ to augment the graph representations derived from the local objects. We aim to construct a frame-specific spatial graph using $F_u^{t'} = F_o^{t'} \cup F_I^{t'}$.

While our methodology is inspired by [44], it differs significantly in its execution. Unlike [44], which assumes static object groups within a video clip and employs a fixed linking score based on appearance and spatial location $F_s^{t'}$, our approach with MiniGPT-4 dynamically tracks objects across the video sequence. This dynamic tracking enhances the robustness and adaptability of our model, particularly

improving its generalizability to longer video sequences.

Graph construction for $V_i^{t'}$: We propose to consider the bounding boxes from $\mathcal{B}_i^{t'}$ and the entire frame $V_i^{t'}$ as constituting the $m + 1$ nodes in the frame-specific graph $\mathcal{G}_i^{t'}(A^{t'}, R^{t'})$, with $A^{t'}$ denoting the node-set, and we put up an edge between two bounding boxes, and the edge weights are defined as follows,

$$R^{t'} = \text{Softmax}\left(\phi_k(F_u^{t'})\phi_v(F_u^{t'})^{\top}\right) \qquad (3)$$

Here, ϕ_k and ϕ_v denote linear transformations and the transpose operation is denoted by $(\cdot)^{\top}$. The obtained $\mathcal{G}_i = \{\mathcal{G}_i^1, \mathcal{G}_i^2 \dots\}$ which contains the object representation and also the spatial representations $F_s^{t'}$ are passed to the Q-DGT module for video feature extraction.

3.3. Question-aware dynamic graph transformer

Following the methodologies proposed in [44], we utilize DGT to capture the complex dynamics of objects from the obtained graphs. However, different from [44], to enhance the relevance of the object dynamics to the specific questions posed, we condition the DGT on the question (Q-DGT), focusing the analysis only on the objects that are essential for answering the questions. This conditional approach ensures that our model's attention is selectively tuned to the pertinent elements of the video content. Furthermore, to enhance the contextual relevance of the visual information extracted by the Q-DGT module, our approach goes beyond the typical refinement processes described in [44], which focuses solely on global representations. We extend refinement to both global and local representations, thereby improving the accuracy and contextual depth of the answer prediction.

In Q-DGT, the question embedding $\tilde{\mathbf{Q}}_i$ is intentionally masked to control the influence of the question representation on the model. This masking helps isolate specific features, mitigating the risk of overfitting by finely tuning the interaction between the question representation and the object dynamics captured by the DGT. Such an approach ensures that only the most relevant dynamics are emphasized, enhancing the model's accuracy and generalizability.

$$\hat{\mathbf{Q}} = \mathbf{M} \odot \tilde{\mathbf{Q}}_i$$

Here, $\mathbf{M}$ is binary mask vector, $\odot$ denotes the element-wise (Hadamard) product.

Q-DGT integrates both a temporal and a spatial graph transformer unit to process the input visual graphs within $\mathcal{G}_i$. $F_s^{t'}$ is the input to the spatial unit that models the spatial relationships within each frame, and $F_u^{t'}$ is the input to the temporal unit that captures relationships across different frames. These units are specifically designed to handle the different dimensions of data - temporal changes over time and spatial relationships within frames. The output of the

Q-DGT is a local representation $\mathcal{F}_{local}^{t'}$ corresponding to the t'-th frame, which is obtained by non-linearly transforming the embeddings from the frame-specific graph through a trainable projection layer with parameters ϕ_{local}. This projection layer is crucial as it ensures that vital information pertinent to the video is preserved and not lost in transformation processes. The local representation formula is:

$$\mathcal{F}_{local_i}^{t'} = \phi_{local}(\text{Q-DGT}(\mathcal{G}_i^{t'}, \hat{\mathbf{Q}})) \qquad (4)$$

Additionally, a global representation $\mathcal{F}_{global}$ is derived by aggregating all the spatial and temporal representations through a global transformer, similar to the approach in [44]. This global transformer incorporates learnable sinusoidal temporal position embeddings to model the sequence of events within the video effectively. The outputs of this transformer are then mean-pooled to produce a comprehensive global representation $\mathcal{F}_{global}$ of the entire video, which encapsulates both the spatial and temporal dynamics across all processed frames. MHSA stands for Multihead Self Attention, and MPool represents the Maxpooling operation.

$$\mathcal{F}_{global_i} = \text{MPool}(\text{MHSA}(\text{Q-DGT}(\mathcal{G}_i, \hat{\mathbf{Q}}))) \qquad (5)$$

For further details regarding DGT, refer [44].

