

Lyra: An Efficient and Speech-Centric Framework for Omni-Cognition

Zhisheng Zhong^{1*} Chengyao Wang^{1*} Yuqi Liu^{1*} Senqiao Yang¹ Longxiang Tang² Yuechen Zhang¹
Jingyao Li¹ Tianyuan Qu¹ Yanwei Li¹ Yukang Chen¹ Shaozuo Yu¹ Sitong Wu¹ Eric Lo¹ Shu Liu³✉ Jiaya Jia^{2,3}
¹CUHK ²HKUST ³SmartMore

*Equal contribution ✉ Corresponding author Code: <https://github.com/dvlab-research/Lyra>

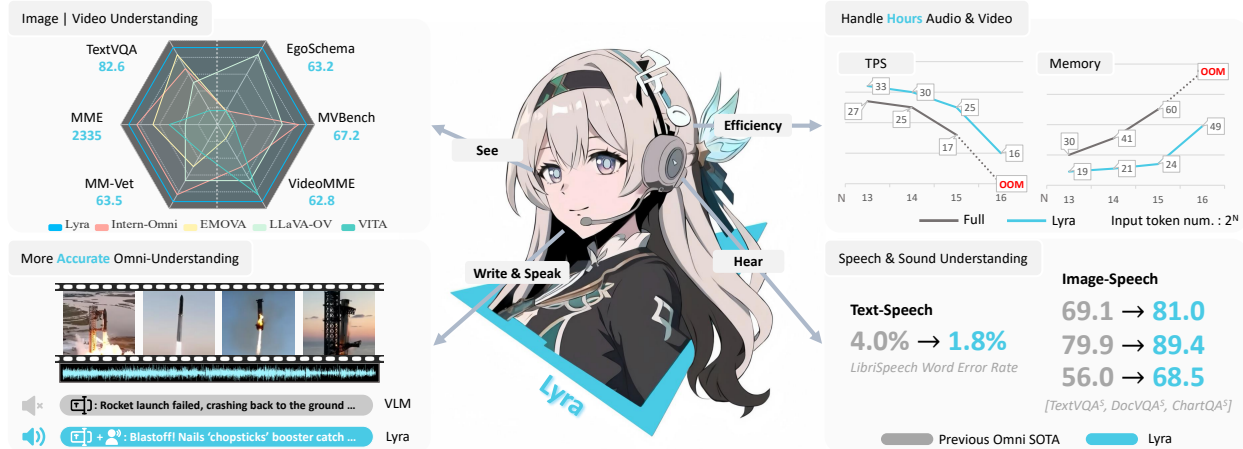


Figure 1. **Overview of Lyra.** Lyra shows superiority compared with leading models in the following aspects: 1. *Stronger performance.* Lyra achieves state-of-the-art results across a variety of modalities understanding and reasoning tasks. 2. *More versatile.* Lyra can directly handle images, videos, and audio tasks even lasting several hours. 3. *More efficient.* Lyra is trained with less data and increases the speed, reduces memory usage, making it suitable for latency-sensitive and long-context multi-modality applications.

Abstract

As *Multi-modal Large Language Models (MLLMs)* evolve, expanding beyond single-domain capabilities is essential to meet the demands for more versatile and efficient AI. However, previous *omni-models* have insufficiently explored speech, neglecting its integration with multi-modality. We introduce *Lyra*, an efficient MLLM that enhances multi-modal abilities, including advanced long speech comprehension, sound understanding, cross-modality efficiency, and seamless speech interaction. To achieve efficiency and speech-centric capabilities, *Lyra* employs three strategies: (1) leveraging existing open-source large models and a proposed multi-modality *LoRA* to reduce training costs and data requirements; (2) using a *latent multi-modality regularizer* and *extractor* to strengthen the relationship between speech and other modalities, thereby enhancing model performance; and (3) constructing a high-quality, extensive dataset that includes 1.5M multi-modal (language, vision, audio) data samples and 12K long speech samples, enabling *Lyra* to handle complex long speech inputs and achieve more robust *omni-cognition*. Compared to other *omni-methods*, *Lyra* achieves state-of-the-art performance on various vision-language, vision-speech, and speech-language benchmarks, while also using fewer computational resources and less training data.

