LLMs Meet Long Video: Advancing Long Video Question Answering with An Interactive Visual Adapter in LLMs

Anonymous ACL submission

Abstract

Long video understanding is a significant and ongoing challenge in the intersection of multimedia and artificial intelligence. Employing large language models (LLMs) for comprehending video becomes an emerging and promising method. However, this approach incurs high computational costs due to the extensive array of video tokens, experiences reduced visual clarity as a consequence of token aggregation, and confronts challenges arising from irrelevant visual tokens while answering video-related questions. To alleviate these issues, we present an *Interactive Visual Adapter (IVA)* within LLMs, designed to enhance interaction with fine-grained visual elements. Specifically, we first transform long videos into temporal video tokens via leveraging a visual encoder alongside a pretrained causal transformer, then feed them into LLMs with the video instructions. Subsequently, we integrated IVA, which contains a lightweight temporal frame selector and a spatial feature interactor, within the internal blocks of LLMs to capture instruction-aware and fine-grained visual signals. Consequently, the proposed video-LLM facilitates a comprehensive understanding of long video content through appropriate long video modelling and precise visual interactions. We conduct extensive experiments on nine video understanding benchmarks and experimental results show that our interactive visual adapter significantly improves the performance of video LLMs on long video QA tasks. Ablation studies further verify the effectiveness of IVA in long and short video understandings.

1 Introduction

The exponential advancement of the Internet and multimedia technologies has resulted in a significant surge in video content production by individuals and enterprises across various domains. The ability to interpret and extract meaningful content from videos is increasingly vital for meeting human demands and promoting the speed of information dissemination (Tang et al., 2023). Therefore, Video Question Answering (Yu et al., 2019; Li et al., 2023b; Castro et al., 2022) (Video QA), which allows users to ask about the content of videos through natural language and receive answers derived from their visual and auditory content, attracts tremendous research interest. Recently, large language models (LLMs) (OpenAI, 2023; Chiang et al., 2023) have demonstrated exceptional efficacy in the domains of human-machine interaction and the handling of extensive contextual information. Capitalizing on these advancements, there is a burgeoning inclination towards integrating LLMs into the realm of video information processing. This approach primarily aims to enhance the interface between users and video content through intelligent question-and-answer (QA) sessions.

The core of this innovation is a strategy that bridges the gap between the visual information in videos and the textual comprehension capabilities of LLMs. This is accomplished through a meticulously designed process that translates video data into a format comprehensible by LLMs, thereby facilitating an advanced question-answering system tailored for video content. The process involves mapping video encoding into the language space of LLMs via a learnable visual mapping network (Wu et al., 2023; Li et al., 2023d; Dai et al., 2023). Essentially, the video is converted into "video tokens", which are then fed into the LLM along with textual tokens of natural language questions. Leveraging the vast knowledge storage and natural language processing prowess of LLMs, this approach effectively handles video QA tasks. For instance, Maaz et al. (2023) performs spatial and temporal pooling for video tokens and feeds them into Vicuna (Chiang et al., 2023) to achieve the interaction between users and video content. Zhang et al. (2023b) utilizes Q-former (Li et al., 2023d) to extract question-relevant video tokens, which are then fed into LLaMA (Gao et al., 2023) to generate the answer.

These LLMs-powered video understanding models (Tang et al., 2023; Song et al., 2023) mainly focus on short video modelling and have achieved a successful performance on short video captioning (Zhang et al., 2023c), question-answering (Jin et al., 2023), and summarization (Tang et al., 2023). However, the core challenges of video processing (Xu et al., 2023) stem from the need to efficiently model long video sequences and precisely respond to questions relevant to the video. Generally, using LLMs to handle long-form video often encounters the following hurdles: *1) high computational costs from a multitude of video tokens; 2) reduced visual clarity as a consequence of token aggregation such as employing average or maximum representation pooling for visual frames; 3) irrelevant visual tokens leading to incorrect answers, notably when question-relevant information is embedded within long temporal cues.* Hence, previous models struggle to handle long-form videos owing to the constrained input capacity for video tokens and the challenge of distilling question-relevant, fine-grained visual features during generation.

To alleviate these issues, we present a long video comprehension method for LLMs, named *Interactive Visual Adapter (IVA)* to achieve in-depth interactions between LLMs and video content. Specifically, we first use the pretrained visual encoder to obtain global and fine-grained frame representations. We construct the temporal video tokens by integrating the global features of frames with temporal video embeddings, which are obtained through a pretrained causal transformer. The whole set of temporal video tokens is fed into the LLM to attain a whole understanding of the video content. Additionally, we design a parameters-sharing Interactive Visual Adapter (IVA) that contains an instruction-aware temporal frames selector and a spatial feature interactor. The selector is used to obtain question-relevant frames based on contextual query embeddings and global encodings of videos. The selected frames are then fed into the spatial interactor to engage with the contextual query embeddings, in which fine-grained representations of frames are used. By doing so, LLMs could achieve in-depth interaction with video content by applying IVA between different layers.

To verify the effectiveness of our method, we conduct extensive experiments on four long video QA and five short video understanding benchmarks. Experimental results indicate that IVA is capable of achieving effective interactions between LLMs and long or short videos. Our contributions are summarized as follows:

- We analyze the challenges of modelling long videos for LLMs and propose an interactive visual adapter for LLMs to handle long videos. It realizes the in-depth interaction between LLMs and long videos based on efficient video tokens and the IVA mechanism.

