InternLM-XComposer2.5-OmniLive: A Comprehensive Multimodal System for Long-term Streaming Video and Audio Interactions

Anonymous ACL submission

Abstract

Creating AI systems that can interact with environments over long periods, similar to human cognition, has been a longstanding research goal. Recent advancements in multimodal large language models (MLLMs) have made significant strides in open-world understanding. However, the challenge of continuous and simultaneous streaming perception, memory, and reasoning remains largely unexplored. Current MLLMs are constrained by their sequence-to-sequence architecture, which limits their ability to process inputs and generate responses simultaneously, akin to being unable to think while perceiving. Furthermore, relying on long contexts to store historical data is impractical for long-term interactions, as retaining all information becomes costly and inefficient. Therefore, rather than relying on a single foundation model to perform all functions, this project draws inspiration from the concept of the **Specialized Generalist AI** and introduces disentangled streaming perception, reasoning, and memory mechanisms, enabling real-time interaction with streaming video and audio input. The proposed framework **InternLM-XComposer2.5-OmniLive (IXC2.5-OL)** consists of three key modules: **(1) Streaming Perception Module**: Processes multimodal information in real-time, storing key details in memory and triggering reasoning in response to user queries. **(2) Multi-modal Long Memory Module:** Integrates short-term and long-term memory, compressing short-term memories into long-term ones for efficient retrieval and improved accuracy. **(3) Reasoning Module:** Responds to queries and executes reasoning tasks, coordinating with the perception and memory modules. This project simulates human-like cognition, enabling multimodal large language models to provide continuous and adaptive service over time. All code and models of **InternLM-XComposer2.5-OmniLive (IXC2.5-OL)** will be publicly available.

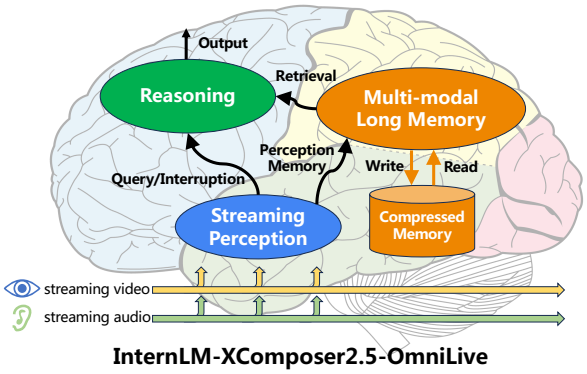

Figure 1: Inspired by human-like cognition and Specialized Generalist AI, we introduce InternLM-XComposer2.5-OmniLive (IXC2.5-OL), a system that facilitates real-time interaction with: (1) a **streaming perception** module supports streaming video and audio inputs; (2) a **multi-modal long memory** module that compresses short-term memory into long-term memory; and (3) a **reasoning** module that answers queries based on retrieved memories.

1 Introduction

The goal of developing AI systems (LeCun, 2022) that can understand and interact with environments over long periods, akin to human cognition, has been a central focus of research for decades. The rise of large-scale data corpora (Lin et al., 2014; Kuznetsova et al., 2020; Schuhmann et al., 2022; Wang et al., 2023) and multimodal large language models (OpenAI, 2023b, 2024; Team, 2023) has driven significant advances in free-form multimodal question answering. Recent developments, such as Mini-Omni (Xie and Wu, 2024a), VideoLLM-Online (Chen et al., 2024a), and VITA (Fu et al., 2024b), have made notable strides toward enabling more natural and immersive online interactions. However, challenges persist in creating systems capable of continuous interaction due to the intrinsic limitations of a single decoder-only large language model architecture.

Existing architectures (Zhang et al., 2024e; Xie and Wu, 2024a; Chen et al., 2024a; Fu et al., 2024b) encounter significant limitations in real-

time and long-term streaming perception, reasoning, and memory. The sequence-to-sequence decoder-only architecture used in current MLLMs forces a switch between perception (e.g., seeing and hearing) and thinking, limiting the simultaneous processing of inputs and outputs. Additionally, existing works (Zhang et al., 2024b; Wang et al., 2024e; Fan et al., 2024) rely on the integration of multimodal memories within context windows. The reliance on long contexts to store historical information proves impractical for long-term use, especially in scenarios requiring continuous AI assistance. Multimodal data, like video streams, can quickly accumulate millions of tokens within a few hours, making it impractical to maintain context over multiple days of service. The cost and inefficiency of storing all historical clues within the context further limit the system's capacity to provide continuous and long-term service. In contrast, the human brain can effortlessly integrate perception and cognition, preserving long-term multimodal memories. This is believed to be closely related to the functional partitioning design of the human brain cortex, where different areas of the cortex are responsible for distinct tasks, such as perception, memory, and cognition.

Inspired by the paradigm of Specialized Generalist AI (Zhang et al., 2024c), we propose a system **InternLM-XComposer2.5-OmniLive (IXC2.5-OL)** composed of fused specialized generalist models for streaming perception, reasoning, and memory, respectively. The system is designed to enable AI models to engage continuously with environments while retaining observations over time. By integrating short-term and long-term multimodal memory, our approach attempts to emulate human-like cognition, enabling more dynamic and sustained interactions.

As shown in Figure 1, the IXC2.5-OL system consists of three key modules: (1) **Streaming Perception Module:** This module processes the multimodal information stream on-the-fly. To ensure perception accuracy and efficiency, the video and audio streams are handled separately. A live video perception model processes the video stream, encoding the information and storing key details in memory. Meanwhile, an audio model recognizes the contents of human speech and other sounds, , barking, knocking, or whistling. It triggers the reasoning process when human queries occur. (2) **Multi-modal Long Memory Module**: This component integrates both long-term and short-term memory, enabling the retrieval of detailed short-term information as well as long-term historical cues. It continuously compresses short-term memories into more information-rich long-term memories to enhance retrieval efficiency and accuracy. (3) **Reasoning Module**: The reasoning module, activated by the perception module, handles queries and performs reasoning tasks. As the component with the most model parameters, it serves as the core of the system's deep cognitive processes.

