ViSpeak: Visual Instruction Feedback in Streaming Videos

Shenghao Fu[1,3,4,†], Qize Yang[3,†], Yuan-Ming Li[1,4], Yi-Xing Peng[1,3,4], Kun-Yu Lin[1,4], Xihan Wei[3], Jian-Fang Hu[1,4*], Xiaohua Xie[1,4,5,6*], Wei-Shi Zheng[1,2,4,6]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China;

[2]Peng Cheng Laboratory, China; [3]Tongyi Lab, Alibaba Group;

[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China;

[5]Guangdong Province Key Laboratory of Information Security Technology, China;

[6]Pazhou Laboratory (Huangpu), China

fushh7@mail2.sysu.edu.cn, qize.yqz@alibaba-inc.com, xiexiaoh6@mail.sysu.edu.cn, hujf5@mail.sysu.edu.cn

ViSpeak: https://github.com/HumanMLLM/ViSpeak

ViSpeak-Bench: https://github.com/HumanMLLM/ViSpeak-Bench

Abstract

Recent advances in Large Multi-modal Models (LMMs) are primarily focused on offline video understanding. Instead, streaming video understanding poses great challenges to recent models due to its time-sensitive, omni-modal and interactive characteristics. In this work, we aim to extend the streaming video understanding from a new perspective and propose a novel task named ***Visual Instruction Feedback*** *in which models should be aware of visual contents and learn to extract instructions from them. For example, when users wave their hands to agents, agents should recognize the gesture and start conversations with welcome information. Thus, following instructions in visual modality greatly enhances user-agent interactions. To facilitate research, we define seven key subtasks highly relevant to visual modality and collect the* ***ViSpeak-Instruct*** *dataset for training and the* ***ViSpeak-Bench*** *for evaluation. Further, we propose the* ***ViSpeak*** *model, which is a SOTA streaming video understanding LMM with GPT-4o-level performance on various streaming video understanding benchmarks. After finetuning on our ViSpeak-Instruct dataset, ViSpeak is equipped with basic visual instruction feedback ability, serving as a solid baseline for future research.*

1. Introduction

Recent Large Video Language Models [4, 17, 23, 35, 68, 69] excel at fine-grained spatial perception, long-term temporal reasoning, and comprehensive spatiotemporal understanding. In the offline setting, the entire video is provided and the complete context can be modeled. However, in streaming video understanding, models can not access the entire video. Video content continuously arrives, and the model must make decisions based on the information available so far while continuously processing incoming future data, which poses great challenges to recent LMMs.

Three key differences exist between streaming and offline video understanding: *First, the question answers in streaming video understanding are time-sensitive.* The answers for the same question "What is happening now?" vary at different timestamps and the model should output the answer at a proper time. *Second, streaming videos are always accompanied by streaming audios*, making problems as omni-modal ones. *Third but most importantly, streaming video understanding is distinguished by its* ***interaction characteristic***. The interaction characteristic encompasses three folds: **1) non-awakening interaction** where users can interact with agents at any time, **2) interruption** where users can stop the answer or change the topic at any time, and **3) proactive output** where agents can also express their mind at a proper time. Despite its significance, the interaction characteristic has *been largely overlooked* by the community. MMDuet [51] and Dispider [41] conducted preliminary explorations on proactive output to point out a specific event when it occurs based on user prompt. However, the prompts do not always exist, especially for an unintentional event or during communications. VITA [16] uses dual models to decide when to respond to instructions in audio but it can not respond to visual contents.

In this work, we dive deeper into the interaction characteristic of streaming video understanding and introduce a new task named **Visual Instruction Feedback** to explore

*: Corresponding authors are Xiaohua Xie and Jian-Fang Hu. †: Equal Contribution. Work was done when Shenghao Fu and Yi-Xing Peng were interns at Alibaba.

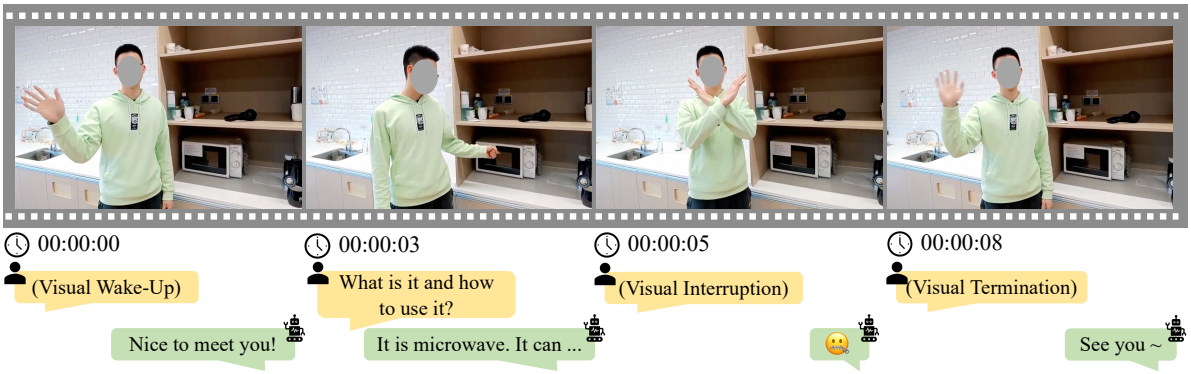

Figure 1. Examples of some actions in Visual Instruction Feedback task, which are Visual Wake-Up, Visual Reference, Visual Interruption, and Visual Termination in order. The content in parentheses is displayed by body language instead of text or speech.

the instructions in the visual modality. We restrict the feedback primarily in conversational scenarios and define it as *a kind of feedback towards visual contents to provide in-time interaction with users and necessary assistance effectively*. In this task, we select seven representative subtasks, including: 1) Visual Wake-Up: users use body language to start the conversation, 2) Anomaly Warning: agents provide in-time warnings or advice based on accidental events, 3) Gesture Understanding: agents respond to gestures from humans in conversations, 4) Visual Reference: users use body language to refer to a specific object, 5) Visual Interruption: users use body language to stop agents speaking, 6) Humor Reaction: agents share feedback to funny things with users, and 7) Visual Termination: users use body language to end the conversation. Examples are shown in Figure 1. To facilitate exploration, we collect the **ViSpeak-Bench** benchmark containing 1,000 videos and 1,000 QA pairs and the **ViSpeak-Instruct** training dataset containing 34k samples. As shown in Table 1, ViSpeak-Bench is the first comprehensive benchmark to evaluate the ability to respond to instructions in visual modality.

