TC-LLaVA: Rethinking the Transfer of LLava from Image to Video Understanding with Temporal Considerations

Mingze Gao[1, 2, 3], Jingyu Liu[2], Mingda Li[2], Jiangtao Xie[4], Qingbin Liu[2], Kevin Zhao[2], Xi Chen[2,*], Hui Xiong[1, 3, *]

[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]Tencent PCG
[3]The Hong Kong University of Science and Technology, China
[4]Dalian University of Technology, China

Abstract

Multimodal Large Language Models (MLLMs) have significantly improved performance across various image-language applications. Recently, there has been a growing interest in adapting image pre-trained MLLMs for video-related tasks. However, most efforts concentrate on enhancing the vision encoder and projector components, while the core part, Large Language Models (LLMs), remains comparatively under-explored. In this paper, we propose two strategies to enhance the model's capability in video understanding tasks by improving inter-layer attention computation in LLMs. Specifically, the first approach focuses on the enhancement of Rotary Position Embedding (RoPE) with Temporal-Aware Dual RoPE, which introduces temporal position information to strengthen the MLLM's temporal modeling capabilities while preserving the relative position relationships of both visual and text tokens. The second approach involves enhancing the Attention Mask with the Frame-wise Block Causal Attention Mask, a simple yet effective method that broadens visual token interactions within and across video frames while maintaining the causal inference mechanism. Based on these proposed methods, we adapt LLaVA for video understanding tasks, naming it Temporal-Considered LLaVA (TC-LLaVA). Our TC-LLaVA achieves new state-of-the-art performance across various video understanding benchmarks with only supervised fine-tuning (SFT) on video-related datasets.

Introduction

By leveraging vast open-source and AI-generated datasets (Lin et al. 2014; Schuhmann et al. 2022; Chen et al. 2023), along with the impressive development of large language models such as GPT (Achiam et al. 2023), LLaMA (Touvron et al. 2023), and GLM (Du et al. 2021), Multimodal Large Language Models (MLLMs) have demonstrated remarkable proficiency in image comprehension tasks (Liu et al. 2024b,a; Li et al. 2023a; Zhu et al. 2023). Given the powerful capabilities of image-pretrained MLLMs, a recently emerging research focus is on transferring these models from single-image tasks to video understanding.

Recently, various approaches (Maaz et al. 2023; Yang et al. 2022; Zhang et al. 2024) have tended to treat a video as a series of concatenated frames in the spatial dimension, thereby transferring video-related tasks back to image-related tasks. However, these methods face two issues as they treat text and visual tokens as the same modality and fed them into the LLMs as a unified input. Firstly, utilizing LLMs' vanilla attention mechanism to uniformly process all tokens overlooks the distinct interactions between visual tokens within individual video frames and those across different frames. Secondly, it neglects the temporal information inherent in the video input, which is crucial for video understanding tasks. Consequently, the constructed video MLLM fails to effectively summarize the dynamic events occurring within videos, reducing the analysis to single frames as if they were still images. For instance, it fails to adequately capture and detail the complex motion changes of the primary subject in the video, particularly in activities such as dancing or gymnastics. This deficiency ultimately results in inaccurate or 'hallucinatory' responses by the model, as depicted in Figure 1.

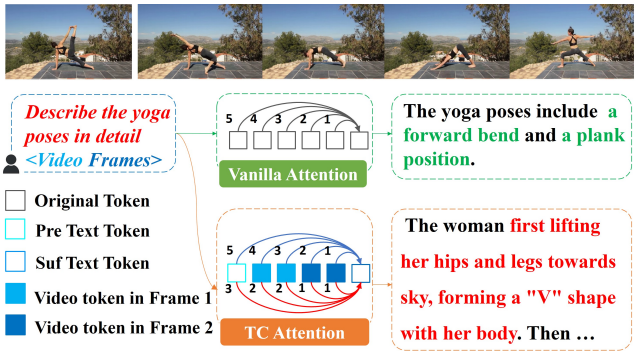


Figure 1: Video language processing with LLaMA (Touvron et al. 2023) and our TC-LLaVA, where arrows represent the attention interactions with this token, and numbers indicate the relative positional distance between tokens. Vanilla Attention uniformly encodes and applies attention interactions to both visual and text tokens. The proposed TC Attention incorporates temporal information encoding and differentiates interactions between visual tokens within and across frames, which are indicated by different colors.

In this paper, we propose Temporal-Considered (TC)

*Corresponding authors.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

LLaVA, a novel video-language framework designed to address the aforementioned issues. The primary innovation is to enhance the temporal awareness of MLLMs and distinguish the attention interactions between text and video modalities through two core strategies. First, we introduce Temporal-Aware Dual RoPE, which assigns each token an independent position id with the original RoPE to preserve global relative positional relationships, while incorporating temporal-aware RoPE to assign the same position id to visual tokens within the same frame and to encode inter-frame relationships to capture the temporal dynamics of videos, as shown in Figure 1. Additionally, we design three different attention masks to optimize token interaction strategies in attention computation, accounting for the distinct characteristics of visual and text tokens. Finally, we select the Frame-wise Block Causal Attention Mask to replace the original causal attention mask, enhancing interaction between visual tokens within and across frames while preserving the causal reasoning paradigm, making it more suitable for causal language model inference.