The proposed system empowers AI with the ability to perceive, think, and memorize simultaneously. By overcoming the limitations of alternating perception and reasoning, IXC2.5-OL seeks to provide continuous, adaptive service, and long-term AI service. The proposed system will not only enhance the performance of AI assistants but will also contribute to the broader AI applications capable of continuously interacting and adapting to dynamic environments.

The **IXC2.5-OL** demonstrates strong performance across both audio and video benchmarks. Among the open-source models, IXC2.5-OL achieves competitive results on audio recognition (ASR) benchmarks such as Wenetspeech (Zhang et al., 2022) for Chinese and LibriSpeech (Panayotov et al., 2015) for English. For video understanding benchmarks, IXC2.5-OL achieves state-of-the-art results among models with less than 10B parameters, obtaining an M-Avg of 66.2% on MLVU (Zhou et al., 2024) and an overall accuracy of 68.7% on MVBench (Li et al., 2024d). Additionally, it demonstrates competitive performance on Video-MME (Fu et al., 2024a) (60.6%) and MMBench-Video (Fang et al., 2024) (1.42). On recent streaming video bench StreamingBench (Lin et al., 2024b), IXC2.5-OL achieves new SOTA results on open-source models (73.79%), highlighting its exceptional capabilities for real-time video interactions.

To foster the development of the multimodal streaming interaction community, alongside the model parameters, the inference and deployment source code, encompassing both the web frontend and backend code, will be released.

2 Related Works

MLLMs for Text-Image Conversation. Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Qwen, 2023; Cai et al., 2024) have garnered significant attention for their re-

markable capabilities in language comprehension and generation. Building on this success, Large Vision-Language Models (LVLMs) (Zhang et al., 2023d; OpenAI, 2023a; Chen et al., 2023; Dong et al., 2024; Lin et al., 2024a) have been developed by integrating LLMs with vision encoders (Radford et al., 2021; Zhang et al., 2024a; Chen et al., 2024e; Zhang et al., 2024f), extending their ability to comprehend visual content and enabling applications like text-image conversations. Earlier LVLMs were primarily designed for single-image, multi-round conversations, whereas recent advancements (Alayrac et al., 2022; Bai et al., 2023; Dong et al., 2024; Zhang et al., 2024d; Li et al., 2024b) have expanded their capabilities to process and understand multi-image inputs.

MLLMs for Video Understanding. In addition to advancements in image understanding, the field of MLLMs has seen growing efforts in video analysis (Li et al., 2023b; Ning et al., 2023; Wang et al., 2024a). To address the complexity of video inputs, existing approaches leverage techniques such as sparse sampling or temporal pooling (Lin et al., 2023; Maaz et al., 2023; Huang et al., 2024; Yu et al., 2024a), compressed video tokens (Li et al., 2023a; Zhang et al., 2023c; Chen et al., 2024c), and memory banks (Song et al., 2023; He et al., 2024; Fan et al., 2024). Additionally, some methods utilize language as a bridge for video understanding (Kahatapitiya et al., 2024; Zhang et al., 2023a). Beyond these video-specific strategies, video analysis can also be framed as interpreting a high-resolution composite image generated from sampled video frames (Kim et al., 2024; Xu et al., 2024; Zhang et al., 2024e). Recent advancements (Chen et al., 2024a; Wang et al., 2024d; Wu et al., 2024; Zhang et al., 2024b) have increasingly focused on online video understanding, aiming to simulate real-world scenarios where AI processes video streams in real-time to comprehend the environment on-the-fly. However, existing solutions still lack the capability to simultaneously perform perception, memory, and reasoning, limiting their applicability for consistent and long-term human-AI interactions.

MLLMs for Audio Understanding. Audio understanding can be effectively modeled as a sequence-to-sequence (Seq2Seq) task (Radford et al., 2023), which enables powerful integration with large language models by incorporating audio tokenizers and encoders (Zhang et al., 2023b; Chu et al., 2023; Zeng et al., 2024). In addition to receiving the audio input, recent research investigates streaming duplex speech models (Wang et al., 2024b; Yu et al., 2024b; Wang et al., 2024c) that allow speakers to interrupt freely. Beyond audio-text models, emerging research delves into audio-visual models (Shu et al., 2023; Li et al., 2024c) and unified architectures that process audio, visual, and text modalities (Zhan et al., 2024; Li et al., 2024e).

MLLMs for Omni-Modal Understanding. Integrating multiple modalities into a single omni-modal foundation model represents a promising research direction. Existing works (Han et al., 2023; Zhan et al., 2024; Xie and Wu, 2024b; Sun et al., 2024) explore models capable of processing omni-modal inputs, typically combining video and audio, to produce outputs in various formats. These outputs include text (Han et al., 2023; Fu et al., 2024b), audio (Xie and Wu, 2024b), and omni-modal contents (Zhan et al., 2024). In the current design of IXC2.5-OL, we handle the audio and video modalities separately to mitigate potential influence during joint training. In future versions, our model will incorporate joint training across all modalities, enabling seamless omni-modality integration.

3 Method

As we briefly introduced in Sec.1, the IXC2.5-OL has three disentangled modules: 1) the Streaming Perception Module for on-the-fly visual and audio information processing, 2) the Multi-modal Long Memory Module for memory integration and retrieval, and 3) the Reasoning Module collect information from the perception and memory module, and handles queries and performs reasoning tasks. All the modules work simultaneously and interact asynchronously.