- The proposed IVA is capable of selecting relevant frames and interacting with their fine-grained spatial features through the internal selector and interactor, respectively. The IVA architecture is lightweight and designed to be shareable between layers of LLMs.

- Experimental results show that LLMs with IVA could achieve powerful performances in long video QA. Ablation studies underscore the critical role and effectiveness of IVA, confirming its significant contribution to enhanced performance.

2 Related Work

Traditional Video Understanding Models The rapid development of deep learning methods possesses superior task-solving capabilities for video understanding. DeepVideo (Karpathy et al., 2014) was the earliest method introducing a Convolutional Neural Network (CNN), for video understanding. Two-stream networks (Feichtenhofer et al., 2016), then integrating Convolutional Neural Networks (CNNs) (Feichtenhofer et al., 2016) and Improved Dense Trajectories (IDT) (Li et al., 2021), enhanced motion analysis in video understanding. For long-form content, Long Short-Term Memory (LSTM) (Yue-Hei Ng et al., 2015) networks were adopted, offering a robust solution for sequential data analysis over extended durations. Additionally, Temporal Segment Network (TSN) (Wang et al., 2016) advanced long-form video understanding by segmenting videos for individual analysis before aggregating insights, enabling more nuanced interpretation. Meanwhile, 3D networks started another branch by introducing 3D CNN to video understanding (C3D) (Tran et al., 2015). The introduction of Vision Transformers (ViT) (Dosovitskiy

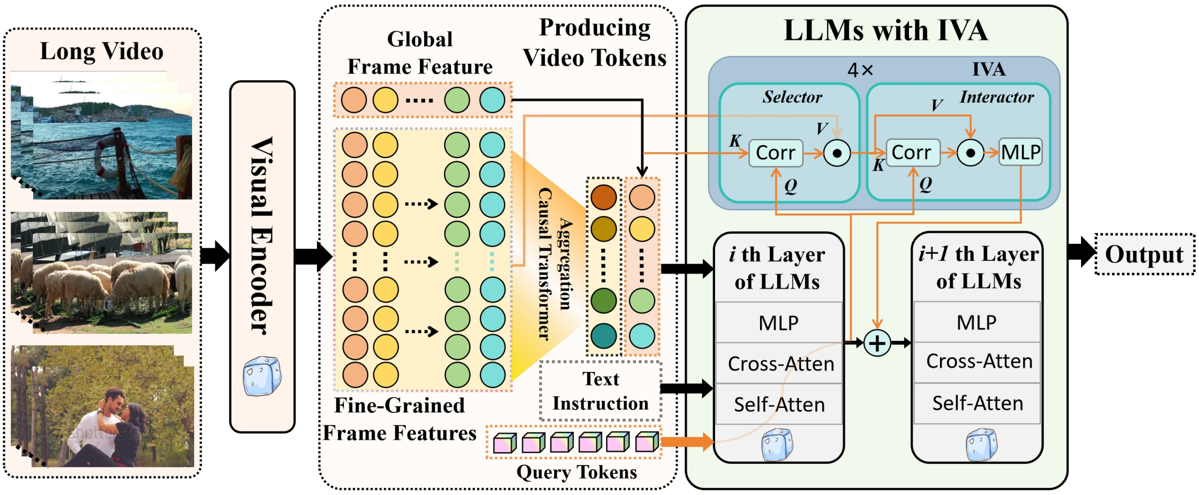


Figure 1: The overview of our framework employing LLMs to handle long video. While producing video tokens, we combine the global features and aggregated fine-grained features to represent a frame, allocating two tokens for each frame. The causal transformer is used to capture temporal relationships across frames and its output will be spliced with spatial feature sequence. The IVA will be inserted between blocks of LLMs to incorporate fine-grained visuals based on an understanding of the long video tokens, text instructions, and query tokens.

et al., 2021; Arnab et al., 2021; Fan et al., 2021) promotes a series of prominent models Among the pioneering efforts in this self-supervised video training domain, VideoBERT (Sun et al., 2019) leverages the bidirectional language model BERT (Kenton and Toutanova, 2019) for self-supervised learning from video-text data. This model, and others following the "pre-training and fine-tuning" paradigm, such as ActBERT (Zhu and Yang, 2020), SpatiotemporalMAE (Feichtenhofer et al., 2022), OmniMAE (Girdhar et al., 2023), showcase the diverse strategies developed to enhance video understanding. Notably, these models have set a foundation for advanced video-language models like maskViT (Gupta et al., 2022), CLIP-ViP (Xue et al., 2022), LF-VILA (Sun et al., 2022), further pushing the boundaries of what's achievable in action classification, video captioning, and beyond. The evolution from VideoBERT to more recent innovations like HiTeA (Ye et al., 2023), and CHAMPAGNE (Han et al., 2023) underscores the rapid advancement in this field.