However, to the best of our knowledge, none of the open-sourced models can perform the Visual Instruction Feedback task even after finetuning on our dataset, especially for the visual interruption subtask, as they adopt a turn-taking chat template and the agent will fully express its mind without interruption before analyzing new user inputs. Thus, we propose the **ViSpeak** model which is finetuned from an existing omni-model using a novel three-stage finetuning procedure. In the first template alignment stage, we adapt the offline model to a streaming input template while preserving the original offline capacities. The template supports taking the user's input and the model's responses as inputs at the same time, making the two input streams fully time-aligned. This template also supports interruption when the model is speaking. In the second streaming finetuning stage, we enhance the model's streaming question-answering ability and proactive output ability. The resulting model achieves SOTA performance on the StreamingBench [29] and OVO-Bench [27], achieving 62.00 and 61.08 overall scores, separately, which are comparable with GPT-4o. Finally, we finetune the model on our collected ViSpeak-Instruct dataset which serves as a solid baseline for the Visual Instruction Feedback task.

In summary, our contributions are three folds:

1. We propose a novel streaming video understanding task named Visual Instruction Feedback, which requires the model to actively respond to visual contents. This task greatly enhances human-agent interactions.
2. To support exploration, we manually collect the ViSpeak-Bench benchmark and the ViSpeak-Instruct training dataset. We also provide some analysis based on the evaluation results of existing models.
3. We also propose a strong baseline ViSpeak for the new task, which is finetuned from an existing omni-modal model with three-stage finetuning. ViSpeak not only preserves offline understanding capacities but also achieves SOTA performance on streaming video understanding benchmarks.

2. Related Work

2.1. Large Multi-Modal Model

To better understand videos [30–32], recent Large Multi-modal Models (LMMs) have rapidly evolved from video understanding models [4, 23, 39, 68, 69] to omni-modal understanding models [16, 17, 26, 35, 59, 70]. With high-quality instruction turning data [3, 4, 48], improved training recipes [23, 48], and well-designed model architecture [28, 34, 46, 56], recent LMMs achieve fine-grained multi-modal alignment and extend their abilities of comprehensive image-level understanding [23, 49], fine-grained region perception [18, 21, 22, 50], long-term temporal reasoning [40, 66, 67], timestamp awareness [33, 47] and even human mind or emotional understanding [20, 53, 58, 70]. Although great progress has been made, recent video LMMs are primarily focused on offline videos where the entire video is provided for understanding.

Benchmark	#Videos	#QA Pairs	Time	Streaming	PO	Visual Instruct	Anno
ActivityNet-QA [62]	800	8,000	✗	✗	✗	✗	Manual
NExT-QA [52]	1,000	8,564	✗	✗	✗	✗	Auto
MVBench [25]	3,641	4,000	✗	✗	✗	✗	Auto
Video-MME [15]	900	2,700	✗	✗	✗	✗	Manual
ET-Bench [33]	7,002	7,289	✓	✗	✗	✗	Manual
StreamingBench [29]	900	4,500	✓	✓	✓	✗	Mixed
ViSpeak-Bench	1,000	1,000	✓	✓	✓	✓	Mixed

Table 1. Comparison between ViSpeak-Bench and other video benchmarks. 'Time' means the dataset is time-sensitive. 'PO' denotes the dataset to evaluate the proactive output ability.

Subtask	#Videos	#QA Pairs	QA Type
Visual Wake-Up	100	100	Open-Ended
Anomaly Warning	200	200	Open-Ended
Gesture Understanding	200	200	Open-Ended
Visual Reference	200	200	Multi-Choice
Visual Interruption	100	100	Open-Ended
Humor Reaction	100	100	Open-Ended
Visual Termination	100	100	Open-Ended
ViSpeak-Bench	1,000	1,000	

Table 2. ViSpeak-Bench benchmark statistics. ViSpeak-Bench contains 7 subtasks with 1,000 videos and 1,000 QA pairs.

2.2. Streaming Video Understanding

In practical human-agent interactions, LMMs should process streaming videos, which has drawn great attention in recent years. Many streaming video understanding benchmarks [27, 29, 55, 60] have been proposed, which have simultaneously spurred the development of many streaming video LMMs. VideoLLM-online [2], as a pioneer model, proposes a LIVE framework to process streaming videos, which uses Streaming Loss to learn when to speak. Subsequently, different models focus on different challenges in streaming video understanding. Flash-VStream [64] and IXC2.5-OL [66] propose some memory mechanisms to handle long context in streaming videos. Dispider [41] and MMDuet [51] focus on proactive output, the former disentangling perception, decision, and reaction while the latter introduces two additional heads. Mini-Omni2 [54] and VITA 1.5 [17] pay more attention to end-to-end real-time speech interaction. STREAMCHAT [55] and StreamingChat [60] aim to tackle the multi-round conversation problem. In this work, we extend the streaming video understanding problem from a new perspective and introduce a new Visual Instruction Feedback task.

3. Visual Instruction Feedback Task

3.1. Task Definition

In this work, we define a new task named **Visual Instruction Feedback** for streaming video understanding. Formally, we define the feedback as *a kind of feedback towards visual contents to provide in-time interaction with users and necessary assistance effectively*. We also restrict the feedback primarily to conversational scenarios. In this task, users may not provide explicit instructions in text or audio format. The agent should analyze visual inputs and express its mind accordingly. Assume a video stream $X_{[0,+\infty)}$ with infinite length. An action $A_{[t_1,t_2]}$ or an event $E_{[t_1,t_2]}$ may appear at any time $[t_1, t_2]$ and the model should recognize them and provide feedback within a limited time span $[t_1, t_2+T]$. As the task focuses on conversational scenarios, the agent should provide responses from a second-person perspective.