To verify the effectiveness of our TC-LLaVA, we evaluate the model on extensive video benchmarks, including MSVD (Xu et al. 2016), MSRVTT (Xu et al. 2016), ActivityNet (Caba Heilbron et al. 2015), TGIF (Li et al. 2016), VCGbench (Maaz et al. 2023) and MVbench (Li et al. 2023d). Comparing with the latest video MLLMs, TC-LLaVA achieves new state-of-the-art performance on these benchmarks at the same model scales, demonstrating the benefits of enhancing visual token interactions within and across frames, as well as the importance of incorporating temporal information in video analysis.

Related Work

Attention in Vision and Language Models

The introduction and evolution of the attention mechanism have significantly enhanced model performance in natural language processing (NLP) and computer vision (CV). The earliest attention mechanism by (Bahdanau, Cho, and Bengio 2014) allowed machine translation models to assign different weights to input sentence parts, improving translation accuracy. (Vaswani et al. 2017) introduced the Transformer model, which uses a self-attention mechanism to enable parallel processing and superior long-range dependency modeling, achieving significant results in multiple NLP tasks. To further optimize the attention computation, (Shaw, Uszkoreit, and Vaswani 2018) propose Relative Position Encoding (RPE) to improve the token interaction by introducing extra position information. Recently, Rotary Position Embedding (RoPE) (Su et al. 2024) is designed for the interaction limitation of RPE by leveraging complex number rotations. In CV, attention mechanisms have proven effective with models like Non-local Neural Networks by (Wang et al. 2018) and Vision Transformer (ViT) (Dosovitskiy et al. 2020), which first applies the Transformer architecture to image classification tasks. There have also been numerous advancements (Liu et al. 2021; Xie et al. 2021; Liu et al. 2022) in attention mechanisms that continually improve the performance of Transformer-based models, enhancing their ability to capture essential features and increasing computational efficiency. Our work continues to delve deeply into improving attention computation in the multimodal domain of video and text, and we propose the TC-Attention method to achieve this goal.

Video Multimodal Large Language Models

Video Multimodal Large Language Models (Video MLLMs) operate by aligning modalities and performing instruction fine-tuning on video data, enabling them to generate responses based on user instructions and input video streams. Recently, Video MLLMs have experienced rapid development. One significant milestone in this field is BLIP2 (Li et al. 2023a), which integrates a frozen vision encoder with a Q-Former to enhance video processing efficiency, demonstrating remarkable zero-shot capabilities in Video Question Answering (VQA) and outperforming existing techniques. Video-ChatGPT (Maaz et al. 2023) introduced video instruction tuning and created a high-quality instructional dataset, setting a new standard for video-based text generation benchmarks. VideoChat (Li et al. 2023b) employed cross-attention mechanisms to condense visual tokens and align user queries with the dialogue context, enhancing interpretative capabilities. Building on this, VideoChat2 (Li et al. 2023d) refined the approach with a multi-stage bootstrapping technique focused on modality alignment and instruction tuning, utilizing a robust collection of high-quality video data. Chat-UniVi (Jin et al. 2024) processes longer videos by introducing a method for compressing tokens in both the spatial and temporal dimensions. LLaMA-VID (Li, Wang, and Jia 2023) introduced an innovative dual-token approach that effectively condenses video representations by segregating context and content tokens, allowing for more efficient compression. VideoLLaMA and VideoLLaMA2 (Zhang, Li, and Bing 2023; Cheng et al. 2024) enhances video understanding by incorporating audio modality information and utilizing a Spatial-Temporal Convolution (STC) connector. ST-LLM (Liu et al. 2024c) intruduce a dynamic masking strategy into MLLM. PLLaVA (Xu et al. 2024) explore the Image-pretrained LLaVA into video tasks with simple spatial pooling. In this paper, we introduce TC-LLaVA, which considers the differences in visual token interactions within and across frames, and directly incorporates temporal position into the causal attention computation to enhance the understanding of model.

Method

Preliminary: Introducing Position Embeddings

While Relative Position Encoding (RPE) (Shaw, Uszkoreit, and Vaswani 2018) incorporates relative positional information into the attention mechanism through a position bias element-addition computation with inter-layer attention map, this approach may limit interaction with attention weights and, consequently, hinder the effective utilization of relative positions. To address this limitation, RoFormer (Su et al. 2024) introduces RoPE, a novel method that more

