Granularity-Adaptive Spatial Evidence Tokenization for Video Question Answering

Hao Jiang[1], Yang Jin[1], Zhicheng Sun[1], Kun Xu[2], Kun Xu[2], Liwei Chen[2], Yang Song[2], Kun Gai[2], Yadong Mu[1]*

[1]Peking University
[2]Kuaishou Technology
jianghao@stu.pku.edu.cn, myd@pku.edu.cn

Abstract

Video question answering plays a vital role in computer vision, and recent advances in large language models have further propelled the development of this field. However, existing video question answering techniques often face limitations in grasping fine-grained video content in spatial dimensions. It mainly stems from the fixed and low-resolution input of video frames. While some approaches using high-resolution inputs partially alleviate this problem, they introduce excessive computational burdens by encoding the entire high-resolution image. In this work, we propose a granularity-adaptive spatial evidence tokenization model for video question answering. Our method introduces multi-granular visual tokenization in the spatial dimension to produce video tokens at various granularities based on the question. It highlights spatially activated patches at low resolutions through a granularity weighting module and then adaptively encodes these activated patches at high resolution for detail supplementation. To mitigate the computational overhead associated with high-resolution frame encoding, a masking and acceleration module is developed for efficient visual tokenization. Moreover, a granularity compression module is designed to dynamically select and compress visual tokens of varying granularities based on questions. We conduct extensive experiments on 11 mainstream video question answering datasets and the experimental results demonstrate the effectiveness of our proposed method.

Introduction

The past few years have witnessed the rapid development of video understanding technologies, which are widely applied in various computer vision tasks, including video question answering (Xiao et al. 2024; Yu et al. 2024), video summarization (Zhu et al. 2020; Jiang and Mu 2022), video spatial and temporal grounding (Wasim et al. 2024; Jiang, Yizhang, and Mu 2024), etc. The advent of pretraining paradigm has promoted the development of multimodal large language models (LLMs), enhancing video understanding capabilities and achieving remarkable progress in generating accurate responses.

Existing works on LLMs-based video question answering can be roughly divided into two categories: the improvement of visual tokenization strategies (Li et al. 2024b; Cheng et al. 2024a; Fei et al. 2024) and the enhancement of alignments between visual tokens and the LLM semantic space (Liu et al. 2024b; Lin et al. 2023; Wang et al. 2023b). However, a limitation of these approaches is that most models learn visual features at a fixed, lower resolution (such as 224×224). This often results in the loss of fine-grained information when dealing with local details in the video, thereby reducing the model's prediction accuracy.

As shown in Figure 1, with a question like "Does the athlete in the video have long hair?", the target area of "athlete's hair" only spatially occupies a small portion of the frame. Lower resolution video frames lack the fine-grained encoding necessary for capturing detailed image content, and fixed resolution fails to highlight specific target areas, allowing irrelevant visual features to dominate the encoded tokens. While some image LLMs (Li et al. 2024c; Liu et al. 2024a) have advanced in capturing high-resolution visual details by dividing high-resolution images into several sub-patches and encoding them separately, these approaches suffer from two major drawbacks. Firstly, their methods require encoding the entire high-resolution image. Since the number of patches increases quadratically with the resolution, these approaches inevitably result in excessive computational overhead, particularly when applied to videos. Secondly, the separate encoding of sub-patches prevents the model from accessing the global semantic context of the image, which impairs the learning of comprehensive high-resolution features.

To address these challenges, in this work, we propose a granularity-adaptive spatial evidence tokenization framework for video question answering. Specifically, our method introduces multi-granular visual tokenization to produce video tokens at various granularities based on the question. It highlights spatially activated patches at low resolutions through a granularity weighting module and then adaptively encodes these activated patches at high resolution for detail supplementation. To mitigate the computational overhead associated with high-resolution frame encoding, a masking and acceleration module is developed for efficient visual tokenization. Compared to existing high-resolution image LLMs (Li et al. 2024c), our method offers the advantage of not requiring the encoding of the entire frame content. Instead, it effectively grounds compact spatial evidence necessary to answer questions from high-resolution frames.

*Corresponding Author.