According to the definition, we summarize seven key subtasks. Before conversations:

1. **Visual Wake-Up (VW)**. Unlike keyword-based (like Siri) or VAD-based [16, 66] wake-up, visual wake-up needs the model response to salutations from users.
2. **Anomaly Warning (AW)**. In this task, models need to identify accidental events (*e.g.* fighting, explosion) or unintentional actions (*e.g.* falling down) and provide in-time warnings, advice, or help.

During conversations:

3. **Gesture Understanding (GU)**. Gestures play a vital role in conversations, even serving as a short response from users, like "OK", "GOOD", "ONE", "TWO". Models should understand human gestures and provide the corresponding feedback.
4. **Visual Reference (VR)**. In many cases, it is difficult to describe an object or where the object is precisely, but it can be done by pointing it out with the fingers, such as "What is this". Models should identify which object is referenced and answer questions from uses.
5. **Visual Interruption (VI)**. When users are not satisfied with the model's response or want to change the topic, they may interrupt the model with some body language, like the stop gesture. Models should stop generating the remaining responses when receiving these signals.
6. **Humor Reaction (HR)**. Humor understanding is one of the key abilities of humans. Reacting properly to funny things provides necessary emotional value to users.
7. **Visual Termination (VT)**. Visual termination is the action to end conversations. Although the actions in wake-up and termination may be the same (*i.e.* wave hands), they can be classified by the context where the action at the beginning of conversations is visual wake-up, otherwise visual termination. Models should be aware of contexts and start or end conversations properly.

Although there are many other scenarios where agents should talk to users actively, for example, sign language (we exclude it due to the technical complexity and its variability across the world), we believe the subtasks above cover common scenarios in daily life. Some examples are shown in Figure 1. More visualizations are shown in Supplement.

3.2. Dataset Construction

Video Collection and Annotation. We collect videos from both open-sourced datasets and our self-collected datasets.

For open-sourced datasets, we use anomaly videos in Holmes-VAU [65] and unintentional videos in OOPS [13] for Anomaly Warning and HumorQA in FunQA [53] for

Humor Reaction. All the datasets above are annotated with timestamps and event descriptions. We simply use GPT-4o to rewrite the annotations in a conversational tone to simulate conversations. For Gesture Understanding, we select 10 common gestures from Jester [36], each with 400 videos.

For other subtasks, we manually record the videos by ourselves. To ensure diversity, we recruit a team of 610 people (346 men and 264 women) with an age ranging from 10 to 70 years old from 5 provinces. For each kind of subtask, we carefully designed diverse conversation scripts and instructed participants to follow these scripts during filming to simulate human-computer interaction scenarios. Videos are recorded in various environments, including homes, offices, factories, warehouses, supermarkets, wild, and many others. In summary, we collect 1,185 videos for Visual Interruption, 4,689 videos for Visual Reference, 1,188 videos for Visual Wake-Up and Termination, and 1,507 videos for Gesture Understanding. For each video, we manually annotate the accurate timestamps for each body language, making the videos suitable for streaming video understanding. The corresponding scripts for each video are used as the annotations. For the Gesture Understanding subtask, in addition to the 10 gestures in Jester [36], we further add 10 gestures commonly used during conversations, with a total of 20 gestures. We also design 5 gestures for Visual Interruption. More details can be found in Supplement.

Dataset Enhancement. Although we have made great efforts to make the dataset large and diverse, the self-collected data still cannot cover infinite scenarios in the real world, especially for gesture understanding. To alleviate the problem, we augment the dataset with some offline video understanding datasets. The motivation is that the model can improve its social intelligence from the perspective of a bystander, simulating a child, who observes the conversations between adults. The offline data have no timestamp annotations and questions are appended at the end of the video.

Specifically, we select SMILE [20] for understanding funny things and IntentQA [24] and Social-IQ [63] for learning body languages. Further, we manually review the videos in Social-IQ [63] dataset and re-annotate some common body languages lasting longer than 1 second in conversations. For each action, we point out what action it is and why the speaker did the action in the context, which helps the model study the meaning of common body language in the wild. In summary, 678 videos with 1,861 annotations are collected. Examples can be found in Supplement.

Quality Verification. While recording videos, we will provide guidelines to participants to ensure the quality of the raw videos. The videos are then sent for human annotation. Low-quality data will be rejected and re-recorded. Some well-performed annotators will conduct spot checks on the annotation results. If significant quality issues are identified during annotation, the corresponding annotator will undergo retraining, and the data will be re-annotated.

Subtask	Data Source	Data Type	#Samples	Ratio
Visual Wake-Up	self-collected data	online	1k	0.03
Anomaly Warning	OOPS [13]	online	3k	0.09
	HIVAU [65]	online	3k	0.09
Gesture Understanding	Jester [36]	online	4k	0.12
	self-collected data	online	4k	0.12
	Social-IQ [63]	offline	2k	0.06
	IntentQA [24]	offline	5k	0.15
	SocialIQA [42]	offline	0.5k	0.02
	self-collected data	offline	1k	0.03
Visual Reference	self-collected data	online	5k	0.15
Visual Interruption	self-collected data	online	1k	0.03
Humor Reaction	FunQA [53]	online	2k	0.06
	SMILE [20]	offline	1k	0.03
Visual Termination	self-collected data	online	1k	0.03
ViSpeak-Instruct			34k	1

Table 3. Task and sample distribution in ViSpeak-Instruct.

Data Partition and Dataset Statistics. With the collected data above, we manually select some representative videos to construct the ViSpeak-Bench evaluation dataset. For Visual Wake-Up, Visual Termination and Visual Interruption, we select 100 videos for each subtask with actions lasting 2 seconds. For Visual Reference, we carefully select 200 videos with multiple objects. The referenced object may appear at any location within the frame and is not necessarily positioned at the center. We formulate this subtask as a multi-choice problem during evaluation and manually annotate each video with three other confusing options which are also displayed in the video, ensuring the answer is highly related to visual reference. For Humor Reaction, we select 100 humorous videos in FunQA that are only relevant to visual content. For Gesture Understanding and Anomaly Warning, we randomly select 200 videos for testing. The remaining data are contained in ViSpeak-Instruct and used for training. The statistics of ViSpeak-Bench and ViSpeak-Instruct are summarized in Table 2 and Table 3.