1. Introduction

With the rapid evolution in Large Language Models (LLMs) [23, 28, 45, 59, 61], empowering the impressive capabilities for multi-modality inputs is becoming an essential part of current Multimodal Large Language Models (MLLMs). However, most current MLLMs are limited to just two modalities: either vision-language [2, 12, 27, 31, 32, 34, 37, 84] or speech-language [11, 16, 69]. OpenAI’s recent release of GPT-4o [46], an advanced omni-modal model, has reignited interest in intelligent assistants capable of fine-grained visual perception, understanding spoken instructions, and generating vocal responses simultaneously. It highlights a strong demand for MLLMs that integrate more functions and modalities, such as visual, language, speech, sound, and even other new abilities [7, 19, 67, 77].

Based on our study, most existing omni-models [7, 16, 19, 77] primarily focus on the relationship between speech and text, without exploring connections between speech and other modalities, such as vision. Consequently, speech-related evaluation metrics are typically limited to text. In this paper (Sec. 4.3), we observe that strong performance in the speech-text modality does not necessarily imply good performance in the speech-vision modality. Thus, we suggest that omni-model evaluation should be speech-centric, expanding its involvement with additional modalities.

To further enhance the speech capabilities of MLLMs, we inevitably encounter the following challenges: First, larger datasets (*e.g.*, the extensive data required to train models like LLaMA3 [15] and Qwen2-VL [64]) are needed for both previous modalities and speech. Second, there is a clear trend toward increasing context length across modalities. More long-context benchmarks for specific modalities are being proposed, including long-document [5, 9] and long-video tasks [6, 18, 35, 66, 70, 80]. Last, building a sufficiently powerful model may demand significant financial and computational resources, which raises environmental concerns related to high carbon emissions.

Combining the above three points, we propose Lyra, an efficient and speech-centric framework for omni-cognition:

Leveraging existing open-source large models. We efficiently start with powerful LLMs and VLMs, like LLaMA3 [15] and Qwen2-VL [64], which already demonstrate strong multi-modal capabilities. Through our multi-modality LoRA module, we can effectively preserve certain strong capabilities of open-source large models in specific modalities with minimal training data, while simultaneously developing their abilities in the speech modality.

Enhancing information interaction between modalities, especially within the speech modality. 1) Considering the implicit correspondence between speech and text, we propose latent cross-modality regularizer. 2) Based on instructions, we identify potential redundancy in context token information across multiple modalities. We further propose latent multi-modality extractor to mine informative tokens, which brings significant advantages in training speed, inference speed and GPU memory efficiency.

High-Quality Datasets for Omni-Cognition. Centered on speech, we have constructed two types of high-quality datasets: To enhance the model’s speech capabilities, we collect and generate a multi-modal dataset of 1.5M text-image-speech samples from diverse public sources, ensuring a rich and varied data foundation; To handle longer speech inputs and demands, we are the first to construct a long speech dataset comprising 12K samples. Through training, our model achieves robust omni-cognitive abilities and can handle long speech inputs lasting several hours.

With these three improvements, Lyra offers the following advantages (Fig. 1). **More versatile:** As shown in Table 1, Lyra now supports both sound and speech understanding and generation, while also handling more complex long speech cases. **More efficient:** Lyra achieves faster training and inference speed across speech, image, and video tasks. Compared to previous models, Lyra has a smaller model size and is trained with less data. **Stronger:** Lyra demonstrates enhanced omni-comprehension capabilities over previous MLLMs, achieving state-of-the-art performance in vision-language and vision-speech and speech-language tasks simultaneously.

Function	Method	Vision		Audio			
		Image	Video	SU	SG	LS	Sound
Vision	LLaVA-OV	✓	✓	✗	✗	✗	✗
	Intern-VL	✓	✓	✗	✗	✗	✗
	Mini-Gemini	✓	✓	✗	✗	✗	✗
Audio	Qwen-Audio	✗	✗	✓	✗	✗	✓
	Mini-Omni	✗	✗	✓	✓	✗	✗
	LLaMA-Omni	✗	✗	✓	✓	✗	✗
Omni	Intern-Omni	✓	✗	✓	✗	✗	✗
	VITA	✓	✓	✓	✗	✗	✗
	Any-GPT	✓	✓	✓	✓	✗	✗
	EMOVA	✓	✗	✓	✓	✗	✗
	Lyra	✓	✓	✓	✓	✓	✓

Table 1. **Function comparison of related work.** SU, SG, and LS represent speech understanding, speech generation, and long speech support, respectively.