3.1 Streaming Perception Module

Besides nature language, the IXC2.5-OL could handle video and audio. To realize this, the Streaming Perception Module contains an Audio Translation Module and a Video Perception Module.

Audio Translation Module contains an audio encoder, an audio projector, and a Small Language Model (SLM). The audio encoder encodes the input audio sample into high-dimension features, and the audio projector further maps the feature to the input space of the SLM. The SLM outputs both the class (e.g. laughing, clapping, or raining) of the audio and the natural language within the audio (i.e. the automatic speech recognition). In practice,

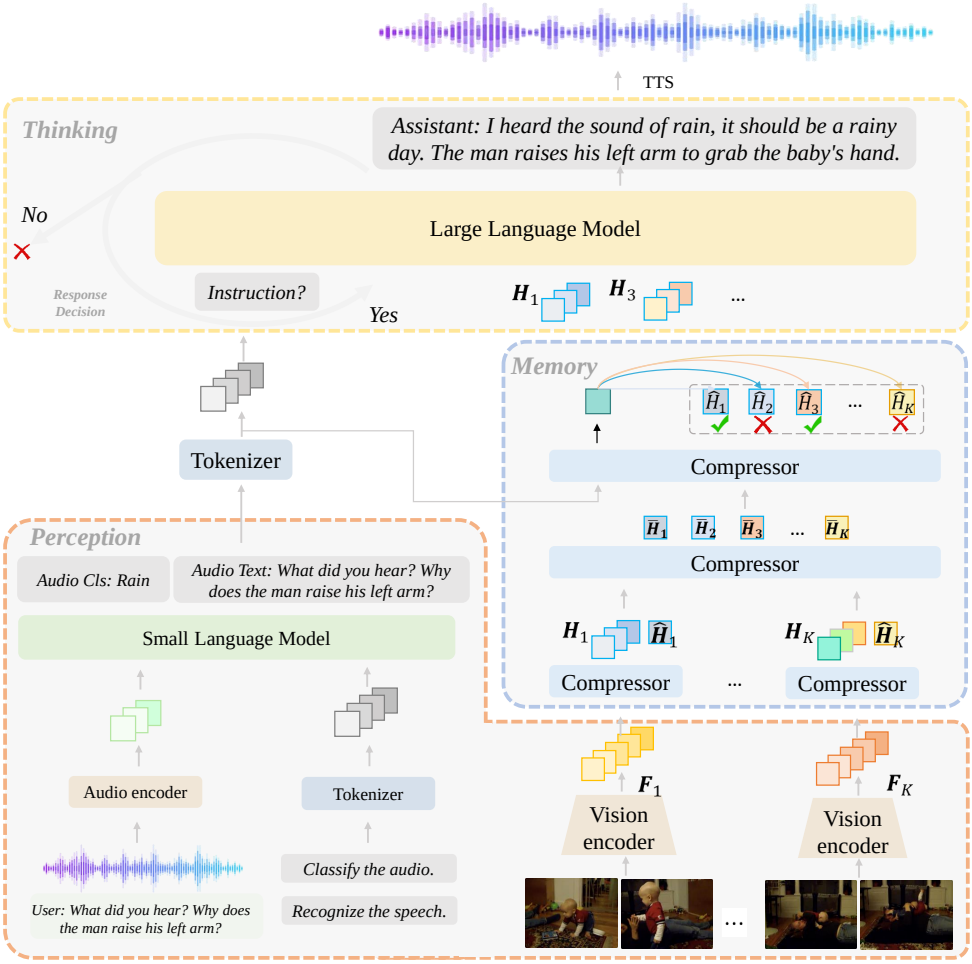

Figure 2: **Pipeline of the InternLM-XComposer2.5-OmniLive. (IXC2.5-OL)**. The IXC2.5-OL is a real-time interacting system that is constructed by three simultaneous modules: 1) the Streaming Perception Module, 2) the Multi-modal Long Memory Module, and 3) the Reasoning Module.

we use the Whisper (Radford et al., 2022) model as the audio encoder and a Qwen2-1.8B (Yang et al., 2024) as the SLM. The training contains two stages and we list the training data in Table 1.

Video Perception Module provides coarse-grained visual information to the Multi-modal Long Memory Module. It processes the real-time video input stream and encodes each frame into semantic features. For efficiency, we use the OpenAI CLIP-L/14 (Radford et al., 2021) In practice.

3.2 Multi-modal Long Memory Module

The Multi-modal Long Memory Module is the core design to handle extremely long video input and helps the Reasoning Module to get rid of millions of tokens from its context window. It shares a similar idea from the VideoStreaming (Qian et al., 2024) that encodes video clips into short-term memories and integrates them into long-term memory. With the given questions, it retrieved the most related video clips for the Reasoning Module. Formally, the Multi-modal Long Memory Module is trained with three tasks:

Video Clip Compression. With features of k_{th} video clip extracted from the Perception Module $F_k \in \mathbb{R}^{TN \times C}$, we initialize its short-term memory $H_k \in \mathbb{R}^{TP \times C}$ by the spatial down-sampling and its global memory $\hat{H}_k \in \mathbb{R}^{1 \times C}$. We realize the compression by the auto-regressive and feature aggregation nature of LLMs:

$$H_k, \hat{H}_k = Compressor([F_k \circ H_k \circ \hat{H}_k]).$$

Memory Integration. Short-term memory represents the detailed information of each short video clip while the model still lacks a macro view of the video. To this end, with the short-term and global memory of a list of video clips, we integrate them into long-term memory by the Compressor in the following format:

$$\bar{H}_1, \bar{H}_2, ..., \bar{H}_k = Compressor([H_1 \circ H_2 ... \circ H_k \circ \hat{H}_1 \circ \hat{H}_2 ... \circ \hat{H}_k]).$$

Table 1: Audio Datasets used in IXC2.5-OL. CLS denotes classification.