Evaluation Metrics. In our task, we evaluate both the timing accuracy $\mathcal{T}_{\text{acc}}$ of the model's feedback (*i.e.*, Time Accuracy) and the quality score $\mathcal{S}$ of its response text (*i.e.*, Text Score), and then derive an overall score $\mathcal{O}$. For Time Accuracy, the model should response within the ground-truth time span $[t_1, t_2 + T]$, where T is the time margin we set. For subtask s, the accuracy of T_{res} is measured based on whether it falls within this time window.

$$\mathcal{T}^s_{\text{acc}} = \frac{1}{N^s} \sum_{i=1}^{N^s} \mathbb{I}\big(T^{(i)}_{\text{res}} \in [t_1, t_2 + T]\big), \tag{1}$$

where N^s is the number of questions in each subtask.

For Text Score, the output of the model should accurately reflect the actions or events within the video and be consistent with historical context. The response must be positive and supportive, providing assistance to the user when necessary. Thus, we use the dialogue history and ground truth as references, designing different prompts for different subtasks. We use GPT-4o as the judge model, scoring the re-

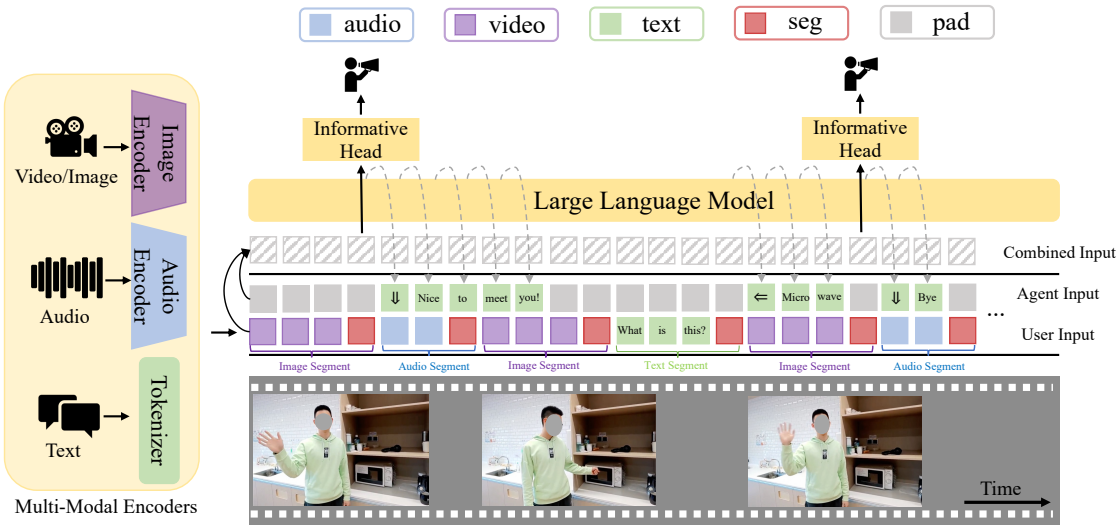

Figure 2. ViSpeak is an omni-modality LMM with multiple encoders and a LLM. To support streaming video analysis, ViSpeak takes two input streams as inputs, one for user inputs and one for self-generated outputs. Two streams will be combined into a single one before sending to LLM. An informative head is trained for visual proactive output.

sponses on a scale from 0 to 5 (see Supplement for more details). Note that, for the visual reference task, we use multi-choice questions for evaluation and rescale the score for this task to range from [0,5], while its response time accuracy is always set to 1.

Finally, the overall score is calculated as:

$$\mathcal{O} = \frac{1}{N}\sum_{s=1}^{N} \mathcal{T}_{\text{acc}}^{s} \times \mathcal{S}^{s}, \tag{2}$$

where N is the number of subtasks.

4. The ViSpeak Model

4.1. Model Architecture

In order to accomplish the Visual Instruction Feedback task, we design a model named **ViSpeak** as shown in Figure 2, which is an omni-model with an image encoder to extract image or video features, an audio encoder for encoding both audios and music, and a large language model to integrate multimodal features and conduct analyses to fulfill the relevant instructions. However, the turn-taking chat template with explicit role control is not suitable for interruption. Inspired by Moshi [10], we design a **two-stream chat template**, one for user inputs and one for agent historical outputs, so that the model can continuously process user inputs while outputting the next tokens. Thus, the model can adjust outputs based on upcoming inputs. Two input streams are combined into a single one before sending to the LLM. By default, we use a linear layer to predict the weights for a weighted sum of the two streams. In the ablation study part, we have tried many kinds of combination methods and found that they perform similarly.

Further, we segment the streaming inputs from users into multiple fragments, such as extracting one frame per second from the video and dividing the audio into 1-second snippets, subsequently organizing these segments in chronological order. Each segment is appended with a special <seg> token and the LLM can only start to speak from it. To differentiate the response towards different kinds of instructions, answers for text, audio, and visual instructions start with "⇐", "⇒", and "⇓" separately, following VITA [16]. For Visual Interruption, the model can simply output "⇓ Stop!" at a <seg> token to stop generation.

When to speak is a key problem in streaming processing. We find that using the original language model head (next token prediction) is sufficient to handle text-answering problems, *i.e.* the model can always output a "⇐" token at the end of a text segment. However, visual proactive output is a more challenging task and next token prediction can not manage the turn-taking problem well. Thus, we train another informative head to predict when to speak following MMDuet [51], which is a binary classification head to predict speaking or not. When the prediction score is above a predefined threshold, the model will respond to a visual instruction. In this work, we do not take the turn-taking problem for audio modality into account for simplicity.

4.2. A Three-Stage Finetuning Recipe

Directly training a strong streaming model from scratch is resource-demanding. Thus we begin with a well-pretrained omni-modal offline model [17] and adopt a three-stage finetuning recipe to train the model.