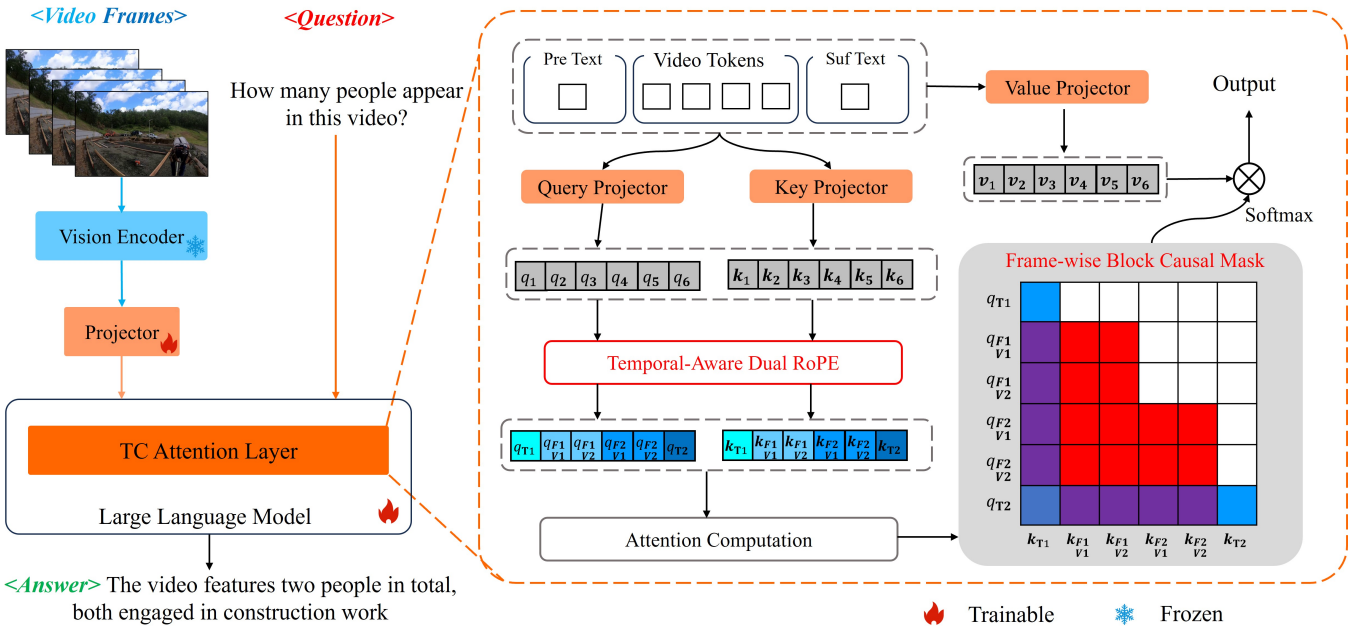


Figure 2: **The framework of TC-LLaVA**. During SFT stage, the projector and the language model (LLM) are unfrozen, while the visual encoder remains frozen. The right part illustrates our TC-attention mechanism in each Transformer layer. After applying Temporal-Aware Dual RoPE, both visual and text tokens acquire additional temporal positional information while preserving the global relative positional relationships. Frame-wise Block Causal Mask aims to enhance the visual tokens interactions within and across frames.

effectively incorporates relative positional information by leveraging complex number rotations.

Specifically, when computing the attention map, the RoPE (Rotary Positional Encoding) technique introduces the multiplication of Euler's formula $e^{i\theta}$ to the query and key vectors as a relative position embedding. For instance, when considering the n-th and m-th query and key vectors q_n and k_m in $\mathbb{R}^{1\times d_{head}}$, RoPE is applied as follows:

$$\mathbf{q}'_n = \mathbf{q}_n e^{in\theta}, \quad \mathbf{k}'_m = \mathbf{k}_m e^{im\theta}. \tag{1}$$

Then, the (n, m)-th component of the attention matrix is calculated as:

$$A_{(n,m)} = \text{Re}[\mathbf{q}'_n \mathbf{k}'^{*}_m] = \text{Re}[\mathbf{q}_n \mathbf{k}^{*}_m e^{i(n-m)\theta}], \tag{2}$$

where $\text{Re}[\cdot]$ denotes the real part of a complex number and * denotes the complex conjugate. By multiplying complex rotations $e^{i\theta n}, e^{i\theta m}$ depending on token position (n, m), RoPE injects relative positions $(n - m)$ into the attention matrix in a rotational form. In practical implementation, RoPE (Su et al. 2024) converts the vectors q_n and k_m from $\mathbb{R}^{1\times d_{\text{head}}}$ to complex vectors $\bar{q}_n$ and $\bar{k}_m$ in $\mathbb{C}^{1\times(d_{\text{head}}/2)}$. This is achieved by treating the $(2t)$-th dimension as the real part and the $(2t + 1)$-th dimension as the imaginary part, where $t \in 0, 1, \ldots, d_{\text{head}}/2$. This method results in the same attention values as $\mathbf{q}_n \mathbf{k}^T_m = \text{Re}[\bar{\mathbf{q}}_n \bar{\mathbf{k}}^*_m]$ while reducing computational overhead. Additionally, RoPE employs multiple frequencies θ_t through the channel dimensions of the query and key vectors as follows:

$$\theta_t = 10000^{-t/(d_{\text{head}}/2)}, \tag{3}$$

This approach allows for more effective integration of relative positional information within the attention mechanism, enhancing the model's capability to process and understand sequential data.