Interaction of the question and graph features in Q-DGT: To integrate textual context effectively, we employ a RoBERTa language model to process the question Q and project the token outputs into a textual information space $Z_{\hat{\mathbf{Q}}}$ using a linear transformation:

$$Z_{\hat{\mathbf{Q}}} = \phi_{\hat{Q}}(\hat{\mathbf{Q}}) = \{z_{\hat{\mathbf{Q}}}^h\}_{h=1}^{\mathcal{H}},$$

where $\phi_{\hat{Q}}$ is a projection matrix in $\mathbb{R}^{768 \times d}$, $\mathcal{H}$ denotes the number of tokens in $\hat{\mathbf{Q}}$, and $z_{\hat{\mathbf{Q}}}^h$ represents the embedded representation of the hth token. The encoded tokens include those representing the words of an open-ended question Q or QA pairs in a multiple-choice format.

Within the DGT framework, the cross-modal encoder $Q\text{-}DGT_{cm}$ processes the textual embeddings $Z_{\hat{\mathbf{Q}}}$ together with the visual embeddings $\tilde{\mathbf{E}}_i^{t'}$ corresponding to the frame $\hat{V}_i^{t'}$ for the ith instance. This integration facilitates a nuanced refinement of both local and global video representations:

$$\mathcal{F}_{local} = Q\text{-}DGT_{cm}(\mathcal{F}_{local}, Z_{\hat{\mathbf{Q}}}) = \mathcal{F}_{local} + \sum_{h=1}^{\mathcal{H}} \alpha_h^1 z_{\hat{\mathbf{Q}}}^h, \qquad (6)$$

$$\mathcal{F}_{global} = Q\text{-}DGT_{cm}(\mathcal{F}_{global}, Z_{\hat{\mathbf{Q}}}) = \mathcal{F}_{global} + \sum_{h=1}^{\mathcal{H}} \alpha_h^2 z_{\hat{\mathbf{Q}}}^h, \qquad (7)$$

where α^1 and α^2 are attention weights. These weights are calculated by applying a sigmoid function σ to the transpose

dot product between $\mathcal{F}_{local}$ or $\mathcal{F}_{global}$ and $Z_{\hat{\mathbf{Q}}}$, emphasizing the dynamic and context-sensitive interactions between the modalities. $\alpha^1 = \sigma(\mathcal{F}_{local}) \odot Z_{\hat{\mathbf{Q}}}$, $\alpha^2 = \sigma(\mathcal{F}_{global}) \odot Z_{\hat{\mathbf{Q}}}$.

3.4. Obtaining the final video features

Our method acknowledges the dynamic relevance of global image context and local properties based on the question posed. To adeptly handle this variability, we introduce an adaptive mechanism that updates the global embedding through a cross-attention process with the local embeddings obtained from the Q-DGT module.

The final representation of the answer leverages cross-attention between the local representations $\{\mathcal{F}_{local}^{t'}\}$ and the global representation $\mathcal{F}_{global}$. Directly merging these representations often leads to redundancy due to overlapping information. To address this issue, our attention mechanism is designed such that $\mathcal{F}_{global}$ serves as the query, while $\{\mathcal{F}_{local}^{t'}\}$ function as both keys and values. This structure allows the model to dynamically emphasize the most relevant details from the local context when updating the global representation. This results in a more discriminative and contextually refined final representation.

$$\mathcal{F}_{final} = (1-\gamma)\mathcal{F}_{global} + \gamma\text{Cross-Att}(\mathcal{F}_{global}, \{\mathcal{F}_{local}^{t'}\}), \tag{8}$$

γ is a weighting constant within the range $[0,1]$. $\mathcal{F}_{final}$ embodies a comprehensive, question-aware representation of the video. It seamlessly integrates the broad contextual overview provided by the global features with the detailed insights offered by the local features.