Stage	Task	Dataset	Data Num
Pretrain	ASR	GigaSpeech (Chen et al., 2021)	8,282,987
		WenetSpeech (Zhang et al., 2022)	17,821,017
SFT	ASR	LibriSpeech (Panayotov et al., 2015)	281,241
		VCTK (Veaux et al., 2017)	44,070
		AISHELL-1 (Bu et al., 2017)	120,098
		AISHELL-4 (Fu et al., 2021)	102,254
		MD-RAMC (Yang et al., 2022)	219,325
		ASCEND (Lovenia et al., 2021)	12,314
		KeSpeech (Tang et al., 2021)	888,428
		DASR (Cornell et al., 2024)	190,732
		CommonVoice (Ardila et al., 2019)	2,813,852
	CLS	FSD50K (Fonseca et al., 2020)	40,966
		AudioSet (Kong et al., 2018)	18,683
		Silence	475

Table 2: Video Datasets used in IXC2.5-OL.

Model	Dataset
Memory Module	ShareGPT4Video (Chen et al., 2024b), Ego4D(Grauman et al., 2022)
	ActivityNet (Fabian Caba Heilbron and Niebles, 2015)
	Semantics Implicit QA
	Reference Implicit QA
IXC2.5	ShareGPT4Video (Chen et al., 2024b)
	ActivityNet (Fabian Caba Heilbron and Niebles, 2015)
	FunQA (Xie et al., 2025), TrafficQA (Xu et al., 2021)
	VideoChat2-IT(Li et al., 2023b), LLaVA-Video (Zhang et al., 2024h)

the $\bar{H} = [\bar{H}_1, \bar{H}_2, ..., \bar{H}_k] \in \mathbb{R}^{k \times C}$ represents the video in a high-compressed way and we denote it as the long-term memory.

Video Clip Retrieval. When users raise questions, the Multi-modal Long Memory Module retrieves the question-related video clips and provides both the video clips and their short-term memory to the Reasoning Module. In practice, we first encode the question to the feature space of the memory. We concatenate the long-term memory with the tokenized question as the Compressor input, and we view the last token of the output features as the memory-space-aligned question feature. Then we calculate the similarity between the question feature and each video's global memory, and select the most related clips for the Reasoning Module.

Implementation Detail. We use Qwen2-1.8B (Yang et al., 2024) as the LLMs and construct several kinds of training data for the three aforementioned tasks. As shown in Table. 2, we train the Video Clip Compression task with short video captioning data from multiple sources, using the same prefix captioning task designed in VideoStreaming (Qian et al., 2024). For the Memory Integration task and Video Clip Retrieval task, besides the off-the-shelf video grounding data, we also construct data for two unique tasks: 'Semantics Implicit Question' and 'Reference Implicit Question'.

The 'Semantics Implicit Question' means the question does not point to some object directly, but mentions the usage or meaning of the object, and the model should find out the object by understanding the implicit question. For example, when the user asks 'How about the weather today?', the model should find out some weather-related object in the past video stream, such as an umbrella, a sun-glass, or something. Another example could be 'I'm hungry, where can I heat my sandwiches?', the model should find the microwave oven it has seen before.

The 'Reference Implicit Question' means the question uses pronouns rather than nouns. For example, 'What is this' means the models should retrieve the current frames, although it does not mention any exact objects.

Both kinds of implicit questions are commonly used in real-world communication while current models failed to handle them, so we construct corresponding training data to empower the model with these capabilities.

3.3 Reasoning Module

The Reasoning Module is initialized by an improved version of InternLM-XComposer2.5 (IXC2.5 in the following for simplified statement) and we add a memory projector to align the memory feature with IXC-2.5. For a given questions and both visual and memory information provided by the Memory Module, we formulate the input as:

Question: $<$ |Que| $>$,

Here is the question related video clip $<$ |Img| $>$;

Here is the question related memory $<$ |Mem| $>$

In real-world usage, there exists some noisy input that should not be answered (e.g., the user says 'enn...' or 'ok...'), the model should keep salient util the next question. To realize this, we add an additional 'Instruction Prediction' process for each question to decide it should be answered or not.

3.4 System Pipeline

As illustrated in Figure 3, the system comprises the Frontend, SRS Server, and Backend Server.

Frontend. The frontend application, developed with JavaScript, enables the camera and microphone to capture video and audio stream inputs, which are then pushed to the SRS server. Concurrently, it establishes a WebSocket connection with the backend to listen for audio outputs and interrupt signals. When audio data is received, the frontend plays it. Upon receiving an interrupt signal, the frontend suspends the audio playback and discards the pending audio.

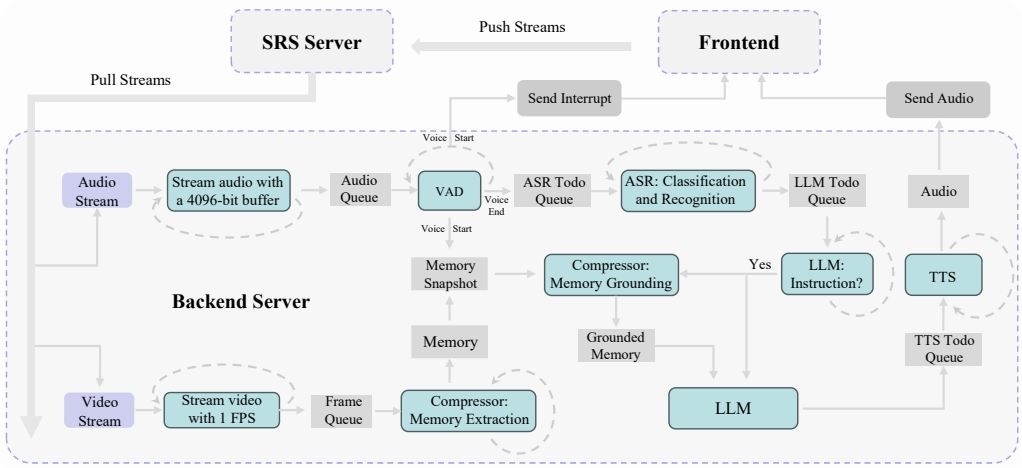

Figure 3: **System pipeline of the IXC2.5-OL**. The system comprises the Frontend, SRS Server, and Backend Server. The Frontend is utilized for capturing video and audio streams and for playing audio from the Backend Server. The SRS Server is employed for managing live streams. The Backend Server is responsible for reading audio and video, extracting memory, and answering questions. The green boxes represent a thread or a process.