Temporal-Aware Dual RoPE

In the RoPE used by most current video-language large language models, the relative distance between the m-th text token T_m and the z-th visual token in the n-th frame $F_n V_z$ is defined as Eqn 4. Each text and visual token is treated as an independent position and assigned a unique position id for embedding. However, this position embedding method fails to distinguish visual tokens within and across different video frames, thereby neglecting the crucial temporal information necessary for effective video understanding tasks. Furthermore, as visual tokens constitute a significant proportion of the total tokens in video understanding tasks, the relative distance $P(T_m) - P(F_n V_z)$ between the generated text tokens and the visual tokens may become substantial. This increased distance can impair the model's ability to fully comprehend the visual information, leading to "hallucinated" responses (Ma et al. 2023).

$$A_{(q_{T_m}, k_{F_n V_z})} = \text{Re}[\mathbf{q}_{T_m} \mathbf{k}_{F_n V_z} e^{i(P(T_m)-P(F_n V_z))\theta}], \tag{4}$$

To address this limitation, we propose a Temporal-Aware Dual Rotary Positional Embedding (TAD-PoPE). It includes one RoPE that retains the global relative position relationships of the visual and textual tokens, and an additional time-aware RoPE to incorporate temporal information pertinent

to the video frames. Specifically, in contrast to the original position ids, the additional RoPE ensures that visual tokens within the same video frame share the same position id. Meanwhile, the temporal order is maintained across different frames, with the position ids incrementing accordingly. The proposed temporal position id is defined as follows:

$$\mathbf{I}_{\mathrm{t}}(n) = \begin{cases} n, & \text{if } n < v_{\mathrm{s}}, \\ v_{\mathrm{s}} + \left\lfloor \frac{n - v_{\mathrm{s}}}{m} \right\rfloor, & \text{if } v_{\mathrm{s}} \leq n \leq v_{\mathrm{e}}, \\ n - \left(v_{\mathrm{e}} - v_{\mathrm{s}} + 1 - \left\lfloor \frac{v_{\mathrm{e}} - v_{\mathrm{s}}}{m} \right\rfloor\right), & \text{if } n > v_{\mathrm{e}}. \end{cases} \tag{5}$$

where v_{s} and v_{e} are the starting and ending position ids of the visual tokens within the global RoPE position id n. m is the number of visual tokens per frame, and $\lfloor . \rfloor$ denotes the floor function, which rounds down to the nearest integer. By scaling the position ids, temporal information is introduced through the adjusted position $\hat{n}$, defined as:

$$\hat{n} = n + \gamma \cdot \mathbf{I}_t(n), \tag{6}$$

where γ is a scaling factor of constant magnitude. This adjustment ensures that temporal information is effectively incorporated into the original position embedding. For both text and visual tokens, the query and key vectors are updated using the adjusted positions $\hat{n}$ and $\hat{m}$:

$$\begin{aligned} \mathbf{q}'_n &= \mathbf{q}_n e^{i\hat{n}\theta} = \mathbf{q}_n e^{i(n+\gamma\cdot\mathbf{I}_t(n))\theta}, \\ \mathbf{k}'_m &= \mathbf{k}_m e^{i\hat{m}\theta} = \mathbf{k}_m e^{i(m+\gamma\cdot\mathbf{I}_t(m))\theta}, \end{aligned} \tag{7}$$

Finally, the attention matrix is calculated as follows:

$$\begin{aligned} A_{(\hat{n},\hat{m})} &= \mathrm{Re}[\mathbf{q}'_n {\mathbf{k}'_m}^*] \\ &= \mathrm{Re}[\mathbf{q}_n e^{i(n+\gamma\cdot\mathbf{I}_t(n))\theta} \mathbf{k}^*_m e^{i(m+\gamma\cdot\mathbf{I}_t(m))\theta}] \\ &= \mathrm{Re}[\mathbf{q}_n \mathbf{k}^*_m e^{i[(n-m)+\gamma(\mathbf{I}_t(n)-\mathbf{I}_t(m))]\theta}] \end{aligned} \tag{8}$$

This formula combines the updated query and key vectors to compute the attention map, incorporating both global positional and temporal information from video frames. By leveraging these aspects, we enhances the MLLM's ability to process and understand the input video comprehensively, resulting in more accurate and contextually appropriate responses.

Frame-wise Block Causal Attention Mask

Another often overlooked key point is the design of attention masks within the transformer layers in large language models. In causal language models like the GPT (Achiam et al. 2023) and LLama (Touvron et al. 2023) series, causal attention masks are employed to ensure that during text aggressive generation, historical token information is not leaked; that is, subsequent tokens can "see" preceding tokens, but preceding tokens cannot "see" subsequent tokens. This design is uniformly applied in such generative models to maintain the unidirectional flow of information, which is crucial for generating coherent and contextually appropriate text.

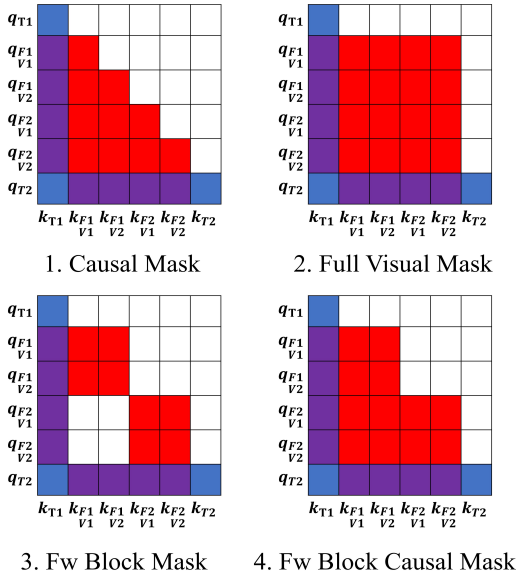


Figure 3: **Variations of Attention Masks**. To explore attention mechanisms for better interactions, we compare **Causal Mask** (1.) with three variants: **Full Visual Mask** (2.), **Frame-wise (Fw) Block Mask** (3.), and **Frame-wise (Fw) Block Causal Mask** (4.). Red indicates pure visual token interactions, blue represents pure text token interactions, and purple denotes interactions between visual and text tokens.