2. Related Work

Multi-modal Large Language Models. Recent advancements in Large Language Models (LLM) and Multi-modal Large Language Models (MLLMs) have pushed the boundaries of human-computer interaction, expanding their capabilities from text-based tasks to complex multi-modality scenarios. Large Language Models, like GPTs [45], LLaMA [15, 61] and Qwen [4, 71], have demonstrated strong capabilities in textual understanding and generation. Building on these foundations, Vision Language Models [31, 34–39, 64, 65, 74] extend LLMs with visual perception capabilities, leveraging advanced encoders [50] and high-resolution techniques to interpret visual inputs. Speech Language Models (SLMs) [52], including SpeechGPT [78] and LLaMA-Omni [16], have introduced real-time speech understanding and generation, with advanced models enabling control over speech styles. Moving further, MLLMs [67] such as AnyGPT [77], VITA [19] and EMOVA [7], integrate vision, text, and audio within a unified architecture, enabling robust interaction across diverse modalities. The abilities and modalities of previous leading MLLMs are listed in Table 1. In contrast, Lyra tackles complex scenarios, enabling seamless, dynamic multi-modal interactions for rich, real-time AI experiences.

Token Reduction for MLLMs. Token reduction techniques aim to improve the efficiency of LLMs and VLMs by minimizing redundant tokens during inference and training. In LLMs, methods like StreamingLLM [68] and FastGen [20] optimize memory usage by selectively retaining essential tokens, while techniques like H₂O [83], ScissorHands [40] and Quest [57] use attention-based scoring to prioritize valuable tokens. In VLMs, approaches such as FastV [8] and VisionZip [73] reduce visual tokens to tackle the high computational cost of image processing. Lyra generalizes token reduction to additional modalities such as video and speech, where token lengths grow with longer contexts. By analyzing the interplay between context and instruction tokens, it incrementally removes redundancy to boost efficiency without sacrificing performance.

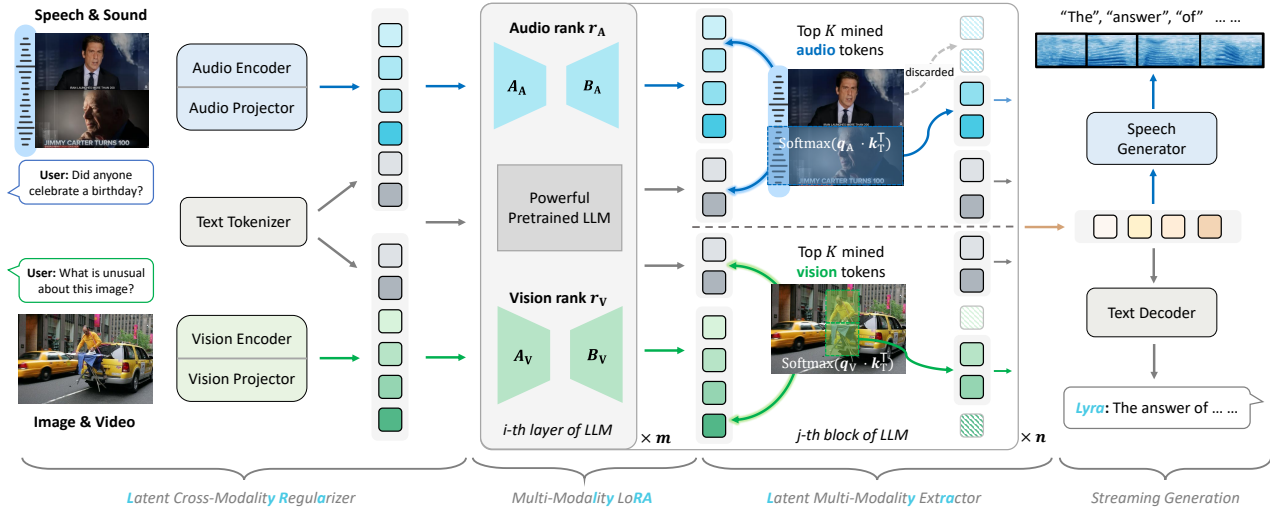


Figure 2. **The framework of Lyra.** Lyra supports multi-modal inputs. When the data contains a speech modality, we use the latent cross-modality regularizer to assist. Data from each modality is processed through encoders and projectors before being sent into the LLM. Within the LLM, multi-modality LoRA and latent multi-modality extraction modules operate synergistically, facilitating the simultaneous generation of both speech and text outputs.