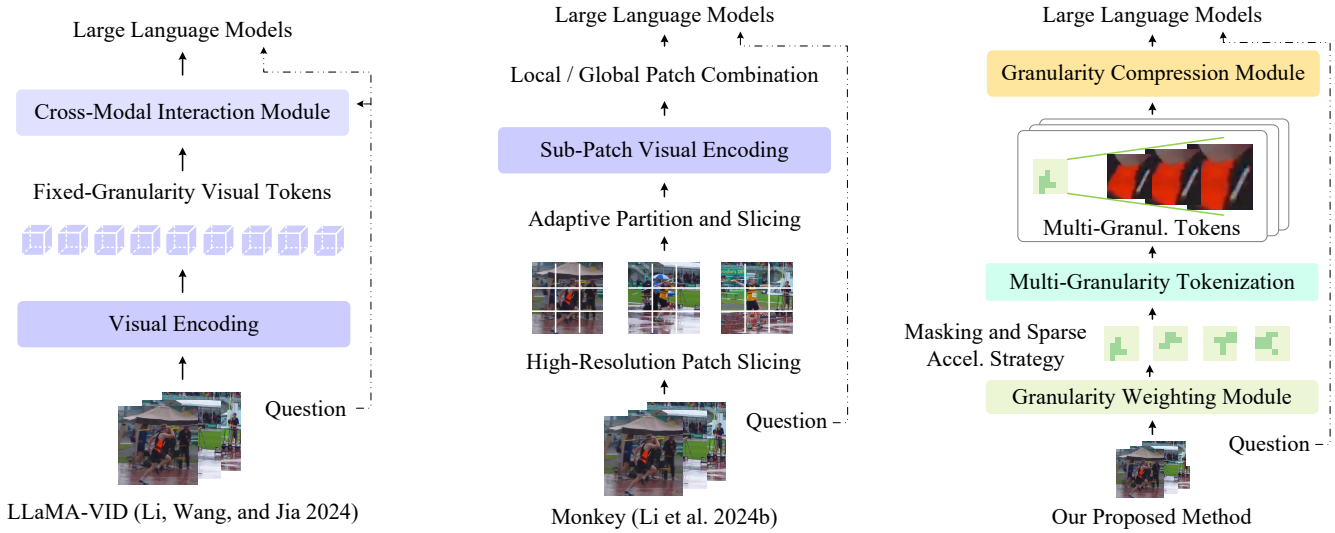

Figure 1: Illustration of the motivation of this paper. Our proposed granularity-adaptive tokenization approach mitigates the loss of local details inherent in previous methods that rely on fixed low-resolution encoding.

In addition, a granularity compression module is designed to dynamically select visual tokens of varying granularities based on the specific question. Extensive experiments on 11 mainstream video question answering datasets verify the effectiveness of the proposed method. We release the code of this work to facilitate future work[1].

Related Work

Multimodal Large Language Models. The outstanding performance of LLMs has sparked interest among researchers in applying them to the multimodal field. Some pioneering approaches propose aligning powerful visual-only and language-only models and achieve superior results (Alayrac et al. 2022; Wang et al. 2023a). With the growing interest in video tasks, researchers have been exploring the extension of LLMs to various video tasks (Li et al. 2024a; Liu et al. 2024b; Cheng et al. 2024a). Many efforts are dedicated to designing effective visual tokenization approaches and aligning video features with LLMs to enhance visual comprehension. For instance, Video-LLaMA (Cheng et al. 2024a) introduced a Video Q-Former to capture temporal changes in visual scenes. However, most existing models encode frames at a fixed, low resolution, limiting their ability to comprehend fine-grained visual content. In contrast, our method proposes to incorporate detailed video content through multi-granular visual tokenizations.

Multi-Granularity Encoding in LLMs. Recent research on LLM-based image understanding has explored to improve the detailed comprehension by using high-resolution images (Li et al. 2024c; Liu et al. 2024a). A common approach involves slicing the high-resolution image into multiple sub-patches, encoding these sub-patches separately, and then inputting them into the LLM for answer prediction. For example, Monkey (Li et al. 2024c) processed input images by dividing them into uniform patches, and LLaVA-UHD (Xu et al. 2024) designed an image modularization strategy that divides native-resolution images into smaller, variable-sized slices for image encoding. However, the existing methods typically require encoding the entire high-resolution image, leading to a substantial computational workload when applied to the video field. In contrast, our approach effectively alleviates the computational burden of encoding high-resolution frames by employing the proposed masking and acceleration strategies.

[1]https://code-website.wixsite.com/videoqa.

Method

Overview

Figure 2 presents an overview of our proposed model, which consists of three key modules: a *granularity weighting module*, a *multi-granularity tokenization module*, and a *granularity compression module*. The granularity weighting module produces spatial activated maps at low resolution based on the input questions. The multi-granularity tokenization module employs the activated patches as queries to query corresponding detailed regions in high-resolution frames. Masking and acceleration techniques are applied to produce fine-grained visual tokens from the high-resolution frames. Finally, the granularity compression module condenses the multi-granularity tokens through question-guided granularity selection.

Granularity Weighting Module

The primary function of the granularity weighting module is to identify spatial activated patches at low resolution guided by the text question. Previous high-resolution image LLMs (Li et al. 2024c; Xu et al. 2024) typically involve encoding entire high-resolution images, thereby enhancing the model's ability to achieve fine-grained understanding. These

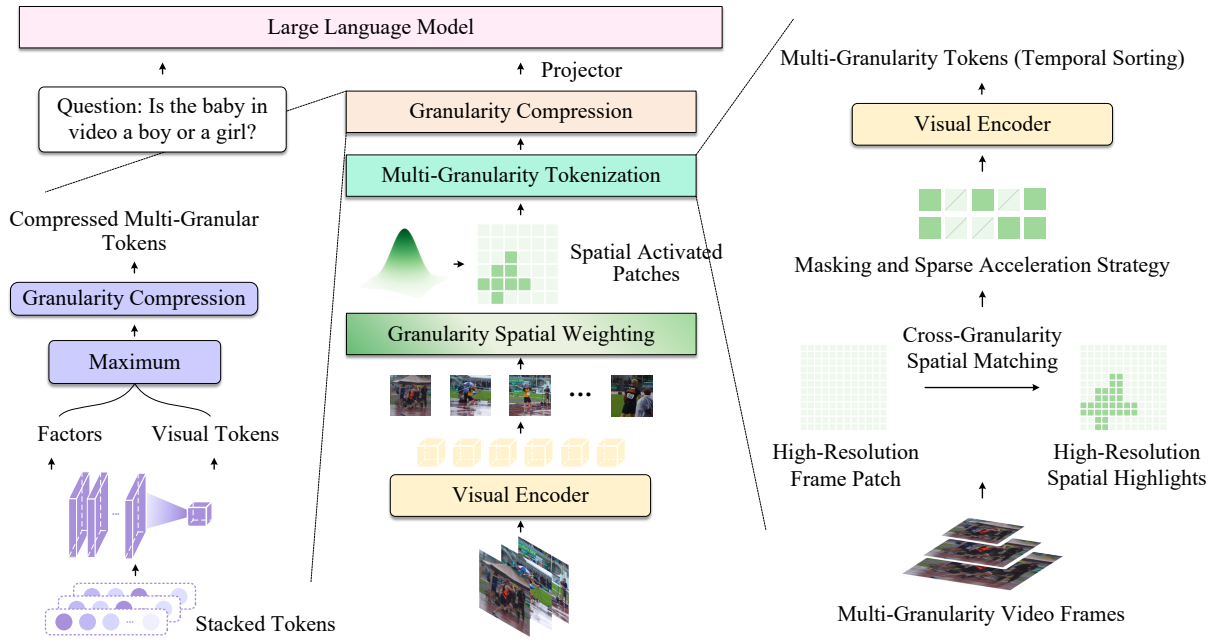

Figure 2: Schematic diagram of the proposed granularity-adaptive spatial evidence tokenization model.