Mathematically, the causal attention mask $M \in \mathbb{R}^{T\times T}$ for a sequence of length T is defined as:

$$M_{ij} = \begin{cases} 0 & \text{if } i \geq j, \\ -\infty & \text{if } i < j. \end{cases} \tag{9}$$

This ensures that each position i only attends to previous positions (including itself), thus implementing the causal attention mechanism. The final attention weights are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V, \tag{10}$$

where Q is the query vectors, K is the key vectors, V is the value vectors, d_k is the dimension of the key vectors, and M is the causal attention mask.

However, for multimodal information involving both visual and textual inputs, the visual modality is only used as a conditional input to the language model. During the unidirectional decoding process of the language model, this design weakens the bidirectional attention interactions obtained from the visual encoder, reducing them to unidirectional attention interactions. To explore the impact of different attention masks, we design three distinct attention masks to enhance and investigate better interactions within visual tokens and between visual and text tokens, as illustrated in Figure 3.

Method	Vision Encoder	LLM Size	MSVD-QA		MSRVTT-QA		ActivityNet-QA		TGIF-QA		Video-ChatGPT					
			Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	CI	DO	CU	TU	CO	Avg.
Video-LLaMA	CLIP-G	7B	51.6	2.5	29.6	1.8	12.4	1.1	-	-	1.96	2.18	2.16	1.82	1.79	1.98
Video-LLaMA2	CLIP-L	7B	70.9	3.8	-	-	49.9	3.3	-	-	3.14	**3.08**	3.69	2.56	**3.16**	3.13
LLaMA-Adapter	ViT-B	7B	54.9	3.1	43.8	2.7	34.2	2.7	-	-	2.03	2.32	2.30	1.98	2.15	2.16
Video-ChatGPT	ViT-L	7B	64.9	3.3	49.3	2.8	35.2	2.7	51.4	3.0	2.50	2.57	2.69	2.16	2.20	2.42
Chat-UniVi	ViT-L	7B	65.0	3.6	54.6	3.1	45.8	3.2	60.3	3.4	2.89	2.91	3.46	2.89	2.81	2.99
MovieChat	CLIP-G	7B	75.2	3.8	52.7	2.6	45.7	3.4	-	-	2.76	2.93	3.01	2.24	2.42	2.67
VideoChat	CLIP-G	7B	56.3	2.8	45.0	2.5	26.5	2.2	34.4	2.3	2.23	2.50	2.53	1.94	2.24	2.29
VideoChat2	UMT-L	7B	70.0	3.9	54.1	3.3	49.1	3.3	-	-	3.02	2.88	3.51	2.66	2.81	2.98
Vista-LLaMA	CLIP-G	7B	65.3	3.6	60.5	3.3	48.3	3.3	-	-	2.44	2.64	3.18	2.26	2.31	2.57
LLaMA-VID	CLIP-G	13B	70.0	3.7	58.9	3.3	47.5	3.3	-	-	2.96	3.00	3.53	2.46	2.51	2.89
IG-VLM LLaVA	ViT-L	7B	**78.8**	**4.1**	63.7	3.5	54.3	3.4	73.0	4.0	3.11	2.78	3.51	2.44	3.29	3.03
ST-LLM	BLIP2	7B	74.6	3.9	63.2	3.4	50.9	3.3	-	-	3.23	3.05	3.74	**2.93**	2.81	3.15
PLLaVA	ViT-L	7B	76.6	**4.1**	62.0	3.5	56.3	**3.5**	77.5	4.1	3.21	2.86	3.62	2.33	2.93	3.12
GPT-4V	Unk	Unk	76.3	4.0	**63.8**	3.5	**57.0**	**3.5**	65.3	3.7	**3.40**	2.80	3.61	2.89	3.13	3.17
TC-LLaVA	ViT-L	7B	**78.8**	**4.1**	63.2	**3.6**	56.8	**3.5**	**78.2**	**4.2**	3.25	2.96	**3.75**	2.91	3.09	**3.19**

Table 1: Results of video question answering on MSVD-QA, MSRVTT-QA, ActivityNet-QA, TGIF-QA, Video-ChatGPT.