SRS Server. SRS (Simple Realtime Server) is a straightforward and efficient real-time video server, adept at supporting a multitude of real-time streaming protocols such as RTMP, WebRTC, HLS, HTTP-FLV, SRT, and others. It is renowned for its ability to reliably receive and deliver audio and video streams.

Backend Server. After establishing a WebSocket connection with the frontend, the backend will pull streaming from the SRS Server and initiate separate threads to read audio and video.

The audio reading thread will segment the audio stream into 4096-bit chunks and enqueue them into the *Audio Queue*. The Voice Activity Detection (VAD) (Gao et al., 2023) thread continuously reads data from *Audio Queue* and detects the start and end of voice activity. Upon detecting the start of voice activity, the backend sends an interrupt signal to the frontend to pause the currently playing audio, and at the same time, dispatches a backup signal to the video process, directing it to save the current memory state. When detecting the end of voice activity, the entire voice segment will be enqueued into *ASR Todo Queue*. The ASR thread continuously reads audio segments from *ASR Todo Queue*, performs background noise classification and voice recognition on them, and then enqueues the results into *LLM Todo Queue* for use by the LLM.

The video reading thread reads video frames at a rate of 1 frame per second and enqueues them into *Frame Queue*. The compressor process reads video frames from the queue, recognizes them, extracts relevant memory, and stores it. Upon receiving a backup signal from the VAD thread, the compressor process will save the current memory state for later retrieval.

The LLM process reads text from the *LLM Todo Queue* and determines whether it is an instruction that requires a response from the model. For texts identified as instructions, the compressor process will use the current instruction and the backed-up memory to perform memory grounding, in order to retrieve memories related to the instruction. The LLM process will then generate a response based on the retrieved memories and the instruction, and enqueue the resulting output into *TTS Todo Queue*. An additional TTS thread (, F5-TTS (Chen et al., 2024d), MeloTTS (Zhao et al., 2023)) will convert the text from the *TTS Todo Queue* into audio and send it to the frontend.

4 Experiments

In this section, we validate the benchmark performance of our InternLM-XComposer2.5-OmniLive (IXC2.5-OL), including both audio and video benchmarks.

4.1 Audio Benchmarks

We evaluate our audio models on two prominent ASR benchmarks: Wenetspeech (Zhang et al., 2022) for Chinese (CN) and LibriSpeech (Panayotov et al., 2015) for English (EN). WenetSpeech includes two test sets: Test_Net, which represents high-quality and relatively clean Chinese speech, and Test_Meeting, which captures more challenging conversational scenarios. LibriSpeech consists of four splits: Dev_clean and Test_clean, which contain clean, high-quality English speech, and Dev_other and Test_other, which include noisier, more complex utterances.

Table 3: **Results on ASR tasks**: "CN" refers to Chinese speech, while "ENG" refers to English speech. The performance is measured using WER ↓ (Word Error Rate).

Method	LLM	Wenetspeech (CN)		Librispeech (ENG)			
		Test_Net ↓	Test_Meeting ↓	Dev_clean ↓	Dev_other ↓	Test_clean↓	Test_other↓
Qwen2-Audio (Chu et al., 2024)	Qwen2-7B (Yang et al., 2024)	7.8	8.4	1.3	3.4	1.6	3.6
Mini-Omni (Xie and Wu, 2024a)	Qwen2-0.5B (Yang et al., 2024)	-	-	4.5	9.7	4.6	9.2
VITA (Fu et al., 2024b)	Mixtral-8x7B (Jiang et al., 2024)	12.2	16.5	7.6	16.6	8.1	18.4
IXC2.5-OL	Qwen2-1.5B (Yang et al., 2024)	9.0	9.2	2.5	5.7	2.6	5.8

Table 4: **Results on MLVU benchmark.** IXC2.5-OL has demonstrated excellent performance, surpassing both open-source models and closed-source APIs, achieving SOTA at the 7B model scale.

Method	Params	Topic Rea.	Anomaly Recog.	Needle QA	Ego Rea.	Plot QA	Action Or.	Action Co.	M-Avg
Closed-source APIs.									
Claude-3-Opus	-	67.2	43.5	21.6	40.2	47.8	18.2	16.7	36.5
Qwen-VL-Max	-	67.4	63.5	40.3	40.9	43.3	25.0	14.8	42.2
GPT-4 Turbo	-	79.5	68.0	45.9	47.4	60.6	26.5	16.1	49.2
GPT-4o	-	87.4	74.5	64.8	57.1	65.1	56.7	46.3	64.6
Open-source models.									
MovieChat (Song et al., 2024a)	7B	29.5	25.0	24.2	24.7	25.8	28.6	22.8	25.8
LLaMA-VID (Li et al., 2025)	7B	50.8	34.5	30.1	32.7	32.5	23.9	27.8	33.2
LLaVA-1.6 (Liu et al., 2024a)	7B	60.6	41.0	43.1	38.4	41.0	25.5	25.7	39.3
ShareGPT4Video (Chen et al., 2024b)	7B	75.8	51.5	47.6	43.2	48.4	34.0	23.3	46.4
VideoLlaMA2 (Cheng et al., 2024)	7B	74.6	64.5	49.9	43.8	45.1	34.0	27.4	48.5
LongVA (Zhang et al., 2024e)	7B	83.3	58.5	69.3	50.0	67.2	38.6	27.2	56.3
IXC2.5 (Zhang et al., 2024d)	7B	-	-	-	-	-	-	-	58.8
InternVL2 (Chen et al., 2024e)	8B	-	-	-	-	-	-	-	64.0
LLaVA-OneVision (Li et al., 2024a)	7B	-	-	-	-	-	-	-	64.7
Video-XL (Shu et al., 2024)	7B	-	-	-	-	-	-	-	64.9
IXC2.5-OL	7B	84.1	68.5	76.6	60.8	75.1	57.1	41.3	66.2