In the first template alignment stage, we adapt the offline model to our streaming input template with the goal

Method	Params	Frames	Omni	Real-Time Visual Understanding										Omni-Source Understanding					Contextual Understanding					Overall
				OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	All	ER	SCU	SD	MA	All	ACU	MCU	SQA	PO	All	
Proprietary MLLMs																								
Gemini 1.5 pro [44]	-	-	✓	83.43	77.94	89.24	81.65	79.17	83.92	83.93	60.32	74.87	49.22	77.39	52.40	50.80	80.40	87.60	67.80	52.80	42.40	59.20	45.10	51.06	70.26
GPT-4o [19]	-	-	✓	80.66	76.98	86.67	73.81	75.95	85.48	75.00	70.66	65.99	43.09	74.54	53.60	32.40	49.00	68.80	50.95	50.40	42.80	52.40	56.86	49.06	64.31
Claude-3.5-sonnet	-	-	✗	82.45	73.77	82.43	82.40	76.39	85.56	61.68	60.73	67.88	47.62	74.04	39.60	35.60	34.40	56.00	41.40	36.00	43.20	34.80	64.71	39.70	60.06
Open-Source Video MLLMs																								
LLaVA-OneVision [23]	7B	32	✗	82.83	77.34	83.23	83.33	72.05	74.77	73.15	68.29	71.10	41.97	74.27	41.20	26.10	43.20	52.80	40.83	35.08	30.40	30.00	29.55	31.68	58.56
MiniCPM-V [61]	8B	32	✗	78.20	71.88	84.18	83.99	75.16	75.39	72.22	56.50	67.14	47.15	72.43	42.00	27.71	40.40	50.80	40.23	37.50	27.20	40.00	22.22	34.09	57.80
InternVL-V2 [6]	8B	16	✗	73.84	65.63	78.80	82.03	71.43	72.90	73.15	63.01	65.44	42.49	70.11	44.80	28.11	47.20	50.80	42.73	35.08	27.20	42.80	40.91	35.40	57.28
Qwen2-VL [49]	7B	1 fps	✗	75.75	79.69	76.58	79.08	74.53	75.08	74.07	65.85	65.16	41.97	71.15	40.80	25.30	41.20	55.60	40.73	34.27	26.40	44.40	22.73	34.24	57.20
LLaVA-Next-Video [68]	32B	64	✗	80.11	71.09	80.70	80.72	71.43	73.21	62.96	59.35	63.17	36.79	69.83	41.60	24.50	44.40	56.40	41.73	34.27	28.80	44.00	18.18	34.58	56.73
Video-LLaMA2 [8]	7B	32	✓	59.95	60.16	62.97	60.46	54.66	46.11	41.67	46.75	48.16	34.72	52.58	43.60	23.29	35.20	41.60	35.92	28.23	26.00	21.20	0.00	23.54	43.30
Ola [35]	7B	64	✓	61.58	71.09	67.19	62.09	62.73	51.71	60.19	52.03	53.82	17.62	56.16	40.80	27.20	23.60	43.20	33.70	30.40	22.80	31.20	11.20	23.90	44.00
VITA 1.5 [17]	7B	16	✓	74.11	78.13	80.76	77.12	73.91	64.17	66.67	58.54	66.57	33.68	68.20	44.00	26.80	42.80	56.80	42.60	31.60	32.80	36.40	23.60	31.10	54.27
Open-Source Streaming MLLMs																								
Flash-VStream [64]	7B	-	✗	25.89	43.57	24.91	23.87	27.33	13.08	18.52	25.20	23.87	48.70	23.23	25.91	24.90	25.60	28.40	26.00	24.80	25.20	26.80	1.96	24.12	24.04
VideoLLM-online [2]	8B	2 fps	✗	39.07	40.06	34.49	31.05	45.96	32.40	31.48	34.16	42.49	27.89	35.99	31.20	26.51	24.10	32.00	28.45	24.19	29.20	30.80	3.92	26.55	32.48
IXC2.5-OL [66]	7B	64	✗	82.83	73.77	78.66	82.95	72.50	76.01	61.11	60.67	71.59	58.85	73.79	-	-	-	-	-	-	-	-	-	-	-
Dispider [41]	7B	1 fps	✗	74.92	75.53	74.10	73.08	74.44	59.92	76.14	62.91	62.16	45.80	67.63	35.46	25.26	38.57	43.34	35.66	39.62	27.65	34.80	25.34	33.61	53.12
ViSpeak (Ours, s2)	7B	1 fps	✓	79.84	88.28	83.28	81.05	76.40	75.08	70.37	65.85	77.34	34.20	74.36	42.80	35.20	61.20	74.80	53.50	38.80	36.80	44.00	38.80	39.60	62.00
ViSpeak (Ours, s3)	7B	1 fps	✓	79.84	71.09	81.39	78.76	74.53	70.09	63.89	64.23	71.39	27.98	70.44	47.20	56.40	61.60	81.20	61.60	49.20	36.40	39.20	50.80	43.90	62.58

Table 4. Performance on StreamingBench [29]. Results for ViSpeak trained after the second and third stage are reported.

Method	Params	Frames	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				Overall
			OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	REC	SSR	CRR	Avg.	
Proprietary MLLMs																		
Gemini 1.5 pro [44]	-	-	87.25	66.97	80.17	54.49	68.32	67.39	70.77	68.59	75.68	52.69	62.32	35.53	74.24	61.67	57.15	65.25
GPT-4o [19]	-	-	69.13	65.14	65.52	50.00	68.32	63.68	63.63	49.83	70.95	55.38	58.72	27.58	73.21	59.40	53.40	58.58
Open-Source Video MLLMs																		
Qwen2-VL [49]	72B	64	72.48	56.88	77.59	52.25	74.26	61.41	65.81	51.52	73.65	63.44	62.87	37.68	60.10	45.00	47.59	58.76
Qwen2-VL [49]	7B	64	69.13	53.21	63.79	50.56	66.34	60.87	60.65	44.44	66.89	34.41	48.58	30.09	65.66	50.83	48.86	52.70
LLaVA-Next-Video [68]	7B	64	69.80	59.63	66.38	50.56	72.28	61.41	63.34	51.18	64.19	9.68	41.68	34.10	67.57	60.83	54.17	53.06
LLaVA-OneVision [23]	7B	64	67.11	58.72	69.83	49.44	71.29	60.33	62.79	52.53	58.78	23.66	44.99	24.79	66.93	60.83	50.85	52.88
InternVL-V2 [6]	8B	16	68.46	58.72	68.97	44.94	67.33	55.98	60.73	43.10	61.49	27.41	44.00	25.79	57.55	52.92	45.42	50.05
LongVU [43]	7B	1 fps	55.70	49.54	59.48	48.31	68.32	63.04	57.40	43.10	66.22	9.14	39.49	16.62	69.00	60.00	48.54	48.48
VITA 1.5 [17]	7B	16	74.50	60.55	70.69	53.37	63.37	58.70	63.53	46.13	54.05	24.19	41.46	37.54	60.73	62.08	53.45	55.49
Open-Source Streaming MLLMs																		
Flash-VStream [64]	7B	1 fps	25.50	32.11	29.31	33.71	29.70	28.80	29.86	36.36	33.78	5.91	25.35	5.44	67.25	60.00	44.23	33.15
VideoLLM-online [2]	8B	2 fps	8.05	23.85	12.07	14.04	45.54	21.20	20.79	22.22	18.80	12.18	17.73	-	-	-	-	-
ViSpeak (Ours, s2)	7B	1 fps	75.17	58.72	71.55	51.12	74.26	66.85	66.28	59.93	48.65	63.98	57.52	33.81	68.52	60.42	54.25	61.08