LLMs for Video Understanding The recent advancement in large language models (LLMs), pre-trained on expansive datasets, has ushered in groundbreaking capabilities in in-context learning (Zhang et al., 2023a) and long-form context modeling (Lyu et al., 2023). This innovation has paved the way for integrating LLMs with computer vision technologies, exemplified by initiatives like Visual-ChatGPT (Wu et al., 2023). These models transcend traditional boundaries by calling vision model APIs (Qin et al., 2023), thereby addressing complex problems within the computer vision domain. Integrating language models with video understanding technologies (Maaz et al., 2023; Zhang et al., 2023d; Li et al., 2023e; Xu et al., 2023; Song et al., 2023) enhances multimodal understanding, facilitating sophisticated processing and interpretation of the intricate interplay between visual and textual data (Ouyang et al., 2022; Liu et al., 2023; Muennighoff et al., 2022; Li et al., 2023g; Zhu et al., 2023; Zhang et al., 2023d). They leverage their extensive multimodal knowledge base and nuanced contextual understanding, mirroring a more human-like comprehension of video. Moreover, the exploration of LLMs in video understanding tasks (Tang et al., 2023) represents a significant stride in analyzing and reasoning about visual data.

3 Methodology

3.1 Overview

Our work primarily introduces an interactive visual adapter for LLMs to handle long videos and answer relevant questions. The overview of workflow is shown in Figure 1. Specifically, given a video V, we first extract frames to obtain the whole sequence frame representations $\mathbf{h}_V = (\mathbf{h}_{I_1}, ..., \mathbf{h}_{I_k}, ..., \mathbf{h}_{I_N})$ via the pretrained image encoder, where $\mathbf{h}_{I_k} = (h_g^{I_k}, h_1^{I_k} ..., h_{576}^{I_k})$ refers to the representations of k th frame and N is the total number of extracted frames. Then, we use a causal transformer to acquire temporal video embeddings from the aggregated spatial representation. The overall video tokens are formed by merging temporal video embeddings and global spatial features $[h_g^{I_1}, h_g^{I_k}, ..., h_g^{I_N}]$, where each frame is represented by two tokens. To enhance the capability of LLMs in leveraging fine-grained visual details from videos, we have developed an Interactive Visual Adapter (IVA) that is integrated into the blocks of LLM. This integration allows LLMs to comprehend the entirety of long videos through efficient video tokens while simultaneously capturing fine-grained visual information facilitated by the IVA.

3.2 Producing Video Tokens

We elaborate on the detailed process employed to produce efficient tokens for long videos, characterized by the extraction of one frame per second. First, we use the self-weighted calculation on the fine-grained feature $\mathbf{h}_f^k = (h_1, ..., h_{576})$ of a frame (k th) to obtain its overall representation, which will be fed into the following causal transformer. This calculation process for the k th frame is given in the following Eq. 1:

$$\begin{aligned} s_f^k &= Softmax(\mathbf{W}(\mathbf{h}_f^k) + \mathbf{b}), \\ \mathbf{h}_t^k &= s_f \mathbf{h}_f^k, \end{aligned} \tag{1}$$

where $s_f \in \mathbf{R}^{1\times 576}$ is the weight distribution and $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters. Hence, we denote the obtained sequence-level frame representation as $\mathbf{h}_f = (\mathbf{h}_f^1, ..., \mathbf{h}_f^N)$.

Causal Transformer is employed to acquire the temporal video embeddings. Specifically, we use a four-layer transformer to facilitate interaction across frames, where a frame only attends to its previous ones. Take the first layer as an example, the specific operation of the causal transformer is presented in Eq. 2:

$$\begin{aligned} \mathbf{h}_s &= SelfAtten(LayerN(\mathbf{h}_f), W^{Mask}) + \mathbf{h}_f \\ \mathbf{h}_s &= LayerN(\mathbf{h}_s), \\ \mathbf{h}_o^1 &= MLP(\mathbf{h}_s) \end{aligned} \tag{2}$$

where $SelfAtten$ and $LayerN$ are the self-attention calculation and the feature normalization. The top output of the causal transformer will be projected into the language model by a linear layer, which is spliced with the global features $\mathbf{h}_g^V = (h_g^{I_1}, h_g^{I_k}, ..., h_g^{I_N})$. These global features of frames will be transferred into language models via a learnable MLP. We denote the final spliced feature to $\mathbf{h}_V = (h_V^1, h_V^2, ..., h_V^{2N})$.

3.3 Interactive Visual Adapter

After obtaining video tokens $\mathbf{h}_V$, and supposing that the textual embeddings of instruction are initiated to $\mathbf{h}_T$ via the frozen word embedding table of LLMs, we concatenate them into a single sequence and fed it into LLMs. Considering fine-grained visual details existing in long videos, we expect that LMMs are capable of capturing the specific fine-grained visual information based on the understanding of instructions and the whole video representations. Hence, we devise a lightweight interactive visual adapter (IVA) to enable LLMs to focus on instruction-relevant fine-grained visuals during content generation.

Concretely, as the bottom part shown in Figure 1, we first introduce learnable dynamic tokens $\mathbf{h}_D = (h_1^D, ..., h_M^D)$ as the query signals and integrate it at the end of the input token sequence. It aims to capture previous instruction and video information via the self-attention mechanism of LLMs, functioning as query tokens to engage with the fine-grained spatial features of videos. Suppose that the output of i th layer of LLMs is $\mathbf{h}^i$. The specific calculation process of IVA between the i and $i+1$ th layers of LLMs is shown in Eq. 3 and 4 in order. Each layer of IVA consists of a selector and an interactor, which are capable of selecting relevant frames and capturing valuable fine-grained visual information. The operational process of the selector is described as follows:

$$\begin{aligned} \mathbf{h}_q^S &= W^q \mathbf{h}_d^i + b^q, \\ \mathbf{h}_k^S &= W^k \mathbf{h}_g^V + b^k, \\ \mathbf{M}^S &= \mathbf{h}_q^S (\mathbf{h}_k^S)^T, \\ \mathbf{h}^S &= Softmax(\mathbf{M}^S/\tau) Trans([\mathbf{h}_{I_1}^f, ..., \mathbf{h}_{I_N}^f]), \end{aligned} \tag{3}$$