3. Lyra

As shown in Fig. 2, the overall architecture of Lyra is composed of four main components: latent cross-modality regularizer, multi-modality LoRA, latent multi-modality extractor, and streaming generation. Lyra is designed as a unified framework, with each component being easily and efficiently extendable to support additional modalities and functionalities. In this paper, Lyra primarily focuses on the three modalities of audio (speech, sound), vision, and language. Therefore, in the following subsections, we will provide a detailed introduction to the mechanisms of the following modules: latent cross-modality regularizer, multi-modality LoRA, and latent multi-modality extractor. Due to space limitations, streaming speech-text generation will be detailed in Appx. B.5. Since speech contexts tend to be lengthy, the integration of long speech capabilities will be discussed at the end of this section. To ensure clarity in the following discussion, let’s define some key notations: the $\mathbf{X}_{[i]}$ be the token of modality- i . For example, $\mathbf{X}_{[\text{text}]}$ represents the text token, $\mathbf{X}_{[\text{image}]}$ represents the image token, $\mathbf{X}_{[\text{video}]}$ represents the video token, $\mathbf{X}_{[\text{speech}]}$, $\mathbf{X}_{[\text{sound}]}$ represents the speech and sound token, respectively.

3.1. Preliminary: Transcript v.s. Raw Speech

It is easy to conceive that the simplest way to enable speech interaction with LLMs is through a cascaded system based on automatic speech recognition (ASR) and text-to-speech (TTS) models, where the ASR model transcribes the user’s speech instruction into text, and the TTS model synthesizes the LLM’s response into speech. However, these cascade-based approaches (transcript) have several limitations, and

the latest SLMs [16, 52, 78] have abandoned this approach and adopted an end-to-end method to integrate the audio modality into the LLM.

Computational efficiency and latency. Cascaded models require loading and running the entire ASR model, *e.g.*, Whisper has a 32-layer encoder and decoder, which increases GPU memory by 4-10 GB compared to ours and slows inference speed by about 2×. However, latency critically impacts the user experience in MLLMs [16, 69].

Greater flexibility and higher potential. Since current SOTA ASR models cannot achieve 100% accuracy (around 90% in real-world scenarios), cascade-based approaches are unable to correct errors during the audio encoding process and instead directly input incorrect transcripts, which limits the upper bound of speech processing capabilities. Furthermore, systematic discrepancies between spoken language patterns and textual representations (*e.g.*, processing multiple-choice question options, more examples are given in Appx. A.2) are inherently challenging. Conventional transcription models, constrained by their *non-end-to-end* training paradigm, exhibit low performance and lack the flexibility required for true speech understanding of MLLM. In summary, we have placed the above specific comparative experimental results in Table 2.

3.2. Latent Cross-Modality Regularizer

For MLLMs, it is crucial to achieve effective alignment between tokens from each modality and LLM. As the view from the speech modality, there is a high degree of informational overlap with the text modality. Specifically, considering only semantic information, speech can be converted into its corresponding transcribed text. However, our exper-

Evaluation	Text-Image	Whisper-transcript	Lyra (LCMR)
DocVQA (9B)	90.0%	84.1%(-5.9%)	85.6%(+1.5%)
A12D (9B)	72.4%	60.8%(-11.6%)	66.4%(+5.6%)
InfoVQA (9B)	61.5%	55.5%(-6.0%)	56.8%(+1.3%)
Inference Time [†]	0.56 s/sample	1.39 s/sample(+148%)	0.73 s/sample(+30%)
DocVQA (74B)	91.0%	86.5%(-4.5%)	89.6%(+3.1%)

Table 2. Performance and efficiency comparison of text, Whisper v3 transcripts, and raw speech inputs on VLM benchmarks.

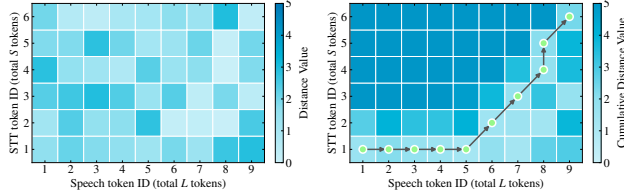


Figure 3. Illustration of the DTW algorithm in our alignment. Our goal is to make the speech tokens as similar as possible to the corresponding translated tokens.

iments from Table 2 (and Table 9 in Appx. B.1) have shown that using speech with naive alignment training as the instruction (for speech instruction and image context) generally yields less effective results compared to using transcribed text (text instruction and image context).