methods, however, lead to the encoding of substantial redundant information that is irrelevant to the specific question, thus wasting computational resources. To address this inefficiency, we propose a granularity weighting module: initially calculating the spatial activated patches necessary for answering the question using the low-resolution frame and then mapping these patches to the high-resolution frame.

Formally, given a video $V = \{v_1, v_2, ..., v_r\}$ and a question $T = \{t_1, t_2, ..., t_s\}$, where r and s represent the number of frames and words respectively, we aim to calculate spatial activated maps $M = \{m_1, m_2, ..., m_r\}$ for each frame. We denote the patches of each frame v_r as $p_r = \{p_r^1, p_r^2, \ldots, p_r^u\}$, where u denotes the number of patches in each frame. These patches p_r are then input into a pre-trained, frozen visual encoder $\mathbf{E}$ to extract features of each frame v_r, resulting in frame visual features $\mathbf{P}_r = \{\mathbf{p}_r^1, \mathbf{p}_r^2, \ldots, \mathbf{p}_r^u\}$. To realize the fusions between visual and textual modalities, following (Li et al. 2024b), an interaction layer is utilized to calculate the fused feature $\mathbf{F}_r$ between frame patches and text question. Afterwards, vector productions are used to obtain the spatial activation of each patch in v_r relative to the question T: $\boldsymbol{d}(r, T) = (\mathbf{M}_{\text{vis}} \cdot \mathbf{F}_r) \cdot (\mathbf{M}_{\text{txt}} \cdot \mathbf{P}_r)^\top$, where $\mathbf{M}_{\text{vis}}$ and $\mathbf{M}_{\text{txt}}$ represent learnable matrices. To produce activated patches in each frame, we select the top $\xi\%$ patches based on $\boldsymbol{d}(r, T)$, and denote these patches as p_r^+. The spatial activated patches $p^+ = \{p_1^+, p_2^+, \ldots, p_r^+\}$ will be input into the multi-granularity tokenization module to perform high-resolution patch queries.

Multi-Granularity Tokenization Module

The primary function of multi-granularity tokenization module is to query the spatial area in high-resolution frames based on the activated map calculated at low resolution. It receives $p_r^+ \in p^+$ produced by the granularity weighting module as input and outputs the visual encoding h_r^+ derived from the high-resolution video frame.

Many previous approaches (Li et al. 2024b; Lin et al. 2023) rely on fixed, relatively low resolutions for frame encoding, which restricts the model's ability to answer questions that pertain to detailed areas of the video. Some advanced high-resolution image LLMs (Li et al. 2024c) divide the entire image into sub-patches and encode them separately, leading to a substantial increase in computational overhead inevitably. To address these issues, we propose a multi-granularity tokenization module. It includes indexing and querying detailed patches based on the spatial activated map, as well as a masking and acceleration strategy for encoding high-resolution frames efficiently. Formally, given p_r^+ produced by the granularity weighting module, we denote the index of spatial activated patches in p_r^+ as set $\mathcal{H}_r$, *i.e.*, $\mathcal{H}_r = \{u \mid p_r^u \in p_r^+, 1 \le u \le u\}, 1 \le r \le r$. Our goal is to query the patch at high resolution $w_h \times w_h$ corresponding to the activated p_r^+ at low resolution $w_l \times w_l$. The elements in set $\mathcal{H}_r$ are queried in the following formula:

$$\begin{aligned} \mathcal{H}_r^{(h)} = \{u^{(h)} \mid u^{(h)} = \ & \varrho(u) \cdot \Theta_w + \zeta(u) \cdot \Theta_h, u \in \mathcal{H}_r\}, \\ \varrho(u) \ = \ & \left\lfloor \frac{u}{w_l/s} \right\rfloor, \ \Theta_w \ = \ \frac{(w_h)^2}{w_l \cdot s}, \\ \zeta(u) \ = \ & u \mod \frac{w_l}{s}, \ \Theta_h \ = \ \frac{w_h}{w_l}, \end{aligned} \tag{1}$$

where $\mathcal{H}_r^{(h)}$ denotes the spatially activated patches at $w_h \times w_h$. $\varrho(u)$ and $\zeta(u)$ represent patch shifts guided by $u \in \mathcal{H}_r$. Θ_w and Θ_h signify shift strides. s indicates the patch size.

In this manner, by using the spatial activated patches in $w_l \times w_l$ as the query, we accomplish the indexing of corre-

sponding patches in $w_h \times w_h$. This approach explicitly highlights significant visual information relevant to text questions, thereby enriching the model's visual encoding with fine-grained details.

