# VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction

Chaoyou Fu[1,2,♠], Haojia Lin[4], Xiong Wang[3], Yi-Fan Zhang[5], Yunhang Shen[3]
Xiaoyu Liu[1,2], Haoyu Cao[3], Zuwei Long[3], Heting Gao[3], Ke Li[3], Long Ma[3],
Xiawu Zheng[4], Rongrong Ji[4], Xing Sun[3,†], Caifeng Shan[1,2], Ran He[5]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]School of Intelligence Science and Technology, Nanjing University
[3]Tencent Youtu Lab, [4]XMU, [5]CASIA
♠ Project Leader † Corresponding Author

Demo Video: Click YouTube Link

Source Code: https://github.com/VITA-MLLM/VITA

## Abstract

Recent Multimodal Large Language Models (MLLMs) have typically focused on integrating visual and textual modalities, with less emphasis placed on the role of speech in enhancing interaction. However, speech plays a crucial role in multimodal dialogue systems, and implementing high-performance in both vision and speech tasks remains a challenge due to the fundamental modality differences. In this paper, we propose a carefully designed multi-stage training methodology that progressively trains LLM to understand both visual and speech information, ultimately enabling fluent vision and speech interaction. Our approach not only preserves strong vision-language capacity, but also enables efficient speech-to-speech dialogue capabilities without separate ASR and TTS modules, significantly accelerating multimodal end-to-end response speed. By comparing against state-of-the-art counterparts across benchmarks for image, video, and speech, we demonstrate that our **omni** model is equipped with both strong visual and speech capabilities, making omni understanding and interaction.

## 1 Introduction

Recent advancements in MLLMs [1, 2, 3, 4, 5, 6, 7, 8, 9] have led to significant progress, particularly in integration of visual and textual modalities. The introduction of visual information into LLMs has notably enhanced model capabilities across various multimodal tasks. However, with the growing appeal of human-computer interaction, the role of the speech modality has become increasingly prominent, especially in multimodal dialogue systems. In such a system, speech not only serves as a key medium for information transmission but also greatly improves the naturalness and convenience of interactions. Consequently, integrating visual and speech modalities to achieve multimodal interactions has emerged as a critical research focus.

The integration of vision and speech in MLLMs is not straightforward due to their inherently differences [10]. For example, visual data, such as images, convey spatial information, while speech data convey dynamic changes in time series. These fundamental differences pose challenges for simultaneous optimization of both modalities, often leading to conflicts during training. For

39th Conference on Neural Information Processing Systems (NeurIPS 2025).


Figure 1: VITA-1.5 enables near real-time vision and speech interaction via an end-to-end framework. It allows you to turn on the camera and have a fluent speech conversation. **Please see our demo video at this YouTube link.**

instance, the inclusion of speech data may degrade performance on vision tasks, and vice versa. In addition, traditional speech-to-speech systems rely on separate modules for Automatic Speech Recognition (ASR) and Text-to-Speech, which can increase latency and reduce coherence, limiting their practicality in real-time applications [11, 12, 13, 14, 15].

In this paper, we introduce VITA-1.5, a multimodal LLM that integrates vision, language, and speech through a carefully designed three-stage training methodology. The training strategy progressively incorporates vision and speech data, relieving modality conflicts while maintaining strong multimodal performance. In the first stage, we focus on vision-language by training visual adapters and fine-tuning the model with descriptive caption and visual QA data. This step establishes the model's foundational visual capabilities, enabling robust image and video understanding. The second stage introduces audio input processing by training an audio encoder using speech-transcription paired data, followed by fine-tuning with speech QA data. This stage equips the model with the ability to understand and respond to audio inputs effectively. Finally, in the third stage, we train an audio decoder to enable end-to-end speech output, eliminating the need for external TTS modules. This allows VITA-1.5 to generate fluent speech replies, enhancing the naturalness and interactivity of multimodal dialogue systems.

We have conducted extensive evaluations on various benchmarks related to image, video, and speech understanding, comparing the results with both open-source and proprietary models. VITA-1.5 demonstrates comparable perception and reasoning capabilities comparable to leading image/video based MLLMs, and shows significant improvements in the speech capability.

## 2 Related Work

Recently, thanks to the rapid development of language models such as GPTs [16, 17], LLaMA [18, 19], Alpaca [20], Vicuna [21], and Mistral [22], researchers have successfully extended text comprehension to multimodal understanding/reasoning through techniques like multimodal alignment and instruction tuning. For example, models such as LLaVA [1], Qwen-VL [23], Cambrian-1 [24], Mini-Gemini [25], MiniCPM-V 2.5 [26], DeepSeek-VL [27], and SliME [28] have made significant advances in image perception and reasoning, while models like LongVA [29] and Video-LLaVA [30] have showcased the latest progress in video understanding. These models are increasingly capable of handling diverse data types, driving the continuous improvement of multimodal perception and understanding capabilities.