3.5. Answer generation

In our framework, we employ distinct strategies for answering objective and subjective questions, leveraging the synthesized representation $\mathcal{F}_{final}$.

For objective questions, the answer prediction $\hat{A}$ is determined by calculating the similarity scores between $\mathcal{F}_{final}$ and a set of pre-encoded answer representations F_A. Here, $A = \{A_l\}_{l=1}^{|A|}$, where $|A|$ represents the number of answer options, and $\mathbf{A}_l$ denotes the RoBERTa-encoded representation of each option l. The prediction is made by identifying the option associated with the highest similarity score:

$$\hat{A} = \arg\max\left((\mathcal{F}_{final})^\top \mathbf{A}\right) \tag{9}$$

We adopt a methodology for subjective questions that enables a video-absent QA scenario, as delineated in prior works [44]. The answer is inferred by evaluating the similarities not only between $\mathcal{F}_{final}$ and A but also between the question representation $\tilde{\mathbf{Q}}$ and $\mathbf{A}$. The final prediction $\hat{A}$ is obtained by taking an element-wise product of these similarity matrices, thereby ensuring that the decision robustly

integrates cues from both the video and the question:

$$\hat{A} = \arg\max\left((\mathcal{F}_{final})^\top \mathbf{A} \odot (\tilde{\mathbf{Q}})^\top \mathbf{A}\right) \tag{10}$$

3.6. Loss objectives

Loss function for multi-choice QA: We employ a composite loss function in multi-choice question answering, where answers are selected from given options. The component L_{vqa} accounts for the interaction between the video, the question, and the multiple-choice options, while L_{vq} pertains solely to the video and the question:

$$L = L_{vqa}(\mathcal{F}_{final}, \tilde{\mathbf{Q}}\otimes\mathbf{A}^+, \tilde{\mathbf{Q}}\otimes\mathbf{A}^-)+\lambda L_{vq}(\mathcal{F}_{final}, \mathbf{Q}^+, \mathbf{Q}^-) \tag{11}$$

A^+ and A^- represent the correct and incorrect answer options, respectively. Similarly, Q^+ and Q^- denote the positive and negative questions associated with a video. The balancing parameter is represented by λ, and the symbol $\otimes$ indicates a concatenation operation.

Loss function for open-ended QA: For open-ended QA, where the answer a is not constrained to predefined options, the loss formulation needs to adapt to the broader scope of potential answers:

$$L = L_{vqa}(\mathcal{F}_{final} \otimes \tilde{\mathbf{Q}}, \mathbf{A}^+, \mathbf{A}^-)+\lambda L_{vq}(\mathcal{F}_{final}, \tilde{\mathbf{Q}}^+, \tilde{\mathbf{Q}}^-) \tag{12}$$

4. Experimental Evaluations

Datasets[†]: : We conduct experiments across various datasets to evaluate different aspects of video understanding. The datasets include NExT-QA [40], STAR-QA [38], and Causal-VidQA [20], which are designed to address complex temporal and causal relationships as well as commonsense reasoning within videos, with a particular focus on temporal dynamics. Additionally, we utilize TGIF FrameQA [14], MSRVTT-QA [45], and ActivityNetQA [4], which concentrate on the recognition of video objects, their attributes, actions, and activities, emphasizing static frame analysis.

4.1. Main results

We compare LGQAVE with several relevant and recent methods from the literature in Table 1 on all the datasets mentioned above. LGQAVE significantly surpasses the previous SOTAs on all tasks defined in previously mentioned datasets, improving the accuracy on an average by 9.29% vs. non-LLM methods like CoVGT [44] and 6.61% vs. LLM models like VideoLlava [23], respectively. Compared to other methods, we paid more attention to the video content related to the question instead of taking all the video

---

[1][†]More about datasets and implementation details in supplementary material.