Table 5: **Results on Video-MME benchmark.** IXC2.5-OL demonstrates performance close to that of the open-source SOTA.

Method	Params	Short	Medium	Long	Overall
Closed-source APIs.					
GPT-4V	-	70.5	55.8	53.5	59.9
Claude 3.5 Sonnet	-	71.0	57.4	51.2	60.0
GPT-4o mini	-	72.5	63.1	58.6	64.8
GPT-4o	-	80.0	70.3	65.3	71.9
Gemini 1.5 Pro	-	81.7	74.3	67.4	75.0
Open-source models.					
ShareGPT4Video (Chen et al., 2024b)	7B	48.3	36.3	35.0	39.9
VideoLlaMA2 (Cheng et al., 2024)	7B	-	-	-	47.9
LongVA (Zhang et al., 2024e)	7B	61.1	50.4	46.2	52.6
Video-XL (Shu et al., 2024)	7B	64.0	53.2	49.2	55.5
VITA (Fu et al., 2024b)	8×7B	65.9	52.9	48.6	55.8
IXC2.5 (Zhang et al., 2024d)	7B	-	-	-	55.8
InternVL2 (Chen et al., 2024e)	8B	-	-	-	56.3
LLaVA-OneVision (Li et al., 2024a)	7B	-	-	-	58.2
mPLUG-Owl3 (Ye et al., 2024)	7B	70.0	57.7	50.1	59.3
MiniCPM-V 2.6 (Yao et al., 2024)	8B	-	-	-	60.9
IXC2.5-OL	7B	72.7	58.2	50.8	60.6

As shown in Table 3, our IXC2.5-OL demonstrates superior performance compared to recent streaming audio LLMs such as VITA and Mini-Omni, particularly achieving lower Word Error Rates (WER) across both CN and EN benchmarks with merely a lightweight 1.5B LLM.

4.2 Video Benchmarks

In Tables 4, 5, 8 and 7, we compare IXC2.5-OL with both closed-source APIs and open-source models on conventional video understanding benchmarks, including MLVU (Zhou et al., 2024), Video-MME (Fu et al., 2024a), MMBench-Video (Fang et al., 2024) and MVBench (Li et al., 2024d). Furthermore, we also assess the performance of different models on the recently proposed StreamingBench (Lin et al., 2024b), which is designed to better evaluate performance for real-time video interactions. The results of this comparison are presented in Table 6. For the video benchmarks, the base model utilizes 64 sampled frames for each video during evaluation.

MLVU MLVU is a comprehensive benchmark designed for evaluating Multimodal Large Language Models in Long Video Understanding tasks. The videos range from 3 minutes to 2 hours and include nine distinct evaluation tasks. Here, we evaluate seven multi-choice tasks, including Topic Reasoning, Anomaly Recognition, Needle QA, Ego Reasoning, Plot QA, Action Order, and Action Count. The detailed comparisons are given in Table 4. The IXC2.5-OL exhibits state-of-the-art (SOTA) performance among closed-source APIs, and open-source models with parameters less than 10 billion, surpassing the previous SOTA by 1.3% for Video-XL, 1.6% for GPT-4o.

Video-MME Video-MME is a high-quality video benchmark. The videos are collected from 6 primary visual domains with 30 subfields to ensure broad scenario generalizability, encompassing both short-, medium-, and long-term videos, ranging from 11 seconds to 1 hour. As demonstrated in Table 5, the IXC2.5-OL exhibits competitive performance on this benchmark, comparable to previous SOTA MiniCPM-V 2.6.

StreamingBench StreamingBench is a streaming video benchmark designed for real-time video evaluation. It comprises 18 tasks, showcasing 900 videos and 4,500 human-curated QA pairs. In this context, we focus on assessing visual understanding in real-time. Table 6 illustrates the comparative analysis, demonstrating that IXC2.5-OL

Table 6: **Results on StreamingBench** for Real-Time Visual Understanding. IXC2.5-OL excels among all open-source models, and falling just short of the Gemini 1.5 Pro.