Firstly, the **Full Visual Mask** modifies the causal attention mask to enable more extensive interactions among visual tokens across different frames. This mask can be represented as follows:

$$M_{ij}^{\text{Full Visual}} = \begin{cases} 0 & \text{if } i \geq j \text{ or } i, j \text{ are visual tokens,} \\ -\infty & \text{otherwise.} \end{cases}$$

Secondly is **Frame-wise Block Mask**, which limits the attention to adjacent visual tokens within the same frame. This is defined as follows:

$$M_{ij}^{\text{Fw Block}} = \begin{cases} 0 & \text{if } i \geq j \text{ and } i, j \text{ within the same frame,} \\ -\infty & \text{otherwise.} \end{cases}$$

Finally, we proposed **Frame-wise Block Causal Attention Mask** (FwBC), which combines the characteristics of the previous causal and block visual attention masks by incorporating broader visual token interactions within the frame while maintaining causal inference mode across video frames. This can be presented as:

$$M_{ij}^{\text{FwBC}} = \begin{cases} 0 & \text{if } i \geq j \text{ or } i, j \text{ within the same frame,} \\ -\infty & \text{otherwise.} \end{cases}$$

By adjusting these masks, we aim to achieve a better balance between visual and textual information integration, enabling MLLMs to distinguish and process both video and text modalities more effectively while enhancing the spatiotemporal global attention to the most critical visual modality information for video understanding tasks. Finally, we utilized ablation experiments to select the Frame-wise Block causal Attention Mask for constructing TC-LLaVA.

Experiments

Experimental Setup

Instruction Tuning Datasets. In alignment with the instruction tuning setting outlined in VideoChat2 (Li et al. 2023b), which integrates data for a variety of video understanding tasks, we utilized an extensive and diverse collection of datasets. Specifically, these include 27k conversation videos from VideoChat (Li et al. 2023c) and Video-ChatGPT (Maaz et al. 2023), 80k classification task samples from Kinetics (Kay et al. 2017) and SthSthV2 (Goyal et al. 2017), 450k captioned data from Webvid (Bain et al. 2021), YouCook2 (Zhou, Xu, and Corso 2018), TextVR (Wu et al. 2023), and VideoChat, 117 reasoning data samples from NextQA (Xiao et al. 2021), CLEVRER (Yi et al. 2019), and 109,000 annotated question answering samples from Webvid, TGIF (Li et al. 2016), and Ego4D (Grauman et al. 2022). In total, we employed 783k video instruction data samples for conducting supervised finetuning (SFT) our TC-LLaVA.

Evaluation Benchmarks. The performance of our trained TC-LLaVA model is assessed using a series of video understanding benchmarks, specifically targeting open-ended Video Question Answer (VideoQA) tasks. These benchmarks include MSVD-QA (Xu et al. 2017), MSRVTT-QA (Xu et al. 2016), Activity-QA (Caba Heilbron et al. 2015), and TGIF-QA (Li et al. 2016), where responses generally consist of single-word answers. The accuracy (with true/false answers) and quality (scored from 0 to 5) of the models' responses are evaluated using GPT-3.5 (OpenAI. 2023). Moreover, we employ the Video-based Generative Performance benchmark (VCG Score), as introduced by VideoChatGPT (Maaz et al. 2023). This benchmark requires longer answers and evaluates five key aspects of video understanding: Correctness of Information (CI), Detail Orientation (DO), Context Understanding (CU), Temporal Understanding (TU), and Consistency (CO). The generative performance is also assessed using the GPT-3.5 model. In addition, we evaluate TC-LLaVA on the multi-choice Question Answering benchmark, MVBench (Li et al.

Method	Vision Encoder	LLM Size	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	Avg.
Video-LLaMA	CLIP-G	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
LLaMA-Adapter	ViT-B	7B	23.0	28.0	51.0	30.0	35.0	35.0	33.5	33.5	21.5	21.5	36.0	29.0	31.5	32.5	44.5	31.5	31.5	22.5	28.0	32.0	31.7
Video-ChatGPT	ViT-L	7B	23.5	26.0	62.0	22.5	26.5	54.0	28.0	30.0	23.0	20.0	31.0	30.0	25.5	39.5	48.5	29.0	40.0	25.0	26.0	35.0	32.7
VideoChat	CLIP-G	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
VideoChat2	UMT-L	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
ST-LLM	BLIP2	7B	66.0	53.5	**84.0**	44.0	58.5	80.5	**73.5**	38.5	**42.5**	31.0	86.5	36.5	56.5	78.5	43.0	44.5	46.5	34.5	41.5	58.5	54.9
PLLaVA	ViT-L	7B	58.0	49.0	55.5	41.0	61.0	56.0	61.0	36.0	23.5	26.0	82.0	**39.5**	42.0	52.0	45.0	42.0	53.5	30.5	48.0	31.0	46.6
GPT-4V	Unk	Unk	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
TC-LLaVA	ViT-L	7B	**71.5**	56.5	67.5	44.5	59.5	**84.0**	70.0	37.0	39.5	39.5	85.5	35.5	**59.5**	**83.5**	**53.5**	42.0	**54.0**	32.0	47.0	**70.0**	**56.6**

Table 2: Results on MVBench multi-choice question answering.

2023d), which consists of 20 tasks that demand nuanced temporal comprehension of videos.