Table 1. Accuracy (%) comparison on NExT-QA [40], TGIF-FrameQA [14], MSRVTT-QA [45] and ActivityNet-QA [4]. Acc@C, T, and D denote accuracy for Causal, Temporal, and Descriptive questions. The **best** and 2nd best results are highlighted in bold and underlined, respectively.

| Methods | Text | NExT-QA Val | | NExT-QA Test | | TGIF-FrameQA | MSRVTT-QA | ActivityNet-QA | Star-QA | Causal-VidQA |
		Acc@C	Acc@T	Acc@D	Acc@All					
VQA-T [47]	DistilBERT	41.66	44.11	59.97	45.30	25.30	40.40	15.70	29.61	40.32
HGA [16]	BERT	46.26	50.74	59.33	51.02	20.70	31.53	14.82	32.27	44.82
HQGA [42]	BERT	48.48	51.24	61.65	51.34	25.40	33.80	17.51	35.83	47.36
ATP [3]	BERT	51.57	52.00	66.80	53.18	26.33	31.76	16.47	39.27	50.14
VGT [43]	BERT	52.28	55.09	61.94	53.68	61.60	39.70	20.40	42.43	53.20
VGT (PT) [43]	BERT	53.43	56.39	59.64	55.70	61.70	3.70	19.70	44.32	54.35
CoVGT [44]	RoBERTa	58.53	57.48	63.82	57.40	61.60	38.30	24.50	46.20	60.80
VideoChat [21]	-	62.30	59.36	64.22	56.27	34.40	45.00	26.50	49.35	66.64
VideoLlama [51]	-	61.53	61.25	66.35	58.41	-	29.60	12.40	53.47	68.35
VideoLlava [23]	-	<u>63.70</u>	<u>63.45</u>	69.10	60.08	70.00	59.20	**45.30**	**62.25**	<u>70.31</u>
LGQAVE	RoBERTa	**68.69**	**68.00**	**74.88**	**66.69**	**72.40**	**63.43**	<u>44.81</u>	<u>61.48</u>	**73.59**

content features and giving answers based on such features. A sampling of frames gained a lot of popularity recently and fine grained frame selection method as shown in [33] works much better than other sampling methods.

Methods like VideoLlava and VideoChat, which do not use graph-based approaches, generalize well for tasks like video summarization and captioning but struggle with reasoning and frame-level questions, particularly in longer videos. In various datasets, we have seen that existing methods struggle to answer reasoning questions on videos with more than 600 lengths, primarily when it is based on a few frames in the video. VideoGPT addresses this by using frame-level captions, but it requires test videos to match the distribution of training videos. Our approach leverages graphs and the LLM model miniGPT, focusing on video understanding without relying on captions, particularly excelling in object recognition. Our methods takes approximately 289GFlops during training and 138GFlops during testing.

Our frame selection vs coarse frame selection [10]: Compared with the coarse frame selection process, which generally employs BridgeFormer [10], our frame selection (FFS) is better in picking up the correct frames related to the question. An average increase of 4.23% is observed with this method alone, as shown in Table 2. BridgeFormer [10] concentrates on nouns and verbs from the question, removing the remaining phrase of the question. In comparison, the fine-grained frame selection process takes all the parts of speech in the question into context, which makes it robust in selecting appropriate frames related to the question.

LGQAVE vs. graph methods + sampling: To show the supremacy of our model, we conducted thorough experiments by including frame sampling methods with the existing graph methods as shown in Table 2. Improved the existing HQGA [42], existing CoVGT [44] by adding the frame sampling modules in their architectures. We made two versions of these architectures, one by adding a coarse frame selection method [10] and another by adding a finer frame selection method [33]. LGQAVE architecture works better than any other graph video question-answering model even after the frame selection process, which shows that making graphs using objects specific to the question and their local

and global representations gives an advantage to our model for a better understanding of the question and answering it.

Table 2. Detailed comparison between LGQAVE and other SOTA methods for frame sampling. CFS [10]: coarse frame selection [33]. FFS: fine frame selection.

| Models | NExT-QA Val | | | |
	Acc@C	Acc@T	Acc@D	Acc@All
HQGA+CFS	50.66	54.11	59.97	48.30
HQGA+FFS	52.27	53.29	62.17	49.40
CoVGT+CFS	61.62	59.08	66.42	59.02
CoVGT+FFS	64.31	62.27	69.50	61.47
LGQAVE (RoBERTa)	**68.69**	**68.00**	**74.88**	**66.69**

Comparison of other graph-based methods with frame selection process: Table 2 compares our LGQAVE model to other state-of-the-art graph-based methods on the NExT-QA validation set. The HQGA models, employing either Coarse Frame Selection (CFS) or Fine Frame Selection (FFS), show limited performance, with `Acc@All` at only 49.40%. CFS, which selects a broad range of frames for a general video overview, often includes irrelevant frames and misses finer details crucial for precise answers, resulting in lower accuracy for HQGA+CFS and CoVGT+CFS. Conversely, FFS targets the most relevant frames, focusing on specific objects or actions linked to the questions, thus improving accuracy. This method filters out extraneous content, concentrating on critical frames and leading to higher `Acc@All` scores for HQGA+FFS and CoVGT+FFS.