Beyond visual modalities, recent years have also witnessed significant progress in incorporating speech capabilities into LLMs, driven by the increasing demand for natural human-computer interaction. The dominant approach has been to cascade ASR, LLM, and TTS modules. This text-centric

approach faces fundamental limitations due to the loss of paralinguistic features like tones and emotions. While works like [31] and [32] have attempted to address these issues by incorporating speech encoders and emotion vectors, they still rely on speech transcription, resulting in substantial latency issues that impact the user experience. The emergence of proprietary models like GPT-4o [33] has demonstrated the possibility of end-to-end speech interaction, inspiring a new wave of research in speech-enabled MLLMs. Following this trend, several notable works have emerged in the open-source community. Models such as Mini-Omni2 [34], LLaMA-Omni [35], and Moshi [36] have explored various strategies for aligning speech modality with LLMs and achieving duplex dialogue capabilities. While these open-source efforts have successfully enabled duplex speech interaction with LLMs, they still lack the capability to handle visual modalities as demonstrated by GPT-4o, limiting their applications in scenarios requiring both visual and speech understanding.

Despite these advances in both visual and speech modalities, a significant gap remains between proprietary and open-source models. Compared to proprietary models that support multiple modalities, including audio, image, and text, e.g., GPT-4o [37] and Gemini-Pro 1.5 [38], most open-source models have primarily focused on image and text modalities [2]. Moreover, few open-source models have involved multimodal interaction capabilities, which is a relatively unexplored area. While works like VITA-1.0 [12] have made initial attempts to introduce speech for human-computer interaction, introducing additional speech data poses challenges to the model's original multimodal abilities. Furthermore, speech generation typically relies on existing TTS systems, which often results in high latency, thus impacting user experience. In this paper, we present VITA-1.5 that leverages refined training strategies, excelling in perceiving data across four modalities (video, image, text, and audio), while also realizing near real-time vision and speech interaction.

# 3 VITA-1.5

## 3.1 Model Architecture

The overall architecture of VITA-1.5 is depicted in Fig. 2. The input side is the same as that of the VITA-1.0 version [12], that is, adopting the configuration of "Multimodal Encoder-Adaptor-LLM". It combines the Vision/Audio Transformer and the Multi-Layer Connector with an LLM for joint training, aiming to enhance the unified understanding of vision, language, and audio. With respect to the output side, VITA-1.5 has its own end-to-end speech module, instead of using the external TTS model like the original VITA-1.0 version.

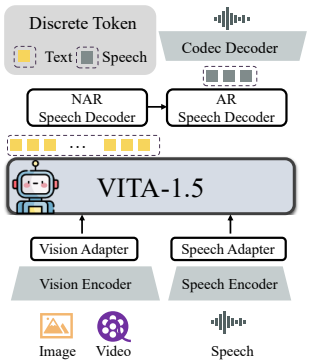


Figure 2: **Overall Architecture of VITA-1.5**. The input side consists of vision and audio encoders, along with their adapters. The output side has an end-to-end speech generation module, rather than directly using an TTS model.

### 3.1.1 Visional Modality

**Visual Encoder.** VITA-1.5 adopts InternViT-300M[1] as the visual encoder, with an input image size of 448×448 pixels, generating 256 visual tokens per image. For high-resolution images, VITA-1.5 employs a dynamic patching [39] strategy to capture local details, improving the accuracy of image understanding.

**Video Processing.** Videos are treated as a special type of multiple-image input. If the video length is shorter than 4 seconds, 4 frames are uniformly sampled; for videos between 4 and 16 seconds, one frame per second is sampled; for videos longer than 16 seconds, 16 frames are uniformly sampled. No dynamic patching is applied to video frames to avoid excessive visual tokens that could hinder processing efficiency.

**Vision Adapter.** A two-layer MLP is used to map the visual features to visual tokens suitable for the subsequent understanding of LLM.

[1] https://huggingface.co/OpenGVLab/InternViT-300M-448px

### 3.1.2 Audio Modality

**Speech Encoder.** Similar to [40], our audio encoding module consists of multiple downsampling convolutional layers (4x downsampling) and 24 Transformer blocks (with a hidden size of 1024). The downsampling layers help reduce the frame rate of the audio features, improving the processing speed of LLM. The audio encoder has about 350M parameters and an output frame rate of 12.5Hz. Mel-filter bank features are used as the input of the audio encoder, with a window size of 25ms and a shift of 10ms [40].

**Speech Adapter.** It consists of multiple convolutional layers with 2x downsampling.

**Speech Decoder.** TiCodec [41] is used as our codec model, customizing a single codebook with a size of 1024. This single-codebook design simplifies the decoding process during the inference phase. The codec model is responsible for encoding continuous speech signals into discrete speech tokens with the frequency of 40Hz, and at the same time has the ability to decode them back into speech signals with the sample rate of 24,000Hz.