Table 5. Performance of various MLLMs on OVO-Bench [27]. Results for ViSpeak trained after the second stage are reported.

of not compensating for its offline multi-modal understanding ability. In this stage, we select 300k text data from Magie [57], 665k image data from ShareGPT4V [3], 1,335k video data from LLaVA-Video [69], 410k audio data from LibriSpeech [38] and WavCaps [37], and 121k cross-modality data from Ola [35], with a total of 2.7M data for training. To save computation, we compress the data by concatenating short samples to a longer one, resulting in 2.0M data. To further enhance the cross-modality feature alignment, we use the audio in video data when available. We further use the CosyVoice2 [12] Text-to-Speech (TTS) method to change a small part of text questions in image and video data into speech following VITA [16]. To make sure the speech is rich in diversity, we select the voice of 5,962 speakers in VoxCeleb2 [9] as the condition for CosyVoice2 to synthesize speech. The training starts with tuning the projector with one quarter of the data and then training the projector and the LLM with LoRA and all data.

In the second streaming finetuning stage, we enhance the model's streaming question-answering ability and proactive output ability. Thus the data should be annotated with timestamps. In this stage, we use 81k data from MMDuet [51] with temporal video grounding task, dense captioning task, and multi-answer question answering task, 42k data from ET-Instruct [33] for temporal action localization task and referred video captioning task, and 42k data from EgoTimeQA [11] for general question answering task. We also sample 500k offline data in stage 1 to enrich the dataset. Finally, the training dataset comprises 657k samples. The informative head is trained at this stage.

Finally, we finetune the model on our collected ViSpeak-Instruct dataset, giving the model the ability to mine the instructions in the visual modality and respond to users actively. The resulting model ViSpeak serves as a solid baseline on ViSpeak-Bench.

5. Experiment

5.1. Implementation Detail

ViSpeak is finetuned from VITA 1.5 [17] due to its high performance on omni-modal data and early open-resourcing, which uses Qwen2 7B [45] as the LLM and InternViT-300M-448px [7] as the visual encoder. The audio encoder is designed by VITA itself and has 341M parameters. In the first stage, we first employ a learning rate of 5e-4 and batch size 256 for MLP adapter pre-training and a learning rate of 1e-4 and batch size 128 for LLM LoRA fintuning. The number of tokens for each image is 256 and the maximum number of images per video is 16. In the second and third stages, the training configurations are the same as those in stage 1 finetuning. However, as streaming video always lasts for a few minutes, we further downsample the image for each frame by a factor of 2, resulting in 64 tokens per image, and increase the maximum number of images per

Method	Params	Frames	Omni	Streaming	Time Accuracy (%)							Text Score								Overall
					AW	VI	HR	VW	VT	GU	**All**	VR	AW	VI	HR	VW	VT	GU	**All**	
Human (Avg)	-	-	-	-	70.00	100.00	90.00	92.00	96.00	98.80	91.13	4.80	2.45	4.58	3.06	5.00	5.00	2.85	3.96	3.69
Human (Max)	-	-	-	-	70.00	100.00	100.00	100.00	100.00	100.00	95.00	5.00	2.71	5.00	3.62	5.00	5.00	3.19	4.22	4.01
Proprietary MLLMs																				
Gemini 1.5 pro [44]	-	-	✓	✗	46.00	60.00	85.00	84.00	48.00	97.00	70.00	3.03	2.34	2.93	1.36	4.66	4.68	2.07	3.01	2.19
GPT-4o [19]	-	-	✓	✗	48.50	82.00	96.00	99.00	100.00	99.50	87.50	3.18	2.27	3.53	1.71	5.00	4.98	2.22	3.27	2.99
Open-Source Video MLLMs																				
InternVL-2.5 [5]	8B	16	✗	✗	41.50	55.50	46.00	96.00	72.00	99.50	68.42	2.93	2.16	3.67	0.74	3.05	4.81	1.26	2.66	1.98
Qwen2.5-VL [1]	7B	1 fps	✗	✗	42.50	78.00	31.00	95.00	85.00	98.50	71.67	2.34	2.31	2.31	1.32	5.00	3.91	1.02	2.60	2.25
Qwen2.5-VL [1]	72B	1 fps	✗	✗	44.50	81.00	77.00	91.00	91.00	93.00	79.58	3.15	2.64	3.36	1.00	5.00	5.00	1.50	3.09	2.62
VITA 1.5 [17]	7B	1 fps	✓	✗	18.00	46.00	40.00	88.00	49.00	97.50	56.42	2.40	2.08	0.57	0.85	4.57	4.49	1.18	2.31	1.54
Ola [35]	7B	1 fps	✓	✗	27.00	67.00	44.00	89.00	69.00	98.50	65.75	2.95	1.81	2.67	0.55	4.71	3.67	1.52	2.55	1.86
FlashVstream [64]	7B	1 fps	✗	✓	34.00	16.00	48.00	75.00	33.00	99.50	50.92	1.75	1.63	1.31	0.67	4.88	4.61	0.70	2.22	1.24
Dispider [41]	7B	16	✗	✓	38.50	70.00	44.00	69.00	100.00	99.50	70.17	2.50	1.75	4.06	0.91	0.61	2.49	2.07	2.06	1.63
ViSpeak (Ours, s3)	7B	1 fps	✓	✓	56.50	72.00	83.00	93.00	79.00	99.00	80.42	3.75	2.63	3.84	1.07	4.95	3.15	3.36	3.25	2.76

Table 6. Performance of various MLLMs on ViSpeak-Bench. Results for ViSpeak trained after the third stage are reported. For human evaluation, we invite 5 participants which are not received relevant training to answer 20% randomly selected questions and we report their average scores and the maximum scores on each subtask.

video to 64 accordingly, to extend the context. By default, videos are sampled at 1 fps. All experiments are conducted on 32 NVIDIA L20 GPUs and the max context length is set to 6,200 due to resource limit. The threshold for the informative head is set to 0.35 for all experiments.