Implementation Details Initialized from the image-pretrained MLLM LLaVA-Next (Zhang et al. 2024), which is based on the Vicuna-7B-v1.5 (Zheng et al. 2024), our TC-LLaVA 7B conduct further video instruction supervised finetuning (SFT) and evaluation on the datasets mentioned above. Following the experimental settings in (Xu et al. 2024), we uniformly sample 16 frames from the raw video as input and use global average pooling to downsample the visual features from a shape of $24*24*d$ to $12*12*d$ where d represents the input feature dimension of the LLM part. During the SFT stage, we employ a batch size of 128 and a learning rate of 2e-5, utilizing a cosine scheduler and a warmup ratio of 0.03. All reported results are evaluated on models trained for 7k steps on 8 NVIDIA A100 GPU. For evaluation, we use the GPT-3.5-turbo-0125 model across benchmarks that require additional scoring or assessment.

Comparison with SOTA

In this section, we compare our TC-LLaVA with recent advanced works, including Video-LLaMA (Zhang, Li, and Bing 2023), LLaMA-Adapter (Zhang et al. 2023), Video-ChatGPT (Maaz et al. 2023), Chat-UniVi (Jin et al. 2024), MovieChat (Su et al. 2020), VideoChat (Li et al. 2023b), VideoChat2 (Li et al. 2023d), Vista-LLaMA (Ma et al. 2023), LLaMA-VID (Li, Wang, and Jia 2023), IG-VLM LLaVA (Kim et al. 2024), ST-LLM (Liu et al. 2024c), PLLaVA (Xu et al. 2024), and GPT-4V (Achiam et al. 2023), across various video understanding benchmarks. The best performance is indicated in **bold**, and the second-best results are indicated with underlining. As shown in Table 1, our TC-LLaVA achieves a new state-of-the-art performance across MSVD-QA, TGIF-QA, and Video-ChatGPT, surpassing GPT-4V by 2.5%, 7.9%, and 0.02%, respectively. Additionally, our TC-LLaVA achieves the best performance across video question-answering benchmarks on the Score metric. Compared to the latest work PLLaVA, which is also initialized from LLaVA-Next and continues using original causal attention mask and RoPE, our TC-LLaVA outperforms it across all five evaluation benchmarks, demonstrating the effectiveness of our proposed methods.

Furthermore, we evaluate TC-LLaVA on MVbench, a multiple-choice video question answering benchmark, focusing on questions that require comprehensive understanding of the entire video. As shown in Table 2, TC-LLaVA achieves state-of-the-art performance in the average MVbench score. Specifically, for time-related tasks such as Action Sequence (AS), Object Existence (OE), Moving Count (MC), Moving Attribute (MA), State Change (SC), Character Order (CO), and Counterfactual Inference (CI), TC-LLaVA demonstrates a significant performance margin of at least 0.5% over other open-source models. Even when compared to GPT-4V, we maintain an edge in average performance across all 20 tasks by 13.1%.

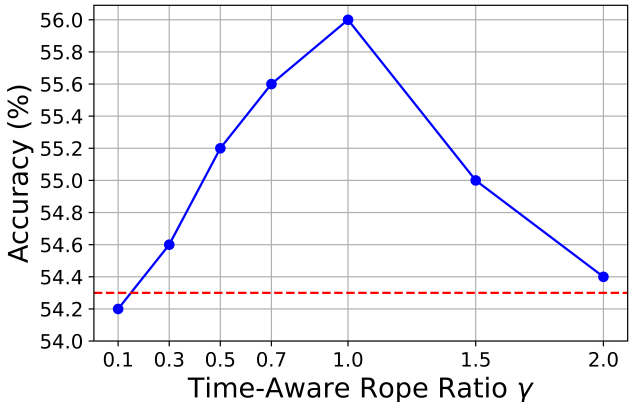


Figure 4: Different ratio γ settings of Time-Aware RoPE on MVbench. The red dashed line in the figure represents the baseline performance, which is the performance without adding the time-aware rope. The blue line shows the performance variations of the model under different ratio settings.

Ablation Studies

In this subsection, we conduct ablation studies to assess the impact of key components. Specifically, we examine the manual ratio settings γ of Time-Aware Dual RoPE and other designs of the attention mask, beyond the proposed Frame-wise Block Causal Mask as shown in Figure 3. For these studies, we use the basic settings as a combination of both the original RoPE and Causal Attention Mask, while keeping the previously mentioned training settings. The evaluation is performed on MVbench. Finally, we present a visualized heatmap comparing the attention weights of our TC-Attention mechanism to the vanilla attention.

Methods	Mvbench	VCGbench
Baseline	54.3	3.09
Full Visual Mask	53.8	3.04
Fw. Block Mask	54.0	3.08
Fw. Block causal Mask	**55.9**	**3.13**
Just Time-Aware RoPE	54.2	3.11
Time-Aware Dual RoPE	**56.0**	**3.15**
TC-LLaVA	**56.6**	**3.19**

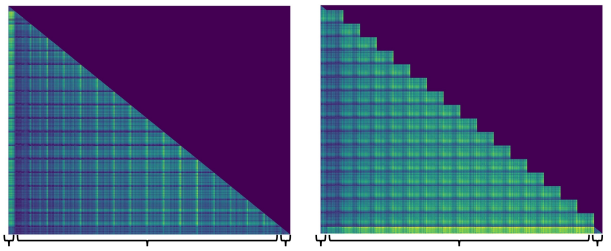
Table 3: Ablation Study of Different Method Settings. The baseline settings use the original RoPE and Causal Mask.