Masking and Acceleration Strategy. Given $\mathcal{H}_r^{(h)}$ calculated in Equation 1, the next step is to compute the visual features of the queried patches $p_r^{(h)} = \{p_r^{u^{(h)}} \mid u^{(h)} \in \mathcal{H}_r^{(h)}\}$. A straightforward approach involves using the visual encoder $\mathbf{E}$ to first compute the features of the entire video frame at resolution $w_h \times w_h$, and then using $\mathcal{H}_r^{(h)}$ to index the feature map to extract desired patch features:

$$\begin{aligned} \mathbf{P}_r^{(h)} = \{\mathbf{p}_r^{u^{(h)}} \mid u^{(h)} \in \mathcal{H}_r^{(h)}\}, \mathbf{p}_r^{u^{(h)}} \in \mathbf{v}_r^{(h)}, \\ \mathbf{v}_r^{(h)} = \mathbf{E}\left(v_r^{(h)}\right), \end{aligned} \tag{2}$$

where $v_r^{(h)}$ denotes all visual patches in frame r at $w_h \times w_h$. $\mathbf{v}_r^{(h)}$ represents all extracted patch features of frame r.

In Equation 2, all patches $v_r^{(h)}$ of r are input into the visual encoder $\mathbf{E}$, and then the encoded frame features $\mathbf{v}_r^{(h)}$ are indexed according to the elements in $\mathcal{H}_r^{(h)}$. The advantage lies in incorporating the visual information of all patches during frame encoding, leading to more comprehensive visual features. However, a significant drawback is that many unactivated visual patches $\{p_r^{u^{(h)}} | u^{(h)} \notin \mathcal{H}_r^{(h)}\}$ also participate in the forward propagation process, resulting in substantial computational overhead particularly as the number of frames increases.

To address this issue, we design a masking and acceleration strategy. Considering that the patches activated by the granularity weighting module occupy only a small portion of the entire video frame, and inspired by recent techniques from foundation model pretraining in computer vision tasks (Li et al. 2023c), we mask the spatial patches p_r^- during visual encoding:

$$p_r^- = \{p_r^u \mid \boldsymbol{d}_u(r,T) < \xi, \boldsymbol{d}_u(r,T) \in \boldsymbol{d}(r,T)\}. \tag{3}$$

Consequently, the corresponding query patches $\{p_r^{u^{(h)}} \mid u^{(h)} \notin \mathcal{H}_r^{(h)}\}$ at $w_h \times w_h$ are also masked. We then focus only on the visual information of the activated patches at high-resolution:

$$\begin{aligned} \mathbf{P}_r^{(h)} = \text{Psort}\big(\{\mathbf{p}_r^{u^{(h)}} \mid u^{(h)} \in \mathcal{H}_r^{(h)}\}\big), \mathbf{p}_r^{u^{(h)}} \in \mathbf{p}_r^{(h)}, \\ \mathbf{p}_r^{(h)} = \mathbf{E}\left(p_r^{(h)}\right), \end{aligned} \tag{4}$$

where Psort refers to sorting the patch features in $\mathbf{p}_r^{(h)}$ according to their original order in $v_r^{(h)}$ within the frame. This operation preserves the spatial arrangement within each frame, preventing any semantic alterations that could result from a disordered patch sequence. Additionally, during the visual encoding of $\mathbf{p}_r^{(h)}$, we employ relative position embedding (Li et al. 2023c) in frames. The relative position of each patch is determined based on its position prior to masking:

$$\mathbf{R}\big(p_r^{u^{(h)}}\big) = \mathbf{R}\big(v_r^{u^{(h)}}\big), \ \forall\, u^{(h)} \in \mathcal{H}_r^{(h)}, \tag{5}$$

where $\mathbf{R}$ denotes the position encoding matrix.

Unlike previous works that encode entire high-resolution images for image understanding (Li et al. 2024c), our strategy sheds light on the proposed multi-granularity visual tokenization idea and offers several advantages. It eliminates the need to encode the entire high-resolution frame and ensures only the desired visual information is activated. This approach reduces computational overhead while maintaining essential visual details for predictions.

Granularity Compression Module

Given the multi-granularity visual representations $\mathbf{P}_r$ and $\mathbf{P}_r^{(h)}$ of frame r, our next step is to input them into the LLM for model prediction. However, before doing so, we note that the granularity of visual tokens used for prediction should depend on the nature of the text question. Some questions focus on the coarse-grained global semantic of the video, while others require fine-grained details of specific visual patches. For example, a question like "How many times does the girl appear in the video?" necessitates only coarse-grained visual cues, making detailed spatial patch activation unnecessary. Conversely, a question like "What is the girl holding in her hand?" requires locating specific details within the frame, thus benefiting from fine-grained signals to predict the answer. Based on the considerations, we propose a granularity compression module to perform visual token selection and compression across various granular tokens.

Formally, we employ the coarse-grained frame token $\mathbf{P}_r$, the fine-grained frame token $\mathbf{P}_r^{(h)}$, and the spatially activated tokens $\mathbf{P}_r^+$ to construct the candidate set $\mathcal{Q}_r$ for visual compression: $\mathcal{Q}_r = \left[\mathbf{P}_r || \mathbf{P}_r^{(h)} || \mathbf{P}_r^+\right]$, where $||$ represents the concatenation operation, and $\mathbf{P}_r^+$ signifies the spatial patches activated by p_r^+: $\mathbf{P}_r^+ = \{\mathbf{p}_r^u \mid p_r^u \in p_r^+, \mathbf{p}_r^u \in \mathbf{P}_r\}$. We leverage the fused feature $\mathbf{F}_r$ between the text and frames as the query, and calculate the compression factors for multi-granular tokens using a simple vector product operation: $\mathbf{F}_{r \leftarrow \mathcal{Q}_r} = \gamma(\text{Pooling}(\mathcal{Q}_r) \cdot \mathbf{F}_r)$, where γ refers to the Softmax function, which is used to normalize the factors for each granularity. Next, we use $\mathbf{F}_{r \leftarrow \mathcal{Q}_r}$ to select tokens of each granularity and apply max pooling operations to obtain the spatially compressed visual tokens:

$$\mathbf{V}_{spatial}^r = \max_{\mathcal{Q}_r} \left(\mathbf{F}_{r \leftarrow \mathcal{Q}_r} \cdot \text{Pooling}(\mathcal{Q}_r)\right). \tag{6}$$

We adopt the maximum operation here to retain the most salient visual information for each granularity based on $\mathbf{F}_{r \leftarrow \mathcal{Q}_r}$. Finally, the concatenation of $\mathcal{Q}_r$ and $\mathbf{V}_{spatial}^r$ are fed into the LLM for model prediction.