where $\mathbf{h}_d^i$ refers to the hidden states $\mathbf{h}^i$ associated with the indices of dynamic tokens. W^q, W^k, b^q, and b^k are learnable parameters. $\mathbf{M}^S$ signifies the distribution score on the frames, which represents the relevant attention distribution. τ is the hyperparameter, which is set to 0.5. "Trans" refers to the transportation of feature dimension. $[\mathbf{h}_{I_1}^f, ..., \mathbf{h}_{I_N}^f]$ represents the fine-grained features of the entire video. The output $\mathbf{h}^S \in \mathbf{R}^{b\times M\times 576\times d_S}$ will be fed into the following interactor as the key value, where d_S represents the dimension of the selector.

For the interactor, the specific calculation progress could be given as Eq. 4.

$$\begin{aligned}\mathbf{h}_q^I &= W^1\mathbf{h}_d^i + b^1,\\ \mathbf{h}_k^I &= W^2\mathbf{h}^S + b^2,\\ \mathbf{M}^I &= \mathbf{h}_q^I(\mathbf{h}_k^I)^T,\\ \mathbf{h}_c^S &= Softmax(\mathbf{M}^I)(W^3\mathbf{h}^S + b^3),\\ \mathbf{h}^S &= MLP(\mathbf{h}_c^S) + \mathbf{h}_c^S\end{aligned} \tag{4}$$

where W^1, W^2, W^3, b^1, b^2, and b^3 are learnable parameters. Overall, we use the same four-layer calculations of the above selector and interactor to facilitate that LLMs interact with fine-grained visual features.

3.4 Training

Stage 1: Pretraining. To endow video tokens with meaningful representation, we first train the causal transformer, linear layers, and other learnable parameters during video tokens production, on massive video-caption pairs from WebVid, a total of 703k video-caption pairs. We freeze the other parameters of the overall model during this process and do not introduce the IVA module.

Stage 2: Video Instruction Tuning. At this stage, the model is required to generate responses that align with various instructions. These instructions often involve complex visual comprehension and reasoning, rather than merely describing visual signals. Note that the conversation data $[Q_1, A_1, ..., Q_r, A_r]$ consists of multiple rounds.

$$X_T^r = \begin{cases} Q_1, & r = 1 \\ \text{Concat}(Q_1, A_1, ..., Q_r, A_r), & r > 1 \end{cases} \tag{5}$$

where r represents the round count. As shown in Eq. 5, when $r > 1$, we concatenate the conversations from all previous rounds with the current instruction as the input for this round. The training objective remains the same as the previous stage. After this stage, the model can generate corresponding responses for various instructions.

4 Experiments

4.1 Data sets

While training the causal transformer, we utilize 702 thousand video-text pairs derived from Valley (Luo et al., 2023), sourced from WebVid (Bain et al., 2021). During the instruction tuning stage, we collect instructional datasets from three sources: a 100K video-text instruction dataset from Video-ChatGPT (Muhammad Maaz and Khan, 2023), a 36K short video-text instruction dataset from Valley-Instruct-73k (Luo et al., 2023), and a 34K multiple-choice QA dataset from NExT-QA (Xiao et al., 2021). Additionally, we assess the generalization of IVA using long and short video benchmarks. Long video benchmarks typically are characterized by videos exceeding one minute in duration. We evaluated our model using four prominent long video evaluation benchmarks: ActivityNet-QA (Yu et al., 2019), Social-IQ 2.0 (Wilf et al., 2023), LifeQA (Castro et al., 2020), WildQA (Castro et al., 2022). For short video benchmarks, the duration of the videos is often measured in several seconds. We evaluate our model against three notable short video evaluation benchmarks: MSVD-QA (Xu et al., 2017), MSRVTT-QA (Xu et al., 2017), and SEED-Bench (Li et al., 2023c).

4.2 Baselines

We mainly compare our models with the following video LLMs that could be extended to handle long videos. **Video-ChatGPT** (Muhammad Maaz and Khan, 2023) encodes frames independently and generates frame-level embeddings. Subsequently, it employs average pooling to transform these embeddings into both temporal and spatial features. These temporal and spatial features are then concatenated to derive video-level features and are fed into the LLM. **Video-LLaMA** (Zhang et al., 2023c) utilizes Vision-Language and Audio-Language to process video frames and audio signals separately. After fine-tuning on image instruction dataset and video instruction dataset, Video-LLaMA exhibited remarkable abilities in comprehending images and videos. **Video-Chat** (Li et al., 2023f) leverages perception tools to convert videos into textual descriptions in real-time, and employs a foundation model named InternVideo to encode videos into embeddings. These textual descriptions and video embeddings are then processed by an LLM for multimodal understanding. **LLaMA-Adapter** (Zhang et al., 2023d) is a lightweight adapter injected into the attention calculation of LLM, which could be used to handle videos, text, and image tasks.