The current LLM can only output text tokens, and the speech generation capability requires the LLM to be able to output speech tokens. To this end, we add two speech decoders after the text tokens following [40]: 1) **Non-Autoregressive (NAR) Speech Decoder**, which processes text tokens globally and models semantic features, with the aim of generating an initial distribution of speech tokens; 2) **Autoregressive (AR) Speech Decoder** generates higher quality speech tokens step by step, based on the speech information produced by the NAR decoder. The final sequence of speech tokens is then decoded into a continuous speech signal flow (waveform) using the speech decoder of the Codec model. We adopt 4 LLaMA decoder layers for both NAR and AR speech decoders, where the hidden size is 896 and the parameter size is about 120M.

## 3.2 Training Data

As shown in Table 1, the training data of multimodal instruction tuning encompass a wide range of categories, such as caption data and QA data, both Chinese and English. During different training phases, subsets of the overall dataset are selectively sampled to serve different objectives. Specifically, the datasets are categorized as follows:

- **Image Captioning Data.** Datasets such as ShareGPT4V [42], ALLaVA-Caption [43], SharedGPT4o-Image[2], and synthetic data are used to train the model to generate descriptive languages for images.
- **Image QA Data.** Datasets like LLaVA-150K[3], LLaVA-Mixture-sample [1], LVIS-Instruct [44], ScienceQA [45], ChatQA [46], and subsets sampled from LLaVA-OV [47], such as general image QA and mathematical reasoning datasets, are utilized to train the model in answering image-based questions and performing visual reasoning tasks.
- **OCR & Diagram Data.** This category supports the model in understanding OCR and diagram content, using datasets such as Anyword-3M [48], ICDAR2019-LSVT[4], UReader [49], SynDOG[5], ICDAR2019-LSVT-QA[6], and corresponding data sampled from LLaVA-OV.
- **Video Data.** Datasets like ShareGemini [50] and synthetic data are used to train the model to handle video inputs and perform tasks such as captioning and video-based QA.
- **Pure Text Data.** This category enhances the model's capability to understand and generate languages, facilitating text-based QA tasks.

In addition to the image and video data listed in Table 1, 110,000 hours of internal speech-transcription paired ASR data, covering both Chinese and English, are incorporated to train the audio encoder and align the audio encoder with the LLM. Furthermore, 3,000 hours of text-speech paired data generated by a TTS system are used to train the speech decoder.

[2] https://sharegpt4o.github.io/
[3] https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K
[4] http://icdar2019.org/
[5] naver-clova-ix/synthdog-en
[6] http://icdar2019.org/

Table 1: Training data of multimodal instruction tuning. The images of the synthetic data come from open-source datasets like Wukong [51], LAION [52], and CC12M [53].

| Data Scenario | QA Type | Dataset Name | Questions (K) | Language |
|---|---|---|---|---|
| General Image | Description | ShareGPT4V | 99.50 | Eng |
| | | ALLaVA-Caption | 697.40 | Eng |
| | | ShareGTP4o-Image | 55.50 | Eng |
| | | Synthetic Data | 593.70 | CN |
| | QA | LLaVA-150K | 218.36 | CN |
| | | LLaVA-Mixture-sample | 1872.10 | Eng |
| | | LVIS-Instruct | 939.36 | Eng |
| | | ScienceQA | 12.72 | Eng |
| | | ChatQA | 7.39 | Eng |
| | | LLaVA-OV General | 1754.65 | Eng |
| | | LLaVA-OV Math Reasoning | 1140.92 | Eng |
| | | Synthetic Data | 212.68 | CN |
| OCR & Diagram | Description | Anyword-3M | 1709.30 | CN |
| | | ICDAR2019-LSVT | 366.30 | CN |
| | | UReader | 100.00 | Eng |
| | | SynDOG-EN | 100.00 | Eng |
| | | SynDOG-CN | 101.90 | CN |
| | QA | ICDAR2019-LSVT-QA | 630.08 | CN |
| | | LLaVA-OV Doc Chart Screen | 4431.50 | Eng |
| | | LLaVA-OV General OCR | 404.20 | Eng |
| General Video | Description | ShareGemini | 205.70 | CN |
| | | Synthetic Data | 569.40 | CN & Eng |
| | QA | Synthetic Data | 4336.30 | CN & Eng |
| Pure Text | QA | Synthetic Data | 1574.20 | CN & Eng |
| Total | | | 22133.16 | CN & Eng |

### 3.3 Three Stage Training Strategies

In order to ensure that VITA-1.5 performs well in tasks involving vision, language, and audio, we have to face a key challenge, i.e., training conflicts between different modalities. For example, adding the speech data could negatively impact the understanding of the vision data, as the features of speech differ significantly from those of vision, causing interference during the learning process. To address this challenge, we devise a three-stage training strategy as shown in Fig. 3. The core idea is to gradually introduce different modalities into the model, allowing it to increase the power of a new modality while maintaining the power of the existing modalities.