During training, next token prediction is employed to optimize the model and calculate loss (Chiang et al. 2023).

Experiments

Datasets and Evaluation Protocols

We evaluate the model performance on 11 mainstream video question answering datasets, including 4 open-ended question answering benchmarks (Xu et al. 2017; Yu et al. 2019; Li et al. 2016), a text generation benchmark (Maaz et al. 2024), 5 multiple-choice benchmarks (Xiao et al. 2021; Wu

Method	LLM Size	MSVD-QA		MSRVTT-QA		ActivityNet-QA		TGIF-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
Video-LLaMA (Cheng et al. 2024a)	7B	51.6	2.5	29.6	1.8	12.4	1.1	-	-
LLaMA-Adapter (Zhang et al. 2024)	7B	54.9	3.1	43.8	2.7	34.2	2.7	-	-
Video-ChatGPT (Maaz et al. 2024)	7B	64.9	3.3	49.3	2.8	35.2	2.7	51.4	3.0
Video-LLaVA (Lin et al. 2023)	7B	70.7	**3.9**	59.2	**3.5**	45.3	3.3	70.0	4.0
LLaMA-VID (Li et al. 2024b)	13B	70.0	3.7	58.9	3.3	47.5	3.3	-	-
Chat-UniVi (Jin et al. 2024a)	7B	65.0	3.6	54.6	3.1	45.8	3.2	60.3	3.4
BT-Adapter (Liu et al. 2024b)	7B	67.5	3.7	57.0	3.2	45.7	3.2	-	-
VideoChat2 (Li et al. 2024a)	7B	70.0	**3.9**	54.1	3.3	49.1	3.3	-	-
Video-LaVIT (Jin et al. 2024b)	7B	73.2	**3.9**	59.3	3.3	50.1	3.3	-	-
MovieLLM (Song et al. 2024c)	7B	63.2	3.5	52.1	3.1	43.3	3.3	-	-
MiniGPT4-V (Ataallah et al. 2024)	7B	72.9	3.8	58.8	3.3	45.9	3.2	67.9	3.7
VideoLLaMA2 (Cheng et al. 2024b)	7B	71.7	**3.9**	-	-	49.9	3.3	-	-
VideoLLaMA2 (Cheng et al. 2024b)	8 × 7B	70.5	3.8	-	-	50.3	**3.4**	-	-
Ours	7B	**73.4**	**3.9**	**59.7**	3.3	**51.4**	**3.4**	**74.9**	**4.1**

Table 1: Performance comparisons with baselines on open-ended question answering benchmarks.

Method	CI	DO	CU	TU	CO	Avg.
Video-LLaMA (Cheng et al. 2024a)	1.96	2.18	2.16	1.82	1.79	1.98
LLaMA-Adapter (Zhang et al. 2024)	2.03	2.32	2.30	1.98	2.15	2.16
Video-ChatGPT (Maaz et al. 2024)	2.50	2.57	2.69	2.16	2.20	2.42
VideoChat (Li et al. 2023b)	2.23	2.50	2.53	1.94	2.24	2.29
Chat-UniVi (Jin et al. 2024a)	2.89	2.91	3.46	2.40	**2.81**	2.89
Video-LLaVA (Lin et al. 2023)	2.87	2.94	3.44	2.45	2.51	2.84
MovieChat (Song et al. 2024a)	2.76	2.93	3.01	2.24	2.42	2.67
BT-Adapter (Liu et al. 2024b)	2.68	2.69	3.27	2.34	2.46	2.69
LLaMA-VID (Li et al. 2024b)	2.96	3.00	3.53	2.46	2.51	2.89
Vista-LLaMA (Ma et al. 2024)	2.44	2.64	3.18	2.26	2.31	2.57
MovieChat+ (Song et al. 2024b)	2.87	2.95	3.10	2.25	2.50	2.73
MovieLLM (Song et al. 2024c)	2.64	2.61	2.92	2.03	2.43	2.53
MiniGPT4-V (Ataallah et al. 2024)	2.93	2.97	3.45	2.47	2.60	2.88
Ours	**3.07**	**3.04**	**3.62**	**2.56**	2.78	**3.01**

Table 2: Performance on text generation benchmark.

et al. 2023; Lei et al. 2018; Li et al. 2023a; Mangalam, Akshulakov, and Malik 2023), and a recently proposed comprehensive video understanding benchmark (Li et al. 2024a). GPT-3.5 is employed to assess open-ended question answering and text generation, with accuracy and score as metrics. For the multiple-choice benchmarks, accuracy is used as the evaluation metric.

Implementation Details

We use Vicuna-7B (Chiang et al. 2023) as the LLM, and the visual encoder is EVA-G (Sun et al. 2023). Pre-trained Q-Former in InstructBLIP (Dai et al. 2023) is employed for feature fusion between frames and questions. w_l and w_h are 224 and 448 respectively. Two-layer MLPs are used to project visual tokens into LLM semantic space. During the pre-training stage, only the projection layer is trained, while in the instruction tuning phase, LLM, Q-Former, and projection layer are trained. Following (Li et al. 2024b), a total of 790K pairs are used in pre-training and 763K samples are utilized for instruction tuning. ξ is set to 0.4.