Our LGQAVE model, enhanced with RoBERTa, significantly outperforms both methods with an `Acc@All` of 66.69%, demonstrating the effectiveness of integrating local and global visual features with advanced textual encoding for more accurate, context-aware video question answering. The results underscore the advantages of LGQAVE, especially when combined with sophisticated language models, in leveraging both detailed visual representations and broader scene context.

5. Ablation analysis

To better understand the contribution of each component in the proposed model LGQAVE, we conducted ablation over the various components, shown in Table 3. The inclu-
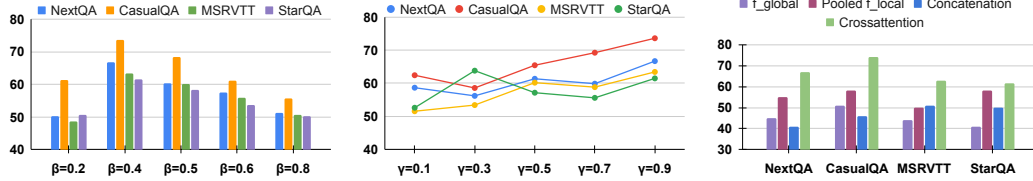
Figure 3. Performance of LGQAVE with change in β and γ parameters on various datasets are shown in the first two plots. Performance of LGQAVE with usage of different combinations of $\mathcal{F}_{local}$ and $\mathcal{F}_{global}$. is shown at the end.
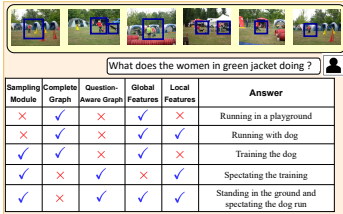

Figure 4. Qualitative answers[†] by LGQAVE model for various ablation configurations on a video from the NextQA dataset.

sion of a sampling strategy markedly enhances our model's performance. Without sampling (Configuration C-1), the model depends solely on global representations, which limits its focus on pertinent frames and leads to reduced accuracy, notably in Acc@All. Introducing sampling in Configuration C-2 improves focus on relevant frames, resulting in significant performance gains across all metrics, particularly in Acc@C (+3.36%) and Acc@T (+3.15%), by filtering out extraneous information.

Table 3. Ablation analysis of the proposed model components of LGQAVE on the NExT-QA dataset.

Conf.	Sampling	miniGPT	local Repr.	Global Repr.	NExT-QA			
					Acc@C	Acc@T	Acc@D	Acc@All
C-1	×	×	×	✓	58.53	57.48	63.82	57.40
C-2	✓	×	×	✓	61.89	60.63	65.37	61.85
C-3	✓	✓	×	✓	65.42	64.79	71.53	64.76
C-4	✓	×	✓	✓	59.26	56.18	57.46	58.13
C-5	✓	✓	✓	✓	**68.69**	**68.00**	**74.88**	**66.69**

The integration of miniGPT in C-3, combined with sampling but excluding local representations, significantly enhances accuracy, particularly in Acc@D (+6.16%), suggesting that miniGPT enriches the model's contextual understanding and response accuracy. In contrast, using graphs with all objects in a frame and employing both local and global representations leads to a severe drop in accuracy, as observed in C-4. In C-5, we leverage local and global representations by cross-attention to balance detailed object-level insights and broader scene context, resulting in the highest accuracy across all metrics. This approach outperforms models that rely solely on global features by +9.29% in Acc@All.

In Figure 3, we analyze the impact of varying the parameter β and γ across four datasets: NextQA, CasualQA, MSRVTT, and StarQA. The study suggests that the optimal β and γ are 0.4 and 0.9, respectively, where the highest performance is observed, with performance declining at higher or lower values. Also, we show that our cross-attention of $\mathcal{F}_{local}$ and $\mathcal{F}_{global}$ gives better accuracy than using $\mathcal{F}_{global}$

or pooled $\mathcal{F}_{local}$ or concatenating them.