Method	Params	Real-Time Visual Understanding										
		OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Overall
Human	-	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46
Closed-source APIs.												
Claude 3.5 Sonnet	-	80.49	77.34	82.02	81.73	72.33	75.39	61.11	61.79	69.32	43.09	72.44
GPT-4o	-	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
Gemini 1.5 Pro	-	79.02	80.47	83.54	79.67	80.00	84.74	77.78	64.23	71.95	48.70	75.69
Open-source models.												
VideoLLM-online (Chen et al., 2024a)	8B	39.07	40.06	34.49	31.05	45.96	32.40	31.48	34.16	42.49	27.89	35.99
VideoLLaMA2 (Cheng et al., 2024)	7B	55.86	55.47	57.41	58.17	52.80	43.61	39.21	42.68	45.61	35.23	49.52
VILA-1.5 (Lin et al., 2024a)	8B	53.68	49.22	70.98	56.86	53.42	53.89	54.63	48.78	50.14	17.62	52.32
LongVA (Zhang et al., 2024e)	7B	70.03	63.28	61.20	70.92	62.73	59.50	61.11	53.66	54.67	34.72	59.96
InternVL2 (Chen et al., 2024e)	8B	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Kangaroo (Liu et al., 2024b)	7B	71.12	84.38	70.66	73.20	67.08	61.68	56.48	55.69	62.04	38.86	64.60
MiniCPM-V 2.6 (Yao et al., 2024)	8B	71.93	71.09	77.92	75.82	64.60	65.73	70.37	56.10	62.32	53.37	67.44
Qwen2-VL (Wang et al., 2024a)	7B	75.20	82.81	73.19	77.45	68.32	71.03	72.22	61.19	69.04	46.11	69.04
LLaVA-OneVision (Li et al., 2024a)	7B	80.38	74.22	76.03	80.72	72.67	71.65	67.59	65.45	65.72	45.08	71.12
IXC2.5-OL	7B	82.83	73.77	78.66	82.95	72.50	76.01	61.11	60.67	71.59	58.85	73.79

Table 7: **Results on MVBench.** IXC2.5-OL shows SOTA results across open-source and closed-source models.

Method	Params	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	Avg
Closed-source APIs.																						
GPT-4V	-	55.5	63.5	72.0	46.5	73.5	18.5	59.0	29.5	12.0	40.5	83.5	39.0	12.0	22.5	45.0	47.5	52.0	31.0	59.0	11.0	43.5
GPT-4o	-	61.5	56.5	72.0	54.0	82.0	62.5	66.5	44.0	36.5	33.5	93.0	54.5	33.5	54.5	53.5	74.5	71.5	32.5	71.0	42.5	57.5
Open-source models.																						
VideoLLaMA (Zhang et al., 2023c)	7B	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0	34.1
VideoChat (Li et al., 2023a)	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
MiniCPM-V 2.6 (Yao et al., 2024)	7B	38.0	43.0	63.0	35.5	67.5	55.5	46.0	35.5	25.5	33.0	77.5	48.0	37.0	54.0	42.5	40.0	31.0	38.0	43.0	40.5	44.7
VideoChat2 (Li et al., 2024d)	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
Qwen2-VL (Wang et al., 2024a)	7B	51.0	58.0	77.5	47.0	64.0	63.0	65.5	40.0	25.5	35.5	77.0	43.5	47.0	62.0	42.0	61.5	49.5	41.5	47.5	41.5	52.0
PLLaVA (Xu et al., 2024)	34B	65.0	53.0	83.5	45.0	77.5	70.0	64.5	38.5	37.5	49.0	89.5	41.5	43.5	70.0	53.0	52.5	65.0	39.5	60.5	58.0	57.8
LLaVA-OneVision (Li et al., 2024a)	72B	63.0	58.0	84.5	46.5	85.5	64.0	73.5	41.5	37.0	69.0	95.0	47.5	47.5	75.5	53.5	52.0	70.5	34.0	64.0	54.5	60.8
InternVL2 (Chen et al., 2024e)	8B	75.0	62.0	83.5	40.5	69.5	96.0	72.0	29.5	58.0	53.0	88.5	39.5	83.0	97.0	51.0	78.5	65.0	33.0	48.0	67.0	64.5
IXC2.5-OL	7B	84.5	81.0	75.0	46.0	81.0	92.0	79.5	36.5	83.0	47.0	90.0	60.5	75.0,	93.0	58.0	60.5	74.0	42.0	53.0	62.0	68.7

Table 8: **Results on MMBench-Video.** IXC2.5-OL shows performance close to the open-source SOTA.

Method	Params	Perception Mean	Reasoning Mean	Overall
Closed-source APIs.				
Claude 3.5 Sonnet	-	1.38	1.35	1.38
Gemini 1.0 Pro	-	1.50	1.39	1.48
Gemini 1.5 Pro	-	1.98	1.86	1.94
GPT-4V	-	1.66	1.45	1.69
1.68				
GPT-4o	-	2.19	2.08	2.15
Open-source models.				
MovieLLM (Song et al., 2024b)	7B	0.81	0.97	0.87
LLaVA-OneVision (Li et al., 2024a)	72B	1.03	0.70	0.94
PLLaVA (Xu et al., 2024)	7B	1.02	1.03	1.03
ShareGPT4Video (Chen et al., 2024b)	7B	1.04	1.03	1.05
VideoStreaming (Qian et al., 2024)	7B	1.13	1.09	1.12
LLaVA-NeXT-Video (Zhang et al., 2024g)	7B	1.14	1.13	1.14
VILA1.5 (Lin et al., 2024a)	13B	1.39	1.28	1.36
InternVL2 (Chen et al., 2024e)	8B	1.30	1.16	1.26
Qwen2-VL (Wang et al., 2024a)	7B	1.46	1.35	1.44
IXC2.5-OL	7B	1.49	1.25	1.42

excels among all open-source models, achieving a 2.67% improvement over the previous state-of-the-art model, LLaVA-OneVision, and falling just short of the closed-source API, Gemini 1.5 Pro. This performance solidifies IXC2.5-OL's remarkable prowess in real-time video interaction.

MMBench-Video MMBench-Video is a free-form QA video benchmark consisting of 600 videos and 2000 QA pairs. The duration of each video varies from 30 seconds to 6 minutes. Given the open-ended nature of the answers, the benchmark utilizes GPT-4-based evaluation to enhance quality in terms of accuracy, consistency, and alignment with human judgment. The results are presented in Table 8. IXC2.5-OL demonstrates state-of-the-art performance on perception tasks and comparable performance on overall evaluations.