5.2. Streaming Video Understanding Benchmarks

In this subsection, we select two large-scale comprehensive streaming video understanding benchmarks for evaluation.

Performance on StreamingBench. StreamingBench [29] is a comprehensive benchmark for streaming video understanding. As shown in Table 4, our ViSpeak model achieves SOTA performance among open-sourced models with only 7B parameters. And the performance is also comparable with GPT-4o, which is a well-known model for its omni-source understanding and interactive ability. And the performance of ViSpeak on omni-source understanding is even higher than GPT-4o (61.60 vs 50.95), demonstrating its outstanding omni-modal comprehensive understanding. Further, with our informative head, our model can speak proactively and get 38.80 scores on PO tasks, while other models should change the proactive problem to an offline one. After the stage 3 finetuning, the proactive output ability is further enhanced and gets 50.80 scores.

Performance on OVO-Bench. OVO-Bench [27] is designed for evaluating the backward tracing ability, the real-time visual perception ability and the forward active responding ability. As shown in Table 5, our ViSpeak model also achieves SOTA performance among open-sourced models and the performance is even higher than that of GPT-4o, showing a great ability to handle time-sensitive characteristics in video streaming understanding.

5.3. ViSpeak-Bench

On ViSpeak-Bench, we evaluate both representative proprietary and open-source MLLMs. We also conduct a human evaluation as a reference. Results are shown in Table 6.

For human evaluation, we find that in most cases, humans are able to provide appropriate responses at a suitable time and achieve the highest score. But the scores on AW, HR, and GU are relatively low. For Anomaly Warning and Humor Reaction, participants overlook some details in their descriptions, leading to a reduction in scores. And participants sometimes fail to accurately describe the gestures depicted in the videos in the gesture understanding subtask.

During testing MLLMs, we observed existing models perform poorly when not given explicit prompts to indicate the exact expected response type, because these models are unaware they are in a conversational scenario. To ensure the reliability of the evaluation of these models, we provide clear prompts for different subtasks (see Supplement for more details). With explicit prompts, all models achieve stable performance. Some observations are concluded as follows: a) Due to its great interactive ability, GPT-4o performs best among all models. b) Within open-sourced offline models, Qwen2.5-VL[1] performs best and a larger model can get more reasonable responses. c) For open-sourced omni-modality models Ola [35] and VITA 1.5 [17], their performances in both time accuracy and text score are inferior to models like InternVL-2.5 [5] and Qwen2.5-VL [1], possibly because they prioritize omni-modality, resulting in a weaker focus on visual understanding. d) For streaming video LMMs, FlashVstream [64] and Dispider [41] still underperform Qwen2.5-VL. We find that FlashVstream tends to speak aggressively, always prior to the actions or events, especially for VI and VT in which the actions are not at the beginning of the video. Additionally, we also use the same prompts for evaluating MMDuet [51] and VideoLLM-online [2], but we find they can not follow the instructions and simply describe the video, *e.g.* "You look at the camera.", which is possibly due to their selected training datasets.

In contrast, without explicit prompts, ViSpeak achieves the highest scores among open-source models, owing to fine-tuning from our strong streaming model. However, we observed that the performance on the anomaly warning and humor reaction subtasks is relatively low, as these tasks exhibit considerable variability in real-world scenarios, and

Method	MME	MVBench	Video-MME
VITA 1.5	2353.5 (1728.9/624.6)	53.95	58
Adaptive Sum	2237.0 (1636.3/600.7)	54.12	55
Linear	2283.4 (1685.5/597.9)	52.95	56
Add	2292.8 (1691.4/601.4)	54.27	55

Table 7. Ablation studies on input stream combination methods

Exp	Head	Joint	Token	Real	Omni	Context	All	PO
(a)	LM	✓	<seg>					30.00
(b)	inform	✗	<seg>	73.88	51.70	37.70	60.91	34.80
(c)	inform	✗	Visual					36.00
(d)	inform	✓	Visual	74.36	53.50	39.60	62.00	38.80

Table 8. Ablation studies on the design of visual proactive speaking control. Performances on StreamingBench are reported. 'Head' denotes using language modeling head or informative head for prediction. 'Joint' denotes whether the head is finetuned with LLM. 'Token' means which token is used for prediction.

understanding humor is difficult for MLLMs without reasoning ability.

5.4. Ablation Study

Effect of different combination methods to combine two input streams. In the ViSpeak model, we propose to use a two-stream chat template to support interactions within streaming videos. Two input streams (one for the user and one for the agent) are combined into a single one before sending to LLM. We design three types of combination methods: 'Adaptive Sum', 'Linear' and 'Add'. The 'Add' method directly adds two streams into a single one along the feature channel dimension. The 'Linear' method first concatenates two streams along the feature channel dimension and then uses a linear layer to reduce the dimension. The 'Adaptive Sum' method first predicts a weight for each stream and then weighted adds two streams. The intuition behind 'Adaptive Sum' is that two inputs may not have equal importance at a specific timestamp. When models are generating responses, they may focus more on their previous output tokens, whereas they pay more attention to user input tokens otherwise. In these experiments, we select MME [14] for image understanding ability evaluation and Video-MME [15] and MVBench [25] for the evaluation of video understanding capacity. As shown in Table 7, after the first template alignment stage, the model can maintain its offline data understanding ability, achieving a performance comparable to our baseline VITA 1.5. Further, we find that different combination methods also perform similarly and we use the 'Adaptive Sum' method by default.