Model	TVQA	MVBench	IntentQA		
			Cau.	Tem.	Avg.
repl./ Monkey	36.1	41.0	59.0	47.4	56.0
Ours	**39.8**	**45.1**	**64.5**	**52.6**	**61.5**

Model	Text Generation Benchmark				
	CI	DO	CU	TU	CO
repl./ Monkey	2.96	3.00	3.57	2.45	2.61
Ours	**3.07**	**3.04**	**3.62**	**2.56**	**2.78**

Table 3: Comparison of our proposed method with high-resolution image LLM baseline (Li et al. 2024c).

Comparisons on Mainstream Benchmarks

Table 1, 2, 4, 7 present performance comparisons across open-ended QA, multiple-choice QA, text generation, and recently proposed video understanding benchmarks. In Table 2 and Table 7, we exclude baselines such as VideoChat2 and VideoLLaMA2 since they leverage significantly larger and more diverse datasets (pretraining: 12.2M samples from 6 datasets *v.s.* 790K samples from 2 datasets; instruction tuning: 2M samples from 34 datasets *v.s.* 763K samples from 6 datasets). The experimental results indicate that our model outperforms the baselines, which can be attributed to the proposed multi-granularity spatial evidence tokenization method. Our approach emphasises specific detailed areas in high-resolution frames, thereby integrating fine-grained information and enhancing model performance. Figure 4 illustrates the qualitative experimental results.

In addition, we provide comparisons with high-resolution LLM baseline, Monkey (Li et al. 2024c), in Table 3. Experimental results verify the effectiveness of our approach.

Ablation Experiments

Table 6 presents the ablation experimental results, where "w/o W.", "w/o C.", and "w/o G." correspond to the removal

Method	LLM	NExT-QA				STAR					TVQA	IntentQA			Ego-S.
		Cau.	Tem.	Des.	Avg.	Int.	Seq.	Pre.	Fea.	Avg.		Cau.	Tem.	Avg.	
Video-LLaVA (Lin et al. 2023)	7B	-	-	-	-	-	-	-	-	-	-	-	-	-	38.4
LLaMA-VID (Li et al. 2024b)	7B	59.3	53.9	70.9	59.4	38.5	41.4	38.3	37.8	39.9	36.2	59.9	50.3	57.4	38.5
Vista-LLaMA (Ma et al. 2024)	7B	-	-	-	60.7	-	-	-	-	-	-	-	-	-	-
Sevilla (Yu et al. 2024)	2.85B	61.3	**61.5**	75.6	63.6	48.3	45.0	44.4	40.8	44.6	38.2	-	-	60.9	-
ViperGPT (Surís et al. 2023)	GPT-3	-	-	-	60.0	-	-	-	-	-	-	-	-	-	-
VideoChat2 (Li et al. 2024a)	7B	61.9	57.4	69.9	61.7	-	-	-	-	-	**40.6**	-	-	-	42.2
Video-LaVIT (Jin et al. 2024b)	7B	-	-	-	-	-	-	-	-	-	-	-	-	-	37.3
MiniGPT4-Video (Ataallah et al. 2024)	7B	-	-	-	-	-	-	-	-	-	36.5	-	-	-	-
LLaVA-NeXT-Video (Liu et al. 2024a)	7B	-	-	-	-	-	-	-	-	-	-	-	-	-	43.9
VideoLLaMA2 (Cheng et al. 2024b)	7B	-	-	-	-	-	-	-	-	-	-	-	-	-	**50.0**
Ours	7B	**62.6**	60.9	**76.3**	**64.2**	**49.0**	**49.7**	**45.4**	**44.7**	**48.8**	39.8	**64.5**	**52.6**	**61.5**	42.8

Table 4: Performance comparison on multiple-choice question answering.

Model	NExT-QA				STAR	TVQA	IntentQA		
	Cau.	Tem.	Des.	Avg.			Cau.	Tem.	Avg.
PR #1	62.6	60.4	75.8	64.0	48.4	39.7	63.5	53.5	60.9
PR #2	59.9	57.9	76.1	61.8	49.5	39.0	62.4	52.6	59.9
PR #3	62.4	60.4	75.9	63.9	48.4	39.5	63.5	53.7	61.0
PR #4	62.2	60.7	76.2	63.9	48.6	39.5	63.3	54.1	60.9
INS #1	61.4	58.8	75.7	62.8	48.8	39.5	62.8	51.8	60.0
INS #2	62.4	60.7	76.7	64.1	49.1	39.8	63.3	53.3	60.8

Table 5: Experiments on various question prompts and system instructions on the performance of model predictions.

Model	MSVD-QA		ActivityNet-QA		TGIF-QA	
	Acc.	Score	Acc.	Score	Acc.	Score
w/o W.	72.0	**3.9**	49.7	**3.4**	72.4	4.0
w/o C.	72.5	**3.9**	49.7	3.3	73.8	4.0
w/o G.	71.9	3.8	49.8	3.3	74.0	4.0
Ours	**73.4**	**3.9**	**51.4**	**3.4**	**74.9**	**4.1**

Table 6: Experimental results of ablation studies.