MVBench MVBench is a video benchmark that emphasizes temporal understanding. It encompasses 20 challenging video tasks that cannot be effectively addressed using a single frame. As shown in Table 7, IXC2.5-OL, despite having a smaller 7B parameter size, has outperformed both the GPT-4 series and the 72B open-source model LLaVA-OneVision, demonstrating its strong capability in understanding video temporal dynamics.

5 Conclusion

We have presented IXC2.5-OL, a real-time streaming model that advances multi-modal text, audio, and visual capabilities with long-term memory. IXC2.5-OL empowers users to engage in dynamic and interactive experiences. Our model's real-time processing enables fluid and responsive interactions, allowing users to engage with ever-changing environments of multimodal data seamlessly, providing a more intuitive and efficient user experience. Our future work will focus on reducing system latency to provide a seamless user experience.

6 Limitations

The limitations of our work stem from the system-level architecture we adopted, which integrates automatic speech recognition, memory extraction and retrieval, large language models, and text-to-speech in a serial interconnected workflow. This sequential processing, where the output of one module serves as the input for the next, inevitably introduces system-level latency. This multi-stage approach may compromise the real-time responsiveness and overall efficiency of the system. Future research should explore the development of end-to-end solutions to mitigate these limitations, thereby enhancing the system's speed and performance while maintaining or improving its functionality.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. 33:1877–1901.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024a. Videollm-online: Online video large language model for streaming video. *Preprint*, arXiv:2406.11816.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, and 1 others. 2024b. ShareGPT4Video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024c. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *Preprint*, arXiv:2411.18211.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2023. Pali: A jointly-scaled multilingual language-image model. *Preprint*, arXiv:2209.06794.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024d. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, and 1 others. 2024e. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Preprint*, arXiv:2404.16821.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Samuele Cornell, Taejin Park, Steve Huang, Christoph Boeddeker, Xuankai Chang, Matthew Maciejewski, Matthew Wiesner, Paola Garcia, and Shinji Watanabe. 2024. The chime-8 dasr challenge for

generalizable and array agnostic distant automatic speech recognition and diarization. *arXiv preprint arXiv:2407.16447*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, and 5 others. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. *Preprint*, arXiv:2403.11481.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*.

E Fonseca, X Favory, J Pons, F Font, and X Serra. 2020. Fsd50k: an open dataset of human-labeled sound events, in arxiv. *arXiv preprint arXiv:2010.00475*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024a. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024b. Vita: Towards open-source interactive omni multimodal llm. *Preprint*, arXiv:2408.05211.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, and 1 others. 2021. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 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*, pages 18995–19012.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language. *Preprint*, arXiv:2312.03700.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *arXiv preprint arXiv:2404.05726*.

Suyuan Huang, Haoxin Zhang, Yan Gao, Yao Hu, and Zengchang Qin. 2024. From image to video, what do we need in multimodal llms? *arXiv preprint arXiv:2404.11865*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. 2024. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. 2024. An image grid can be worth a video: Zero-shot video question answering using a vlm. *Preprint*, arXiv:2403.18406.

Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley. 2018. Audio set classification with attention model: A probabilistic perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320. IEEE.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4. *IJCV*, 128(7):1956–1981.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. 2024b. Aria: An open multimodal native mixture-of-experts model. *Preprint*, arXiv:2410.05993.

Jungang Li, Sicheng Tao, Yibo Yan, Xiaojie Gu, Haodong Xu, Xu Zheng, Yuanhuiyi Lyu, Linfeng Zhang, and Xuming Hu. 2024c. SAVEn-Vid: Synergistic audio-visual integration for enhanced understanding in long video context. *arXiv preprint arXiv:2411.16213*.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2023b. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024d. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, and 8 others. 2024e. Ocean-omni: To understand the world with omni-modality. *Preprint*, arXiv:2410.08565.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2025. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

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

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024a. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

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

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024b. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, and 1 others. 2021. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. *arXiv preprint arXiv:2112.06223*.

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

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

OpenAI. 2023a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023b. Gpt-4v(ision) system card.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *Preprint*, arXiv:2405.16009.

Qwen. 2023. Introducing Qwen-7B: Open foundation and human-aligned models (of the state-of-the-arts).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*.

Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024a. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and 1 others. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. 2024b. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024. video-salmonn: Speech-enhanced audio-visual large language models. *Preprint*, arXiv:2406.15704.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, and 1 others. 2021. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. 2023. V3det: Vast vocabulary visual detection dataset. In *The IEEE International Conference on Computer Vision (ICCV)*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024b. A full-duplex speech dialogue scheme based on large language models. *arXiv preprint arXiv:2405.19487*.

Xiong Wang, Yangze Li, Chaoyou Fu, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024c. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. 2024d. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. *Preprint*, arXiv:2411.17991.

Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. 2024e. Videollamb: Long-context video understanding with recurrent memory bridges. *Preprint*, arXiv:2409.01071.

Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. 2024. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Preprint*, arXiv:2408.16730.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2025. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer.

Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *Preprint*, arXiv:2410.11190.

Li Xu, He Huang, and Jun Liu. 2021. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9878–9888.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *Preprint*, arXiv:2404.16994.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, and 1 others. 2022. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024a. Self-chained image-language model for video localization and question answering. volume 36.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2024b. SALMONN-omni: A codec-free llm for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *Preprint*, arXiv:2402.12226.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-CLIP: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023a. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023b. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

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

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024b. Flash-vstream: Memory-based real-time understanding for long video streams. *Preprint*, arXiv:2406.08085.

Kaiyan Zhang, Biqing Qi, and Bowen Zhou. 2024c. Towards building specialized generalist ai with system 1 and system 2 fusion. *arXiv preprint arXiv:2407.08642*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, and 8 others. 2024d. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, and 1 others. 2023d. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024e. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. 2024f. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Preprint*, arXiv:2406.19389.

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

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024h. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. Melotts: High-quality multi-lingual multi-accent text-to-speech.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. MLVU: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.