Effect of different designs to control visual proactive output. In this work, we jointly train an informative head with the LLM to control visual proactive output. In Table 8, we ablate different designs. In Exp (a) to (c), we first train the model except the informative head. Then, we freeze the LLM and train the informative head in Exp (b) and (c). We find that using the language modeling head for proactive control gets limited performance with only 30.00 scores. Training an informative head following MMDuet [51] on the frozen LLM can get 34.80 scores. We further find that the last visual token in a segment contains more visual cues than the <seg> token so training the informative head based on the visual token can further improve the proactive output score to 36.00. Since the LLM in Exp (a) to (c) are frozen, the performance of other tasks in StreamingBench is the same across these experiments. In Exp (d), we jointly train the informative head with LLM and get the highest proactive output performance. We find that other tasks in StreamingBench are also improved by co-training. We speculate that the informative head makes the model aware of the action boundary thus improving the performance on other tasks.

Dataset	HR (Text Score)	GU (Text Score)	Overall
ViSpeak-Instruct	1.07	3.36	2.76
w/o offline data	1.02	3.17	2.70

Table 9. Ablation studies on the offline data in ViSpeak-Instruct. Performance on ViSpeak-Bench are reported.

Effect of the different dataset composition of ViSpeak-Instruct. Since there are many kinds of gestures in conversations and the gesture in different contexts has different meanings. To enhance the gesture understanding ability, we use some offline data during training, as well as for humor reactions. As shown in Table 9, using offline data can increase the generalization ability and get higher scores.

6. Conclusion

In this work, we extend the streaming video understanding problem with a new Visual Instruction Feedback task, which requires the model to respond to visual contents actively. To facilitate research, we define seven key subtasks and collect the ViSpeak-Bench for evaluation and the ViSpeak-Instruct for training. To solve this problem, we first adapt an offline omni-modal LMM to our designed chat template, and then finetuning it to get a SOTA streaming LMM. This model is evaluated on two comprehensive streaming benchmarks and gets GPT-4o-level performance. Finally, we finetune the SOTA streaming model on ViSpeak-Instruct and get the ViSpeak model which serves as a strong baseline on ViSpeak-Bench for future research. We hope our work can provide deeper insights into streaming video understanding and human-agent interaction.

Acknowledgments. This work was supported partially by the NSFC(92470202, U21A20471), Guangdong NSF Project (No. 2023B1515040025), the Major Key Project of PCL (PCL2024A06), and the Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026). This work was also supported by Alibaba Research Intern Program.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[2] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024. 3, 6, 7

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 2, 6

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024. 1, 2

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 6

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 6

[10] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. 5

[11] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, 2024. 6

[12] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024. 6

[13] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, 2020. 3, 4

[14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 8

[15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3, 8

[16] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 1, 2, 3, 5, 6

[17] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 1, 2, 3, 5, 6, 7

[18] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *CVPR*, 2025. 2

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7

[20] Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. Smile: Multimodal dataset for understanding laughter in video with language models. *arXiv preprint arXiv:2312.09818*, 2023. 2, 4

[21] Qing Jiang, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, Lei Zhang, et al. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024. 2

[22] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 2

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 6

[24] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11963–11974, 2023. 4

[25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 3, 8

[26] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025. 2

[27] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *CVPR*, 2025. 2, 3, 6, 7

[28] Yuan-Ming Li, An-Lan Wang, Kun-Yu Lin, Yu-Ming Tang, Ling-An Zeng, Jian-Fang Hu, and Wei-Shi Zheng. Techcoach: Towards technical keypoint-aware descriptive action coaching. *arXiv preprint arXiv:2411.17130*, 2024. 2

[29] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. 2, 3, 6, 7

[30] Kun-Yu Lin, Jia-Run Du, Yipeng Gao, Jiaming Zhou, and Wei-Shi Zheng. Diversifying spatial-temporal perception for video domain generalization. In *NeurIPS*, 2023. 2

[31] Kun-Yu Lin, Henghui Ding, Jiaming Zhou, Yi-Xing Peng, Zhilin Zhao, Chen Change Loy, and Wei-Shi Zheng. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *CoRR*, abs/2403.01560, 2024.

[32] Kun-Yu Lin, Jiaming Zhou, and Wei-Shi Zheng. Human-centric transformer for domain adaptive action recognition. *TPAMI*, 2025. 2

[33] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. In *NeurIPS*, 2024. 2, 3, 6

[34] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. In *ICLR*, 2025. 2

[35] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025. 1, 2, 6, 7

[36] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *ICCV*, 2019. 4

[37] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 6

[38] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015. 6

[39] Yi-Xing Peng, Qize Yang, Yu-Ming Tang, Shenghao Fu, Kun-Yu Lin, Xihan Wei, and Wei-Shi Zheng. Actionart: Advancing multimodal large models for fine-grained human-centric video understanding. *arXiv preprint arXiv:2504.18152*, 2025. 2

[40] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *NeurIPS*, 2024. 2

[41] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *CVPR*, 2025. 1, 3, 6, 7

[42] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019. 4

[43] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 6

[44] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6, 7

[45] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[46] An-Lan Wang, Bin Shan, Wei Shi, Kun-Yu Lin, Xiang Fei, Guozhi Tang, Lei Liao, Jingqun Tang, Can Huang, and Wei-Shi Zheng. Pargo: Bridging vision-language with partial and global views. In *AAAI*, 2025. 2

[47] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024. 2

[48] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 2

[49] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 6

[50] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *ECCV*, 2024. 2

[51] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. *arXiv preprint arXiv:2411.17991*, 2024. 1, 3, 5, 6, 7, 8

[52] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 3

[53] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *ECCV*, 2024. 2, 3, 4

[54] Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024. 3

[55] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *ICLR*, 2025. 3

[56] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024. 2

[57] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024. 6

[58] Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502*, 2025. 2

[59] Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025. 2

[60] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *ICLR*, 2025. 3

[61] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6

[62] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 3

[63] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, 2019. 4

[64] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024. 3, 6, 7

[65] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171*, 2024. 3, 4

[66] Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, et al. Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024. 2, 3, 6

[67] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 2

[68] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1, 2, 6

[69] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2, 6

[70] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025. 2