4.3 Evaluation Metrics

For open-ended video QA tasks, we employ ChatGPT-Assistant to evaluate the performance following Video-ChatGPT (Muhammad Maaz and Khan, 2023). First, we input the question, the pre-

Method	ActivityNet-QA		Social-IQ 2.0		LifeQA		WildQA	
	Accuracy	score	Accuracy	score	Accuracy	score	Accuracy	score
Video-LLaMA (Zhang et al., 2023b)	12.4	1.1	48.2	2.8	28.8	2.3	**57.5**	3.2
Video-Chat (Li et al., 2023f)	26.5	2.2	-	-	-	-	-	-
LLaMA-Adapter (Zhang et al., 2023d)	34.2	2.7	-	-	-	-	-	-
Video-ChatGPT (Maaz et al., 2023)	35.2	2.7	51.6	3.2	31.2	2.5	54.9	3.3
Baseline (w/o IVA) w/o Spatial Token	38.4	3.0	49.8	3.1	26.3	2.2	51.4	3.0
Baseline (w/o IVA) w/o Causal Token	38.2	2.9	49.9	3.1	30.6	2.4	51.7	3.1
Baseline (w/o IVA)	40.8	3.0	53.0	3.3	30.9	2.5	54.4	3.2
IVA (LQ=8, NI=8)	41.6	3.0	54.0	3.6	46.5	2.8	51.2	3.1
IVA (LQ=16, NI=8)	42.1	3.0	64.9	3.9	50.5	3.0	53.5	3.2
IVA (LQ=32, NI=8)	41.9	3.0	57.1	3.7	**51.9**	**3.1**	53.7	3.2
IVA (LQ=16, NI=4)	42.2	3.0	63.3	3.9	50.1	3.0	52.5	3.2
IVA (LQ=16, NI=16)	42.3	3.0	55.4	3.7	50.0	3.0	55.1	**3.3**
IVA (LQ=16, NI=8)-272K	**46.8**	**3.1**	**68.0**	**4.0**	48.1	2.9	50.9	3.1

Table 1: **Comparison between different methods on 4 long video QA datasets.** LLM with IVA achieves the best performance on long videos compared to baselines and strong video LLMs. "LQ" refers to the length of query tokens and "NI" represents the number of interactions between LLMs and IVA. "-272K" indicates that we introduce additional training data of long video datasets like LifeQA and Social-IQ based on the original short video data.

Method	MSVD-QA		MSRVTT-QA		SEED^{AR}	SEED^{AP}	SEED^{PU}
	Accuracy	score	Accuracy	score	Accuracy	Accuracy	Accuracy
Valley	-	-	-	-	31.3	23.2	20.7
Video-LLaMA	51.6	2.5	29.6	1.8	-	-	-
LLaMA-Adapter	54.9	3.1	43.8	2.7	-	-	-
Video-Chat	56.3	2.8	45.0	2.5	34.9	**36.4**	27.3
Video-ChatGPT	**64.9**	**3.3**	49.3	2.8	27.6	21.3	21.1
Baseline (w/o IVA)	54.5	3.2	49.6	2.9	22.5	23.5	24.8
IVA (LQ=8, NI=8)	53.2	3.2	47.6	2.9	32.0	31.8	27.5
IVA (LQ=16, NI=8)	55.7	3.2	49.1	2.9	**35.2**	32.0	**34.2**
IVA (LQ=32, NI=8)	53.0	3.2	47.2	2.9	32.2	32.1	28.8
IVA (LQ=16, NI=4)	55.0	3.2	47.8	2.9	32.5	31.7	26.0
IVA (LQ=16, NI=16)	53.3	3.1	47.1	2.8	31.8	29.4	31.0
IVA (LQ=16, NI-8)-272K	58.6	3.2	**50.2**	**2.9**	32.2	30.0	31.6

Table 2: **Comparison between different methods on 5 zero-shot short video QA benchmarks.** Benchmark names are abbreviated due to space limits. MSVD-QA(Xu et al., 2017); MSRVTT-QA(Xu et al., 2017); SEED^{AR}: SEED-Bench Action Recognition(Li et al., 2023c); SEED^{AP}: SEED-Bench Action Prediction(Li et al., 2023c); SEED^{PU}: SEED-Bench Procedure Understanding(Li et al., 2023c).

dicted answer, and the correct answer into ChatGPT. Second, we request ChatGPT to verify the accuracy of the predicted answer, expecting a binary response of 'yes' for correct predictions or 'no' for incorrect ones. Additionally, we require ChatGPT to rate the quality of the predicted answer on a scale from 0 to 5, where 5 indicates a perfect match. Finally, we determine the overall accuracy by counting the number of 'yes' responses and calculate the overall score by averaging all quality scores. This evaluation employs the "gpt-3.5-turbo" version of ChatGPT.

4.4 Implementation Details

We employ the AdamW optimizer (Kingma and Ba, 2014) in conjunction with a cosine learning rate scheduler to train our model. We first utilize 2 A100 GPUs to train visual-language MLP with 2 million image-text pairs with a global batch size of 256 and a base learning rate of 2e-4. Subsequently, we train the causal transformers using 703K video-text pairs data on the same two GPUs, employing a global batch size of 24 and a base learning rate of 3e-4. Transitioning to the video instruction tuning stage, we scale up to 8 A100 GPUs with a global batch size of 64. Here, we leverage LoRA to efficiently fine-tune the language model LLaMA. In our implementation, we set the rank to 128 and alpha to 256, maintaining a learning rate of 1e-4 for both LoRA and IVA parameters. Given the pretraining visual-language MLP and causal transformers, we adopt a smaller learning rate of 2e-5.