#### 3.3.1 Stage 1: Vision-Language Training

**Stage 1.1 Vision Alignment.** In this stage, our goal is to bridge the gap between vision and language. The features of the former are extracted from the pre-trained vision encoder InternViT-300M, and the latter is introduced through the LLM. We use 20% of the descriptive caption data from Table 1 for training, where only the visual adapter is trainable, while the other modules are frozen. This approach allows the LLM to initially align the visual modality.

**Stage 1.2 Vision Understanding.** In this stage, our goal is to teach the LLM to transcribe image content. Toward this end, we use all the descriptive caption data from Table 1. During this process, the encoder and adapter of the visual module, as well as the LLM, are trainable. The focus is to enable the model to establish a strong connection between vision and language by learning from descriptive texts about images, allowing it to understand image content via generating natural language descriptions.

**Stage 1.3 Vision SFT.** Following Stage 1.2, the model has acquired a basic understanding of images and videos. However, the instruction following ability is still limited, and it is difficult to cope with the visual QA task. To achieve this, we use all the QA data from Table 1 while retaining 20% of the descriptive caption data to increase the diversity of the dataset and the complexity of the tasks.

During training, the encoder and adapter of the visual module, as well as the LLM, are trainable. The key objective of this stage is to enable the model not only to understand visual content but also to answer questions following instructions.