Time-Aware RoPE Ablation Firstly, Maintaining global Rotary Position Embedding (RoPE) is crucial for preserving the global positional relationships between tokens. LLaVA treated each token in an image as having an independent position. When transitioning from image to video understanding tasks, we aim to retain the characteristics of these pre-trained weights while introducing time-aware RoPE to incorporate temporal information. If we entirely abandon the use of RoPE, it could result in a partial loss of the capabilities encoded in the pre-trained LLMs, ultimately affecting the final performance.

Secondly, RoPE employs a rotational invariant mechanism, which contrasts with the linear and fixed positional embedding schemes of absolute and learnable embeddings. These inherent differences can hinder RoPE's effective scalability when integrating it with other positional encoding techniques, potentially resulting in suboptimal performance or conflicting representations.

Finally, we explore the impact of the hyperparameter γ in Time-Aware Dual RoPE. As shown in Figure 4, we evaluate TC-LLaVA on MVbench by setting the manual ratio γ across [0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0]. Compared to the baseline setting, which uses a single global RoPE (indicated by the red dashed line), introducing our Time-Aware RoPE increases performance, particularly when γ is close to 1.0, achieving the best performance at 56.0%. However, further increasing γ slightly reduces the final performance. We think this occurs because increasing γ too much might distort the original global position relationships encoded by the original RoPE, leading to suboptimal integration of spatial and temporal information. In the end, we choose γ as 1.0 for TC-LLaVA's experimental setting across the entire paper.

Attention Mask and Combination Ablation We further explore other attention mask variances mentioned above. As shown in Table 3, using Full Visual and Frame-wise (Fw.) Block Masks enhances visual token interactions within frames but weakens or sacrifices causal relationships. This is crucial for video understanding, as future frames should be able to reference previous frames, but previous frames should avoid seeing future frames, similar to the way text sequences are handled in autoregressive generation. Our Fw. Block Causal Mask achieves better performance by considering both enhancing visual interactions and preserving the causal relationships between tokens. When combined with Time-Aware Dual RoPE, our TC-LLaVA demonstrates superior performance, scoring 56.6% on MVbench and 3.19 on VCGbench. This combination improves the Attention Module, the core component of the LLM, resulting in a more comprehensive and effective video understanding model.



(a) Vanilla attention (b) TC attention (Ours)

Figure 5: Comparison of attention weights with corresponding attention mask for Vanilla attention (a) and TC attention (b). Lighter colors represent higher weights.

Attention Visualization Finally, we illustrate the attention weights of both our TC-Attention and Vanilla Attention. For this experiment, we compare the video-finetuned LLaVA and TC-LLaVA by inputting the same video test samples and visualizing the average attention weights of different heads in the final decoding layer of the LLM. In the visualization of Figure 5, brighter colors represent higher weights while the darker color represent lower weights. The attention weights assigned to visual tokens are markedly more comprehensive and greater in our TC-Attention. This indicates that, unlike Vanilla Attention, which only focuses on the last few visual tokens of each frame, our TC-Attention attends to every visual token within and across frames. Additionally, the proposed TC-Attention assigns greater attention weight to subsequent text (user input), resulting in a considerably more substantial impact of visual tokens on language tokens. This demonstrates the effectiveness of TC-Attention in integrating visual and textual information, enhancing the model's overall understanding and performance.

Conclusion

In this work, we present TC-LLaVA, rethinking the attention design in large language models (LLM) for video tasks. We introduce two core components to achieve this: Temporal-Aware Dual RoPE, incorporating temporal information into the attention module while maintaining the global position information between visual and text tokens, and Frame-wise Block Causal Attention Mask, enhancing the interaction of visual tokens within frames while preserving causal relationships across video frames. By conducting simple supervised finetuning (SFT) on video-related instruction datasets, our TC-LLaVA achieves new state-of-the-art performance across various video understanding benchmarks, showcasing the effectiveness of these methods. As LLMs continue to scale up, their powerful performance has led to the negligence of some design details. We hope our work encourages researchers to rethink these design aspects.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, 5842–5850.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

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

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kim, W.; Choi, C.; Lee, W.; and Rhee, W. 2024. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

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.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023c. 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.; Wang, L.; and Qiao, Y. 2023d. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. arXiv:2311.17005.

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.; and Jia, J. 2023. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, R.; Li, C.; Tang, H.; Ge, Y.; Shan, Y.; and Li, G. 2024c. ST-LLM: Large Language Models Are Effective Temporal Learners. *arXiv preprint arXiv:2404.00308*.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, F.; Jin, X.; Wang, H.; Xian, Y.; Feng, J.; and Yang, Y. 2023. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

OpenAI. 2023. Chatgpt. In *https://openai.com/index/chatgpt*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Su, H.; Shen, X.; Xiao, Z.; Zhang, Z.; Chang, E.; Zhang, C.; Niu, C.; and Zhou, J. 2020. Moviechats: Chat like humans in a closed domain. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 6605–6619.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wu, W.; Zhao, Y.; Li, Z.; Li, J.; Zhou, H.; Shou, M. Z.; and Bai, X. 2023. A large cross-modal video retrieval dataset with reading comprehension. *arXiv preprint arXiv:2305.03347*.

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.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

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, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35: 124–141.

Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and Tenenbaum, J. B. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.