4.5 Main Results

We present the performance of the models on four long video QA benchmarks and five short video

Model	#DataSize	SIQ2	LifeQA	WildQA	Avg.
Video-ChatGPT	100k	48.2/2.8	28.8/2.3	57.5/3.2	44.8
Video-LLaMA	100k	51.6/3.2	31.2/2.5	54.9/3.3	45.9
Baseline(w/o IVA)	100k	53.0/3.3	30.9/2.5	54.4/3.2	46.1
IVA (LQ=16, NI=8)	100k	53.9/3.1	33.0/2.5	57.4/3.3	48.1 ↑ 2.0
Baseline(w/o IVA)	170k	56.7/3.2	31.7/2.3	52.2/3.1	46.9
IVA (LQ=16, NI=8)	170k	64.9/3.9	50.5/3.0	53.5/3.2	56.3 ↑ 6.4
Baseline(w/o IVA)	272k	59.3/3.3	34.5/2.4	50.5/3.1	48.1
IVA (LQ=16, NI=8)	272k	68.0/4.0	48.1/2.9	50.9/3.1	55.7 ↑ 7.6

Table 3: Performance comparison when increasing the size of instruction data.

Model	Total Second↓	Average Second↓
Video-ChatGPT	96.83	0.9683
Video-LLaMA	127.69	1.2769
Baseline(w/o IVA)	89.18	0.8918
Baseline(IVA, LQ=16, NI=8)	90.68	0.9068

Table 4: Comparison of model efficiency. All models are tested using the same 100 samples randomly selected from the evaluation set.

QA benchmarks. In long video QA benchmarks, our model achieved state-of-the-art (SOTA) results compared to the previous pure video LLMs, except WildQA. Especially on the LifeQA and Social-IQ 2.0 evaluation datasets, our model achieved significantly higher results, surpassing the previous SOTA accuracy by **18.0** and **7.4** percentage points, respectively. In short video QA benchmarks, our model also demonstrated strong capabilities across some evaluation datasets, especially in procedure understanding. We analyze the specific performance of WildQA in Appendix B. Overall, IVA significantly enhances the capability of LLMs to analyze and interpret long videos, maintaining high-performance levels without compromising the understanding and reasoning abilities of short videos.

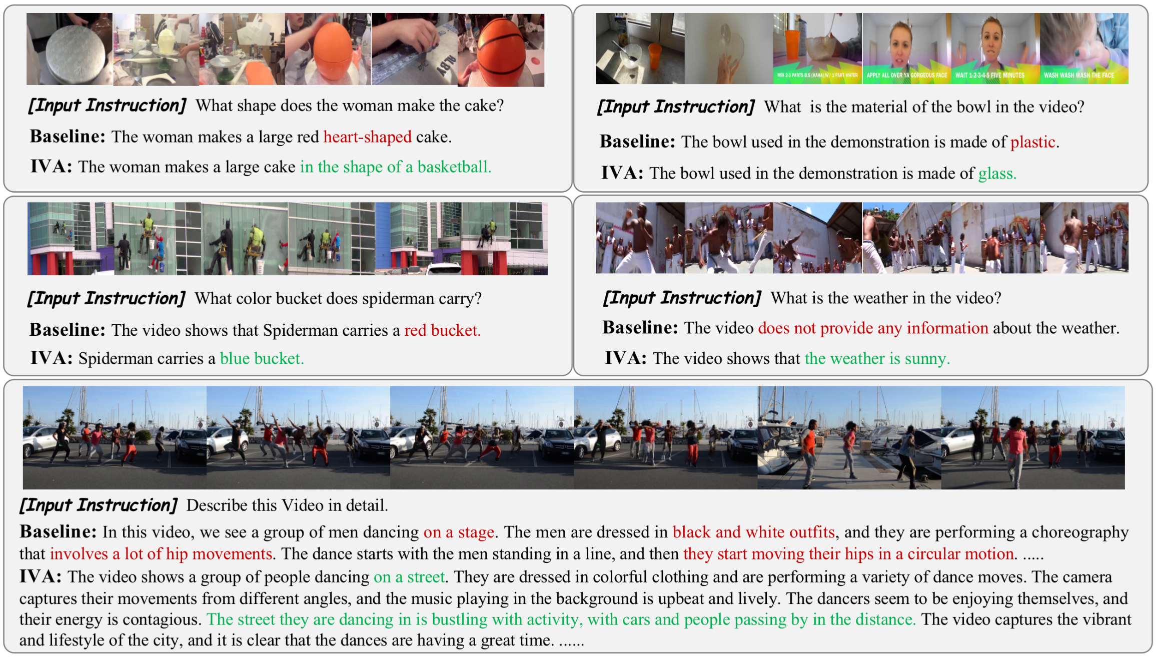
4.6 Ablation Study

Effect of IVA. From Tables 1 and 2, introducing the IVA module significantly improves visual understanding in long video datasets (Social IQ2, LifeQA, ActivityNet-QA) and short video datasets. Notably, our model achieved over a 20% improvement on LifeQA compared to the baseline, highlighting IVA's effectiveness. Comparing baselines without causal or spatial tokens confirms the efficiency of our video tokens production. The experimental results in Table 3 further suggest the beneficial impact of IVA when introducing more video instruction data.