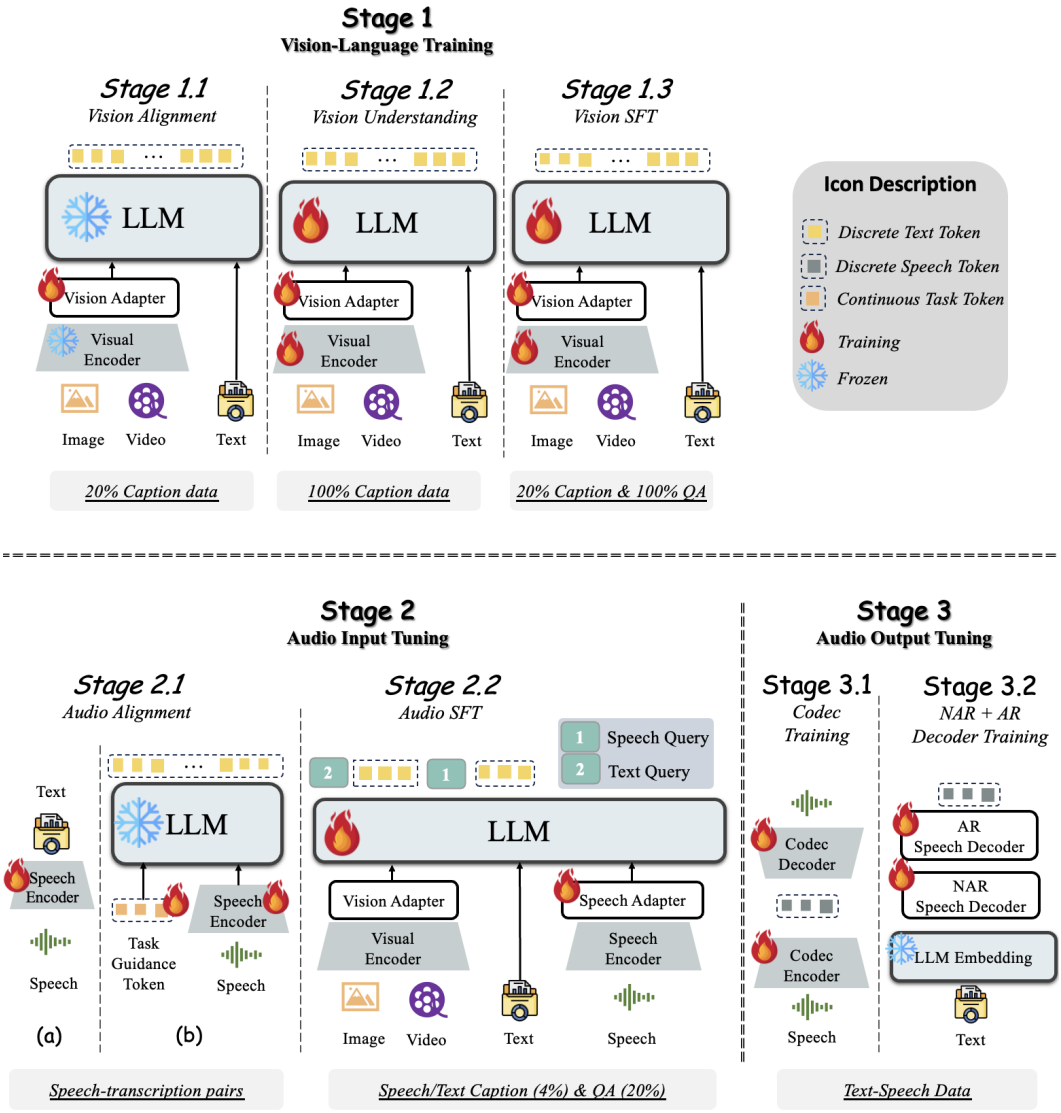


Figure 3: **Training Pipeline of VITA-1.5**. The training process is divided into three stages to incrementally incorporate vision and audio into the LLM while relieving modality conflicts. Stage I focuses on **Vision-Language Training**, including vision alignment (Stage 1.1, using 20% caption data from Table 1), vision understanding (Stage 1.2, using 100% caption data), and instruction tuning for visual QA (Stage 1.3, using 20% caption data and 100% QA data). Stage 2 introduces **Audio Input Tuning**, with audio alignment (Stage 2.1, utilizing 11,000 hours of speech-transcription pairs) and instruction tuning for speech QA (Stage 2.2, sampling 4% caption data and 20% QA data). Finally, Stage 3 focuses on **Audio Output Tuning**, including the training of the codec model (Stage 3.1, using 3,000 hours of text-speech data) and speech decoder training (Stage 3.2). The percentages shown in the image correspond to the data sampling ratios specified in Table 1.

#### 3.3.2 Stage 2: Audio Input Tuning

**Stage 2.1 Audio Alignment.** After completing the training of Stage 1, the model has developed a strong foundation in image and video understanding. In this stage, our goal is to reduce the discrepancy between audio and language based on Stage 1, enabling the LLM to understand audio inputs. The training data consists of 11,000 hours of speech-transcription pairs. We follow a two-step approach: (a) *Speech Encoder Training*: We adopt a training framework used in common speech recognition systems, using a Connectionist Temporal Classification (CTC) loss function [54] to train the speech encoder. The aim is for the encoder to predict the transcription text from the speech input. This step ensures that the audio encoder can extract speech features and map them to the text representation space. (b) *Speech Adapter Training*: After training the speech encoder, we integrate it with the LLM, using an audio adapter to introduce audio features into the input layer of the LLM. The training objective at this stage is to enable the LLM to output the transcription text of the speech data.

Besides, in step (b), we introduce special trainable input tokens to guide the speech understanding process. These tokens provide additional contextual information that guides the LLM used for the QA task to perform the ASR task.

**Stage 2.2 Audio SFT.** The focus of this stage is to introduce the QA functionality with speech questions and text answers. To achieve this, we sample 4% of the caption data and 20% of the QA data from Table 1. In terms of data processing, approximately half of the text-based questions are randomly replaced with their corresponding speech versions, generated using a TTS system.

In this stage, both the visual encoder and adapter, the audio encoder and adapter, as well as the LLM are trainable, aiming to improve the model's adaptability with multimodal inputs. In addition, we add a classification head to the LLM's output. This head is used to distinguish whether the input comes from speech or text. As a result, the model can more accurately interpret speech inputs and process different modalities efficiently and flexibly.

#### 3.3.3 Stage 3: Audio Output Tuning

In the first two stages of training, the VITA-1.5 model has effectively developed its multimodal understanding capabilities. However, a crucial capacity, i.e., speech output, remains absent, which is essential for its role as an interactive assistant. To introduce speech output functionality without compromising the model's fundamental abilities, we draw on the strategy [40], using 3,000 hours of text-speech data and employing a two-step training approach (see Fig. 3).

**Stage 3.1 Codec Training.** The goal of this step is to train a codec model with a single codebook using speech data. The encoder of the codec model has the ability to map speech to discrete tokens, while the decoder can map the discrete tokens back to speech stream. During the inference phase of VITA-1.5, only the decoder is used.

**Stage 3.2 NAR + AR Decoder Training.** The training of this stage uses text-speech paired data, where the text is fed into the tokenizer and the embedding later of the LLM to obtain its embedding vectors, and the speech is fed into the encoder of the codec model to obtain its speech tokens. The text embedding vectors are sent to the NAR speech decoder to get global semantic features, and then the features are sent to the AR speech decoder, which predicts the corresponding speech tokens. Note that the LLM is frozen during this stage, thus the multimodal performance is not affected.

## 4 Evaluation

### 4.1 Vision-Language Evaluation

**Baselines.** We compare a series of open-source MLLMs, including VILA-1.5 [55], LLaVA-Next [56], CogVLM2 [57], InternLM-XComposer2.5 [58], Cambrian-1 [24], MiniCPM-V-2.6 [26], Ovis1.5 [59], InternVL-Chat-1.5, InternVL-2 [60], LLaVA-OV [47], and Video-LLaVA [30], SliME [28], and LongVA [29], as well as 5 closed-source MLLMs, including GPT-4V[7], GPT-4o[8], GPT-4o-mini, Gemini 1.5 Pro [38], and Claude 3.5 Sonnet[9].