Fig 4 highlights the impact of various model configurations on the task of answering video questions. The sampling module, made of graphs and global and local features, was individually assessed for their contribution to the model's overall performance. By isolating each module, the study reveals that using the question-aware object interaction graphs in combination with Local Features significantly improves the accuracy of the model's predictions. For instance, it enables the model to generate more specific and contextually appropriate answers, such as distinguishing between actions like "Training the dog" and "Spectating the dog run." This suggests that incorporating both question-awareness and fine-grained local features plays a crucial role in understanding video content.

6. Takeaways

We present LGQAVE, a novel framework that addresses limitations in existing VideoQA approaches by enhancing multi-modal integration and focusing on semantic visual concepts relevant to the questions. Using cross-attention, LGQAVE identifies the most pertinent video frames for each query, surpassing traditional frame sampling techniques. Our approach generates precise video representations by capturing object dynamics through spatial graphs and grounding them in question semantics via the MiniGPT model. Q-DGT refines these representations, ensuring global and local video content is optimally encoded. An additional cross-attention module synthesizes final video embeddings conditioned on the questions, leading to more accurate answer generation by the language model. Extensive evaluations across benchmarks show that LGQAVE significantly improves accuracy in multi-choice and open-ended VideoQA tasks, suggesting future opportunities to leverage advanced graph-based and attention mechanisms for multi-modal integration.

Acknowledgments

---

[2][†]More qualitative results in supplementary material.

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. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 2

[3] Shyamal Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in Video-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 7

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 6, 7

[5] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 2

[6] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. Mobilevlm v2: Faster and stronger baseline for vision language model. *ArXiv*, abs/2402.03766, 2024. 2

[7] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 1

[8] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733, 2023. 3

[9] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6576–6585, 2018. 3

[10] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridgeformer: Bridging video-text retrieval with multiple choice questions. *arXiv preprint arXiv:2201.04850*, 2022. 7

[11] Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 599–603, 2018. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[13] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *CoRR*, abs/2403.19046, 2024. 2

[14] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 3, 6, 7

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ArXiv*, abs/2102.05918, 2021. 3

[16] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:11109–11116, 04 2020. 7

[17] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *ArXiv*, abs/2311.08046, 2023. 2

[18] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 3

[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. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[20] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022. 6

[21] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 7

[22] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 2, 3

[23] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6, 7

[24] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 2, 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[26] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. Roberta: A robustly optimized bert pretraining approach. arxiv [preprint](2019). *arXiv preprint arXiv:1907.11692*, 1907. 4

[27] Haoyu Lu, Mingyu Ding, Nanyi Fei, Yuqi Huo, and Zhiwu Lu. Lgdn: Language-guided denoising network for video-language modeling. *Advances in Neural Information Processing Systems*, 35:25198–25211, 2022. 2

[28] 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 and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 3

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[30] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *ArXiv*, abs/2406.09418, 2024. 3

[31] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024. 2

[32] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 3

[33] Vidyaranya Nuthalapati and Anirudh Tunga. Coarse to fine frame selection for online open-ended video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 353–361, October 2023. 7

[34] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[37] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019. 2

[38] Bo Wu and Shoubin Yu. Star: A benchmark for situated reasoning in real-world videos. *ArXiv*, abs/2405.09711, 2024. 6

[39] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4818–4829, June 2024. 3

[40] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 1, 6, 7

[41] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 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, 2024. 2

[42] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2804–2812, 2022. 7

[43] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. 7

[44] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, and T. Chua. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, 45(11):13265–13280, nov 2023. 2, 3, 4, 5, 6, 7

[45] 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. 6, 7

[46] Hu Xu, Gargi Ghosh, Po-Yao (Bernie) Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2021. 2

[47] 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. 7

[48] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 2

[49] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[50] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 1

[51] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 7

[52] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. 3

[53] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, Yueting Zhuang, Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, volume 2, page 8, 2017. 1

[54] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. 4

[55] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. 1

[56] 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. 2, 3

[57] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2