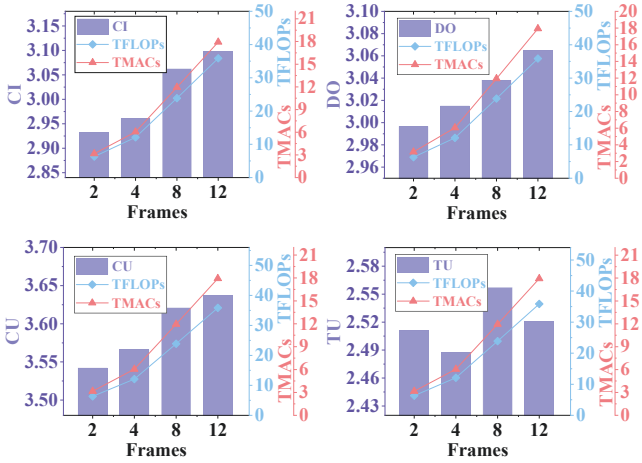

Figure 3: Experiments on text generation performance and model complexity with varying numbers of video frames.

of the granularity weighting module, granularity compression module, and multi-granularity tokenization module, respectively. The results verify the effectiveness of each designed module. Table 5 illustrates the effect of various question prompts and system instructions on model performance. PR #1: "Please only answer the best option's letter from the given choices." PR #2: "Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Based on your observations, select the best option that accurately addresses the question. Only give the best option." PR #3: "Answer with the option's letter from the given choices directly and only give the best option. Let's think step by step." PR #4: "Answer with the option's letter from the given choices directly and only give the best option. Please be critical." INS #1: "Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Based on your observations, select the best option that accurately addresses the question." INS #2: "A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons." Experimental results show that the model's performance may vary slightly with different prompts, though the overall impact remains minimal. For different instructions, INS #2 outperforms INS #1, likely because INS #2 is more closely aligned with the instructions used during training, leading to better performance.

Hyperparameter study. Figure 3 illustrates the effect of numbers of frames on text generation performance. The results indicate that a higher number of frames generally leads to improved performance, whereas fewer frames (*e.g.*, 2 frames) tend to result in suboptimal outcomes. Figure 5 shows a qualitative experiment of the activated patches at low resolution with a ratio of 0.4.

Complexity Analysis. Table 8 provides an analysis of model complexity on ActivityNet-QA. The term "w/o accl."

Model	LM	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	Avg.
LLaMA-Adapter	7B	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	25.0	31.5	22.5	28.0	32.0	31.7
BLIP-2	3B	24.5	29.0	33.5	17.0	42.0	51.5	26.0	31.0	25.5	26.0	32.5	25.5	30.0	40.0	42.0	27.0	30.0	26.0	37.0	31.0	31.4
Otter-I	7B	34.5	32.0	39.5	30.5	38.5	48.5	44.0	29.5	19.0	25.5	55.0	20.0	32.5	28.5	39.0	28.0	27.0	32.0	29.0	36.5	33.5
MiniGPT-4	7B	16.0	18.0	26.0	21.5	16.0	29.5	25.5	13.0	11.5	12.0	9.5	32.5	15.5	8.0	34.0	26.0	29.5	19.0	9.9	3.0	18.8
InstructBLIP	7B	20.0	16.5	46.0	24.5	46.0	51.0	26.0	37.5	22.0	23.0	46.5	42.5	26.5	40.5	32.0	25.5	30.0	25.5	30.5	38.0	32.5
LLaVA	7B	28.0	39.5	63.0	30.5	39.0	53.0	41.0	41.5	23.0	20.5	45.0	34.0	20.5	38.5	47.0	25.0	36.0	27.0	26.5	42.0	36.0
Otter-V	7B	23.0	23.0	27.5	27.0	29.5	53.0	28.0	33.0	24.5	23.5	27.5	26.0	28.5	18.0	38.5	22.0	22.0	23.5	19.0	19.5	26.8
mPLUG-Owl-V	7B	22.0	28.0	34.0	29.0	29.0	40.5	27.0	31.5	27.0	23.0	29.0	31.5	27.0	40.0	44.0	24.0	31.0	26.0	20.5	29.5	29.7
VideoChatGPT	7B	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	29.0	33.0	29.5	26.0	35.5	32.7
Video-LLaMA	7B	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0	34.1
VideoChat	7B	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0	35.5
LLaMA-VID	7B	46.0	42.0	62.5	37.5	55.0	54.5	40.5	33.0	21.0	26.5	87.5	47.5	23.5	43.5	43.0	29.0	42.0	34.5	40.0	34.5	42.2
Video-LLaVA	7B	41.0	47.0	54.5	41.5	54.5	51.0	46.0	34.5	32.0	28.5	86.5	42.0	22.5	52.0	40.5	30.0	45.0	32.0	40.5	37.0	42.9
GPT-4V	-	55.5	63.5	72.0	46.5	73.5	18.5	59.0	29.5	12.0	40.5	83.5	39.0	12.0	22.5	45.0	47.5	52.0	31.0	59.0	11.0	43.5
Ours	7B	49.5	59.5	57.0	46.0	54.5	53.5	60.0	35.0	25.5	27.5	84.0	40.5	34.5	44.5	40.5	42.5	45.5	28.5	40.5	35.0	**45.1**
VideoChat2	7B	66.0	47.5	83.5	49.5	60.0	58.0	71.5	42.5	23.0	23.0	88.5	39.0	42.0	58.5	44.0	49.0	36.5	35.0	40.5	65.5	51.1
Ours w./ Inst.	7B	53.7	68.0	71.0	47.0	52.5	75.3	64.5	33.0	20.0	30.5	82.5	36.0	60.5	72.5	48.5	42.0	45.0	34.0	38.0	55.0	**51.5**

Table 7: Performance on MVBench. Gray rows: Models adopt more instruction tuning data (2M *v.s.* 763K).