[7] https://openai.com/index/gpt-4v-system-card/

[8] https://openai.com/index/hello-gpt-4o/

[9] https://www.anthropic.com/news/claude-3-5-sonnet

Table 2: **Evaluation on Image Understanding Benchmarks.** VITA-1.5 shows performance comparable to the leading open-source models and advanced closed-source counterparts. MMB refers to MMBench, MMS to MMStar, Hal to HallusionBench, MathV to MathVista, and OCR to OCRBench. Note that after the training of Stages 2 (Audio Input Tuning) and 3 (Audio Output Tuning), VITA-1.5 retains almost its original visual-language capabilities in Stage 1 (Vision-Language Training).

| Method | LLM | MMB | MMS | MMMU | MathV | Hal | AI2D | OCR | MMVet | MME | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VILA-1.5 | Vicuna-v1.5-13B | 68.5 | 44.2 | 41.1 | 42.5 | 39.3 | 69.9 | 460.0 | 45.0 | 1718.2 | 52.1 |
| LLaVA-Next | Yi-34b | 77.8 | 51.6 | 48.8 | 40.4 | 34.8 | 78.9 | 574.0 | 50.7 | 2006.5 | 58.3 |
| CogVLM2 | Llama3-8B-Instruct | 70.7 | 50.5 | 42.6 | 38.6 | 41.3 | 73.4 | 757.0 | 57.8 | 1869.5 | 58.8 |
| InternLM-Xcomposer2 | InternLM2-7B | 77.6 | 56.2 | 41.4 | 59.5 | 41.0 | 81.2 | 532.0 | 46.7 | 2220.4 | 61.2 |
| Cambrian | Nous-Hermes-2-Yi-34B | 77.8 | 54.2 | 50.4 | 50.3 | 41.6 | 79.5 | 591.0 | 53.2 | 2049.9 | 61.4 |
| InternVL-Chat-1.5 | InternLM2-20B | 79.7 | 57.1 | 46.8 | 54.7 | 47.4 | 80.6 | 720.0 | 55.4 | 2189.6 | 65.1 |
| Ovis1.5 | Gemma2-9B-It | 77.3 | 58.1 | 49.7 | 65.6 | 48.2 | 84.5 | 752.0 | 53.8 | 2125.2 | 66.9 |
| InternVL2 | InternLM2.5-7b | 79.4 | 61.5 | 51.2 | 58.3 | 45.0 | 83.6 | 794.0 | 54.3 | 2215.1 | 67.3 |
| MiniCPM-V 2.6 | Qwen2-7B | 78.0 | 57.5 | 49.8 | 60.6 | 48.1 | 82.1 | 852.0 | 60.0 | 2268.7 | 68.5 |
| Proprietary | | | | | | | | | | | |
| GPT-4V | - | 65.5 | 50.4 | 59.3 | 48.2 | 39.3 | 71.4 | 678.0 | 49.0 | 1790.3 | 58.5 |
| GPT-4o mini | - | 76.0 | 54.8 | 60.0 | 52.4 | 46.1 | 77.8 | 785.0 | 66.9 | 2003.4 | 66.3 |
| Gemini 1.5 Pro | - | 73.9 | 59.1 | 60.6 | 57.7 | 45.6 | 79.1 | 754.0 | 64.0 | 2110.6 | 67.2 |
| GPT-4o | - | 82.8 | 61.6 | 62.8 | 56.5 | 51.7 | 77.4 | 663.0 | 66.5 | 2328.7 | 69.3 |
| Claude3.5 Sonnet | - | 78.5 | 62.2 | 65.9 | 61.6 | 49.9 | 80.2 | 788.0 | 66.0 | 1920.0 | 69.3 |
| Open Source | | | | | | | | | | | |
| VITA-1.0 | Mixtral-8x7B | 71.8 | 46.4 | 47.3 | 44.9 | 39.7 | 73.1 | 678.0 | 41.6 | 2097.0 | 57.8 |
| VITA-1.5 (Stage 1) | Qwen2-7B | 77.1 | 59.1 | 53.1 | 66.2 | 44.1 | 80.3 | 752.0 | 51.1 | 2311.0 | 67.1 |
| VITA-1.5-Audio (Stage 3) | Qwen2-7B | 76.7 | 59.9 | 52.1 | 66.2 | 44.9 | 79.3 | 732.0 | 49.6 | 2352.0 | 66.8 |

Table 3: **Evaluation on Video Understanding Benchmarks.** Although VITA-1.5 still lags behind models like GPT-4o and Gemini-1.5-Pro, it performs comparably to many open-source models. Note that after the training of Stages 2 (Audio Input Tuning) and 3 (Audio Output Tuning), VITA-1.5 retains almost its original visual-language capabilities in Stage 1 (Vision-Language Training).