Length of Query Tokens. Comparing the experimental results of IVA (LQ=8, NI=8) and IVA (LQ=16, NI=8) in Tables 1 and 2, we observed a significant decrease in evaluation results across various benchmarks when reducing the length of query tokens (16 → 8). Regarding the comparison between IVA (LQ=16, NI=8) and IVA (LQ=32, NI=8) in long video benchmarks, we noted a slight decrease in performance on the first two benchmarks when increasing the length. However, while there was a slight improvement in LifeQA, it did not conclude an overall performance enhancement. In contrast, in the short video benchmarks, there was a downward trend in results across all benchmarks. Overall, increasing the length of query tokens may not lead to performance improvement. Moreover, reducing the length of query tokens may result in the loss of crucial visual information, consequently leading to performance degradation.

Impact of increasing long video instruction data. We explore enhancing the model's performance by introducing more long video data in Table 3. We use the training sets of Social IQ2 and LifeQA, the video instruction part of the MIMIC-IT dataset(Li et al., 2023a), along with the open-ended QA training data from NExT-QA, to the existing 170K training data, forming a new 272K training dataset. From the results of IVA(lQ=16, NI=8)-272K in Tables 1 and 2, we observe a significant improvement in the Social IQ2 with the inclusion of more training data. However, there is little difference in the results on the remaining datasets, and in some cases, a certain degree of decline is observed. This may be attributed to the somewhat indiscriminate addition of new datasets, leading to a certain imbalance in the proportions of different data. The quality of the evaluation dataset will also affect the evaluation results, e.g., many questions in WildQA are typically wide-ranging and the criteria for answers can vary. We analyze this situation in Appendix B. Additionally, the training set of LifeQA only consists of 1,383 instances, which is relatively small in proportion to the total data, thus not providing sufficient improvement. While introducing more instruction data, IVA provides a bigger boost according to the results in Table 3.

Number of interaction between IVA and LLMs. We conduct experiments with both doubled and halved number of interaction layers. The detailed injection layers are shown in Appendix A. Upon analyzing the results of IVA(LQ=16, NI=8) and IVA(LQ=16, NI=4) in Tables 1 and 2, we observe that this reduction results in a significant decrease in its performance on most long video datasets, especially on the Action Prediction and Procedure



[Input Instruction] What shape does the woman make the cake?

Baseline: The woman makes a large red heart-shaped cake.

IVA: The woman makes a large cake in the shape of a basketball.

[Input Instruction] What is the material of the bowl in the video?

Baseline: The bowl used in the demonstration is made of plastic.

IVA: The bowl used in the demonstration is made of glass.

[Input Instruction] What color bucket does spiderman carry?

Baseline: The video shows that Spiderman carries a red bucket.

IVA: Spiderman carries a blue bucket.

[Input Instruction] What is the weather in the video?

Baseline: The video does not provide any information about the weather.

IVA: The video shows that the weather is sunny.

[Input Instruction] Describe this Video in detail.

Baseline: In this video, we see a group of men dancing on a stage. The men are dressed in black and white outfits, and they are performing a choreography that involves a lot of hip movements. The dance starts with the men standing in a line, and then they start moving their hips in a circular motion. .....

IVA: The video shows a group of people dancing on a street. They are dressed in colorful clothing and are performing a variety of dance moves. The camera captures their movements from different angles, and the music playing in the background is upbeat and lively. The dancers seem to be enjoying themselves, and their energy is contagious. The street they are dancing in is bustling with activity, with cars and people passing by in the distance. The video captures the vibrant and lifestyle of the city, and it is clear that the dances are having a great time. ......

Figure 2: Five cases illustrate the comparative performances of our IVA Model and Baseline. Red and green words represent the inaccurate and accurate statements, respectively.

Understanding of SEED-Bench. Moreover, the experimental results also indicate that increasing the number of layers (8 → 16) in the IVA interaction likewise causes a slight degradation in the model's performance. Given that there is no significant improvement observed when increasing the interaction times between IVA and LLMs, we set it to 8 as the standard for experimentation.

Comparison of Computational Efficiency. We evaluate our baseline (w/o IVA), IVA (LQ=16, NI=8), and comparative models Video-ChatGPT and Video-LLaMA on 100 samples from the WildQA dataset in Table 4. All models used llama-7b or its variants with identical generation parameters. Results show: 1) Our method has the fastest inference speed, significantly surpassing Video-LLaMA. 2) With shared IVA and 8 interactions using query tokens of length 16, the increase in inference speed is marginal. Our framework outperforms previous baselines in both inference efficiency and accuracy.

4.7 Case Study

We present four Video QA cases and one detailed description example in Figure 2. Upon examining the initial two examples, we observe that the model augmented with IVA exhibits enhanced proficiency in recognizing particular actions associated with specific frames. In response to specific queries, it could discern objects such as the 'basketball-shaped cake', which solely appears towards the video's conclusion, and the 'glass bowl,' present solely in the video's opening segment. Furthermore, the fourth question-answering example illustrates that IVA augments the model's reasoning ability, enabling it to deduce the prevailing weather conditions based on the lighting conditions within the video. These indicate the effectiveness of IVA in incorporating fine-grained visuals of long videos. Meanwhile, the bottom detailed description example reveals that when confronted with lengthy video descriptions, IVA could refine the perceptual acuity of LLMs, resulting in more precise recognition of elements such as the environment and colors.

5 Conclusion

In this study, we identify the principal obstacles in long video understanding and introduce an Interactive Visual Adapter (IVA) to facilitate dynamic interaction between LLMs and long videos. The IVA incorporates a selector module for identifying relevant temporal frames within long videos based on specific instructions and tokens, along with an interactor module that isolates detailed spatial visual features within long videos. The empirical results demonstrate that our IVA significantly improves LLMs' ability to comprehend and reason about long video content.