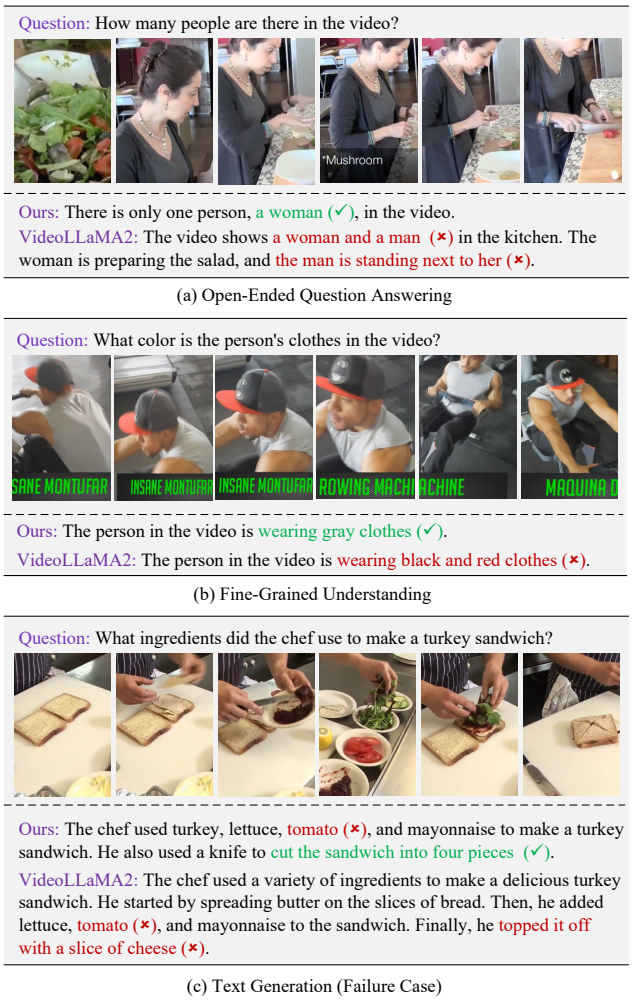

Figure 4: Qualitative experimental results on open-ended question answering and text generation benchmarks.

Model	TMACs	TFLOPs	ΔParams Num
LLaMA-VID	16.84	33.70	**217.04M**
Ours w/o Accl.	26.69	53.40	-
Ours	**11.97**	**23.94**	222.81M

Table 8: Analysis on model complexity.

Figure 5: Qualitative experiments on the granularity weighting module, illustrating the spatial activated patches.

denotes the absence of the proposed acceleration strategy. The results show that our multi-granularity encoding approach requires less computational effort and exhibits reduced model complexity.

Conclusion

In this paper, we propose a granularity-adaptive spatial evidence tokenization method that enhances visual encoding by highlighting specific regions at multiple granularity levels, thereby improving the model's ability to capture fine-grained details. To mitigate the complexity of encoding high-resolution frames, we incorporate masking and acceleration strategies. Extensive experiments on 11 mainstream datasets validate the effectiveness of our proposed method.

Acknowledgements

This work is supported by a grant from Kuaishou (DJHL-20240809-115) and an internal grant of PKU (2024JK28).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhu, D.; Ding, J.; and Elhoseiny, M. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024a. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024b. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Jiang, H.; and Mu, Y. 2022. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16388–16398.

Jiang, H.; Yizhang, Y.; and Mu, Y. 2024. Transferable Video Moment Localization by Moment-Guided Query Prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2516–2524.

Jin, P.; Takanobu, R.; Zhang, W.; Cao, X.; and Yuan, L. 2024a. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13700–13710.

Jin, Y.; Sun, Z.; Xu, K.; Chen, L.; Jiang, H.; Huang, Q.; Song, C.; Liu, Y.; ZHANG, D.; Song, Y.; et al. 2024b. Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization. In *Forty-first International Conference on Machine Learning*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1369–1379.

Li, J.; Wei, P.; Han, W.; and Fan, L. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11963–11974.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023c. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23390–23400.

Li, Y.; Song, Y.; Cao, L.; Tetreault, J.; Goldberg, L.; Jaimes, A.; and Luo, J. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4641–4650.

Li, Y.; Wang, C.; Jia, J.; et al. 2024b. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024c. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26763–26773.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, R.; Li, C.; Ge, Y.; Li, T. H.; Shan, Y.; and Li, G. 2024b. BT-Adapter: Video Conversation is Feasible Without Video Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13658–13667.

Ma, F.; Jin, X.; Wang, H.; Xian, Y.; Feng, J.; and Yang, Y. 2024. VISTA-LLAMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13151–13160.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024a. Moviechat:

From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Song, E.; Chai, W.; Ye, T.; Hwang, J.-N.; Li, X.; and Wang, G. 2024b. MovieChat+: Question-aware Sparse Memory for Long Video Question Answering. *arXiv preprint arXiv:2404.17176*.

Song, Z.; Wang, C.; Sheng, J.; Zhang, C.; Yu, G.; Fan, J.; and Chen, T. 2024c. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Surís, D.; Menon, S.; Vondrick, C.; et al. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11888–11898.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023a. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wang, Y.; Zhang, R.; Wang, H.; Bhattacharya, U.; Fu, Y.; and Wu, G. 2023b. Vaquita: Enhancing alignment in llm-assisted video understanding. *arXiv preprint arXiv:2312.02310*.

Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2024. VideoGrounding-DINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18909–18918.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2023. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xiao, J.; Yao, A.; Li, Y.; and Chua, T.-S. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13204–13214.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Xu, R.; Yao, Y.; Guo, Z.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.-S.; Liu, Z.; Sun, M.; and Huang, G. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9127–9134.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2024. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *International Conference on Learning Representations*.

Zhu, W.; Lu, J.; Li, J.; and Zhou, J. 2020. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30: 948–962.