| Method | LLM | Video-MME w/o sub | Video-MME w/ sub | MVBench | TempCompass |
|---|---|---|---|---|---|
| Video-LLaVA | Vicuna-v1.5-13B | 39.9 | 41.6 | | 49.8 |
| SliME | Llama3-8B-Instruct | 45.3 | 47.2 | - | - |
| LongVA | Qwen2-7B | 52.6 | 54.3 | - | 57.0 |
| VILA-1.5 | Llama3-8B-Instruct | - | - | - | 58.8 |
| InternLM-XComposer-2.5 | InternLM2-7B | - | - | - | 62.1 |
| LLaVA-OneVision | Qwen2-7B | 58.2 | 61.5 | 56.7 | 64.2 |
| InternVL-2 | InternLM2.5-7b | - | - | - | 66.0 |
| MiniCPM-V-2.6 | Qwen2-7B | 60.9 | 63.7 | - | 66.3 |
| Proprietary | | | | | |
| GPT-4o-mini | - | 64.8 | 68.9 | - | |
| Gemini-1.5-Pro | - | 75.0 | 81.3 | - | 67.1 |
| GPT-4o | - | 71.9 | 77.2 | - | 73.8 |
| Open Source | | | | | |
| VITA-1.0 | Mixtral-8x7B | 55.8 | 59.2 | - | 62.3 |
| VITA-1.5 (Stage 1) | Qwen2-7B | 56.8 | 59.5 | 56.8 | 65.5 |
| VITA-1.5 (Stage 3) | Qwen2-7B | 56.1 | 58.7 | 55.4 | 66.7 |

**Evaluation Benchmarks.** To assess the image perception and understanding capabilities of VITA-1.5, we utilize several evaluation benchmarks, including MME [61], MMBench [62], MMStar [63], MMMU [64], MathVista [65], HallusionBench [66], AI2D [67], OCRBench [68], and MMVet [69]. These benchmarks cover a wide range of aspects, including general multimodal capabilities (e.g., MME, MMBench, and MMMU), mathematical reasoning (MathVista), hallucination detection (HallusionBench), chart (AI2D) and OCR (OCRBench) understanding, providing a comprehensive evaluation results. For video understanding, we use representative evaluation benchmarks including Video-MME [70], MVBench [71], and TempCompass [72].

**Vision-Language Capabilities.** Table 2 presents a comparison of VITA-1.5's image understanding performance. After the training of the three stages, VITA-1.5 performs comparably to the most advanced open-source models and even surpasses some closed-source models like GPT-4V and GPT-4o-mini. This result highlights the robust capabilities of VITA-1.5 in image-language tasks. As shown in Table 3, VITA-1.5 shows comparable performance to the top open-source models in the evaluation of video understanding. The notable gap compared to proprietary models suggests that VITA-1.5 still has significant room for improvement and potential for further enhancement in video understanding. Please note that after the training of Stages 2 (Audio Input Tuning) and 3 (Audio Output Tuning), VITA-1.5 retains almost its original visual-language capabilities in Stage 1 (Vision-Language Training).

Table 4: **Evaluation on ASR Benchmarks.** VITA-1.5 has demonstrated strong performance in both Mandarin and English ASR tasks. It outperforms specialized speech models, achieving better results in both languages.

| Model | CN (CER↓) | | | Eng (WER↓) | | | |
|---|---|---|---|---|---|---|---|
| | **aishell-1** | **test net** | **test meeting** | **dev clean** | **dev other** | **test clean** | **test other** |
| Wav2vec2-base | - | - | - | 6.0 | 13.4 | - | - |
| Mini-Omni2 | - | - | - | 4.8 | 9.8 | 4.7 | 9.4 |
| Freeze-Omni | 2.8 | 12.6 | 14.2 | 4.2 | 10.2 | 4.1 | 10.5 |
| VITA-1.0 | - | 12.2 | 16.5 | 7.6 | 16.6 | 8.1 | 18.4 |
| VITA-1.5 | **2.2** | **8.4** | **10.0** | **3.3** | **7.2** | **3.4** | **7.5** |

### 4.2 Speech Evaluation

**Baselines.** The following three baseline models are used for comparison: Wav2vec2-base [73], Mini-Omni2 [74], Freeze-Omni [40], and VITA-1.0 [12].

**Evaluation Benchmarks.** *The Mandarin Evaluation Sets* consists of three datasets: aishell-1 [75], test net [76], and test meeting [77]. These datasets are used to evaluate the model's performance on Mandarin speech. The evaluation metric is the Character Error Rate (CER). *The English Evaluation Sets* include four datasets: dev-clean, dev-other, test-clean, and test-other [78], which are used to evaluate the model's performance on English speech. The evaluation metric is Word Error Rate (WER). The evaluation results in Table 4 indicate that VITA-1.5 achieves leading accuracy in both Mandarin and English ASR tasks. This demonstrates that VITA-1.5 has successfully integrated advanced speech capability to support multimodal interaction.