Limitations

Our work, while contributing insights into long video understanding and question-answering through employing LLMs, is subject to several limitations that warrant further investigation:

- **Optimization for Longer Videos**: Our current methodology demonstrates proficient performance in processing videos ranging from a few seconds to two minutes. However, the challenge of comprehensively understanding longer videos remains. Specifically, the optimization of video token length and the integration method of the Interactive Visual Adapter (IVA) within LLMs require further refinement to enhance their effectiveness and efficiency in handling extended content.

- **Impact of Interaction Frequency and Query Token Length**: The stability of the IVA can be influenced by the frequency of interactions and the length of query tokens. These factors often occur in the development of multimodal large models, where a delicate balance must be struck between achieving high performance and maintaining operational efficiency, particularly in the context of long video interaction and encoding.

- **Accuracy and Appropriateness of Generated Responses**: Another limitation is the potential for LLMs to generate responses that may be inaccurate, contain harmful content, or be factually incorrect. This issue stems from the inherent unpredictability in the response generation process of LLMs, underscoring the need for mechanisms that can ensure the reliability and appropriateness of the output.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. LifeQA: A real-life dataset for video question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France. European Language Resources Association.

Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai G. Burzo, and Rada Mihalcea. 2022. In-the-wild video question answering. In *COLING*, pages 5613–5635, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10406–10417.

Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. 2022. Maskvit: Masked visual pre-training for video prediction. In *The Eleventh International Conference on Learning Representations*.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. *arXiv preprint arXiv:2303.09713*.

Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. 2023. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023c. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *ICML*.

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

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

Xinyu Li, Yanyi Zhang, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. 2021. Vidtr: Video transformer without convolutions. *arXiv e-prints*, pages arXiv–2104.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. 2023g. Lmeye: An interactive perception network for large language models. *arXiv preprint arXiv:2305.03701*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

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

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*.

OpenAI. 2023. Chatgpt. *OpenAI Blog*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. 2022. Long-form video-language pre-training with multimodal temporal contrastive learning. *Advances in neural information processing systems*, 35:38032–38045.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. 2023. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. https://github.com/abwilf/Social-IQ-2.0-Challenge.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 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 (CVPR)*, pages 9777–9786.

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

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. 2023. Retrieval-based video language model for efficient long video question answering. *arXiv preprint arXiv:2312.04931*.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. Clipvip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416.

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

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702.

Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, and Jianhua Yao. 2023a. Dnagpt: A generalized pretrained tool for multiple dna sequence analysis tasks. *bioRxiv*, pages 2023–07.

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

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

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023d. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

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

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.

A Inserting IVA in Different Layers

NI	Corresponding Decoder Layers
4	0, 8, 16, 24
8	0, 4, 8, 12, 16, 20, 24, 28
16	0, 2, 4, 6, 8, 10, ..., 22, 24, 26, 28, 30

Table 5: **Ablation Study on Injection Layers for IVA. NI**: Number of Inserting Layers. The incorporated inserting layers were positioned before the respective decoder layers.

In this section, we detail the methodology behind our ablation studies focusing on the variation in the Number of Injection Layers. Our experiments were structured around three different setups, where the injection layers were configured to be 4, 8, and 16 in number. To ensure a uniform distribution, these

Injection Layers were interspersed throughout the decoder layers of the language model evenly. We utilized the Vicuna-7B model as our experimental framework, which is equipped with 32 decoder layers. The specific layers of the decoder that received the Injection Layers are outlined in Table 5, providing a clear reference to how the integration was achieved in each experimental setup

B Discussion of the lower performance of IVA on the WildQA dataset.

We meticulously reviewed 652 question-and-answering pairs within the WildQA dataset, benchmarking various models against these inquiries. Of these, 319 questions were of the "what" variety, for instance, "What is the man doing?" or "What clothes is the woman wearing?" These questions are typically wide-ranging and the criteria for answers can vary, leading to inconsistencies in evaluation outcomes.

Additionally, a comparative analysis of "What"-specific accuracy versus overall sample accuracy revealed a direct correlation between IVA's performance on "what" questions and its overall accuracy. However, IVA prioritizes visual information comprehension and its responses depend on visual elements. After introducing more data for training, as the experimental results are shown in Tables 1 and 3, improved performance across other datasets but not on WildQA further indicates the occasional disparity between its responses stemming mainly from video contents and the standard answers' format.

Example: **Question**: What types of vehicles are being affected? **Answer**: Cars, buses, motorcycles, mopeds. **Prediction**: The video shows that cars and boats are being swept away by the flood water.

Model	**What-Acc (319 samples)**	**All-Acc (652 samples)**
Video-Chatgpt	43.88 / 2.91	54.9 / 3.3
Video-LLaMA	43.26 / 2.79	57.5 / 3.3
Baseline (w/o IVA)	40.12 / 2.74	52.2 / 3.1
IVA (LQ=16, NI=4)	40.50 / 2.80	52.5 / 3.2
IVA (LQ=16, NI=8)	41.69 / 2.78	53.5 / 3.2
IVA (LQ=16, NI=16)	42.95 / 2.83	55.1 / 3.3
IVA (LQ=16, NI=8)-272K	39.23 / 2.73	50.9 / 3.1

Table 6: Comparison of models on What-Acc and All-Acc metrics.