## 5 Conclusion and Future Work

In this paper, we has presented VITA-1.5, a multimodal LLM designed to integrate vision and speech through a carefully crafted three stage training strategy. By relieving the inherent conflicts between modalities, VITA-1.5 achieves robust capabilities in both vision and speech understanding, enabling efficient speech-to-speech interactions without relying on separate ASR or TTS modules. Extensive evaluations demonstrate that VITA-1.5 performs competitively across multimodal benchmarks. We hope that VITA-1.5 can promote the progress of open-source models in the field of real-time multimodal interaction. Although VITA-1.5 has made some contributions, such as multi-modality joint training, end-to-end architecture, response latency, and basic performance, there are two major areas that can be improved in our future work:

1. Personalized MLLM. Currently, VITA-1.5 is generic and do not incorporate individual preferences during interaction. For example, after learning about personal preferences in the interaction, the content and manner of answers can be adjusted accordingly.

2. Long-term memory. The process of human-computer interaction can last 10 minutes or even several hours, in which case it is important for the human-computer interaction in real scenarios.

## Impact Statement

This paper studies the technology of large models to enhance their technical level. Its influence is the same as that of the research on other large models and will not be repeated here.


## Acknowledgments

This work is funded by National Natural Science Foundation of China (Grant No. 62506158 and No. 62441234), Fundamental Research Funds for the Central Universities, AI & AI for Science Project of Nanjing University (No. 2024300529), and CCF-Tencent Rhino-Bird Open Research Fund.

## References


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

[2] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

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

[4] Jinsong Zhang, Minjie Zhu, Yuxiang Zhang, Zerong Zheng, Yebin Liu, and Kun Li. Speechact: Towards generating whole-body motion from speech. *IEEE TVCG*, 2025.

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

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

[7] Hao Wen, Hongbo Kang, Jian Ma, Jing Huang, Yuanwang Yang, Haozhe Lin, Yu-Kun Lai, and Kun Li. Dycrowd: Towards dynamic crowd reconstruction from a large-scene video. *IEEE TPAMI*, 2025.

[8] Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *AAAI*, 2025.

[9] Xiaoguang Tu, Zhi He, Yi Huang, Zhi-Hao Zhang, Ming Yang, and Jian Zhao. An overview of large ai models and their applications. *Visual Intelligence*, 2024.

[10] Dan Oneață and Horia Cucu. Improving multimodal speech recognition by data augmentation and speech representations. In *CVPR*, 2022.

[11] V Madhusudhana Reddy, T Vaishnavi, and K Pavan Kumar. Speech-to-text and text-to-speech recognition using deep learning. In *ICECAA*. IEEE, 2023.

[12] 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.

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

[14] Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang. Geometry-guided dense perspective network for speech-driven facial animation. *IEEE TVCG*, 2021.

[15] Jinsong Zhang, Xiongzheng Li, Hailong Jia, Jin Li, Zhuo Su, Guidong Wang, and Kun Li. Logavatar: Local gaussian splatting for human avatar modeling from monocular video. *CAD*, 2025.

[16] OpenAI. Gpt-4 technical report. 2023.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[20] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[21] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[22] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[23] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[24] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

[25] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[26] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[27] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[28] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.

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

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

[31] Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Qian Chen, and Lei Xie. E-chat: Emotion-sensitive spoken dialogue system with large language models. *arXiv preprint arXiv:2401.00475*, 2023.

[32] Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. Paralinguistics-enhanced large language modeling of spoken dialogue. *arXiv preprint arXiv:2312.15316*, 2023.

[33] OpenAI. https://openai.com/index/hello-gpt-4o/, 2024.

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

[35] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

[36] 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.

[37] OpenAI. Hello gpt-4o. 2023.

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[39] 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.

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

[41] Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chu Yuan Zhang, and Junzuo Zhou. Fewer-token neural speech codec with time-invariant codes. In *ICASSP*. IEEE, 2024.

[42] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[43] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024.

[44] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

[45] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

[46] Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*, 2024.

[47] 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.

[48] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023.

[49] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

[50] Share. Sharegemini: Scaling up video caption data for multimodal large language models, June 2024. https://github.com/Share14/ShareGemini.

[51] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *NeurIPS*, 2022.

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

[53] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[54] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[55] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

[56] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[57] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

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

[59] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

[60] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

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

[62] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[63] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[64] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[65] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[66] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

[67] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 2021.

[68] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[69] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[70] Chaoyou Fu, Yuhan Dai, Yondong 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.

[71] 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.

[72] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[73] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020.

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

[75] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *O-COCOSDA*. IEEE, 2017.

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

[77] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP*, 2022.

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

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theorem and lemma.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is open-source, and the data processing methods have been clearly defined. The experimental settings are also included, making the experiments reproducible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[NA]

   Justification: The evaluations are performed with temperature=0, resulting in minimal variance. Due to the high computational cost of training from scratch, it is difficult to provide results from multiple reruns (5-10).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This work meets the requirements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work is not applicable to this.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The corresponding works are cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work is not applicable to this.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work is not applicable to this.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work is not applicable to this.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This work is not applicable to this.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.