To address this, we aim to make the tokens from the speech modality as similar as possible to the corresponding transcribed text tokens before feeding them into LLM, thereby minimizing the loss of relevant information. Another challenge arises from the variable length of speech: a sentence can be spoken quickly or slowly while retaining the same meaning in the text modality, leading to length discrepancies. In general, the tokens produced by a speech encoder (like Whisper) tend to be much longer than the corresponding text tokens (speech-to-text, STT), *i.e.*, $\mathbf{X}_{[\text{speech}]} \in \mathbb{R}^{d \times L}$, $\mathbf{X}_{[\text{STT}]} \in \mathbb{R}^{d \times S}$, $L > S$, d is the token dimension. We define the latent distance between the l -th speech token and the s -th SST token as:

$$\text{dist}(l, s) = -\log \left[\text{softmax}(\mathbf{X}_{[\text{speech}]_l} \mathbf{X}_{[\text{STT}]_s}^\top / \tau) \right], \quad (1)$$

Where τ is the temperature. To get the minimum distance between two different length tokens, we follow the Dynamic Time Warping (DTW) algorithm:

$$\mathbf{D}_{l,s} = \text{dist}(l, s) + \min\{\mathbf{D}_{l,s-1}, \mathbf{D}_{l-1,s}, \mathbf{D}_{l-1,s-1}\}. \quad (2)$$

The illustration is shown in Fig. 3. We define the latent cross-modality regularization loss as $\mathcal{L}_{\text{LCMR}} = \frac{1}{L+S} \mathbf{D}_{L,S}$. Finally, the total loss of the system becomes the combination of two losses: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{LCMR}}$, where \mathcal{L}_{CE} is the cross-entropy loss on LLM output, and λ is a loss weight hyper-parameter. LCMR introduces additional supervised learning to optimize the audio encoder for more linguistically meaningful semantic features. Furthermore, LCMR is more suitable for powerful unified audio encoders: *The naive Whisper encoder can only extract semantic features*

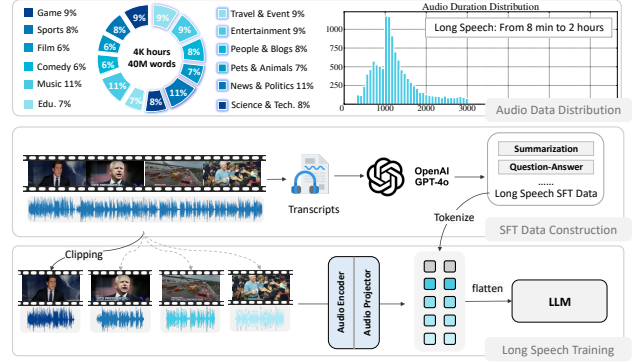


Figure 4. Long speech capability integration pipeline. Top: The proportion of question and speech categories in our long speech SFT dataset. Middle: Our pipeline for generating instruction-following data for long speech. Bottom: Our long speech SFT pipeline. Long speech segments will be clipped and flattened.

from the speech and audio, which is why we proposed LCMR to Whisper. Currently, many new and powerful audio encoder models, such as Moshi [14], SpeechTokenizer [82], and Encodec [13], produce two types of embedding features from the input audio: the *semantic features* and the *acoustic features*. Under such a unified situation, LCMR can also be extended and applied to enhance the semantic features learning, making the audio understanding more powerful (refer to Table 2 and Appx. B.1, Table 9).

3.3. Multi-Modality LoRA Pipeline

The current open-source VLM (such as Qwen2-VL) is already quite powerful. With limited data quantity and quality, jointly training vision-speech-language modalities may reduce the model’s original capabilities. Therefore, we adopt an efficient multi-modality LoRA [26] pipeline. Revisiting the notation introduced at the beginning of this section, we represent $\mathbf{X}_{[i]}$ as the token of modality- i . The modality- i can be text, image, video, speech token, and sound. Since our model involves joint training across multiple modalities, here we define $\mathbf{X}_{[M]}$ can be any combination of the above different modality tokens. The output of multi-modality LoRA can be written as:

$$\mathbf{H} = (\mathbf{B}_{[M]} \mathbf{A}_{[M]} + \mathbf{W}) \mathbf{X}_{[M]}, \quad (3)$$

where \mathbf{W} is the original weight of LLM, $\mathbf{A}_{[M]}$ and $\mathbf{B}_{[M]}$ is low-rank adapter of combination- M . During training, our Multi-Modality LoRA is integrated into each layer of the LLM. Because each modality is trained using LoRA, the process is highly efficient, achieving strong performance with minimal data while preserving much of the original model’s visual capabilities.

3.4. Latent Multi-Modality Extractor

As MLLMs expand their functionality and accommodate longer contexts, efficiently using tokens within a limited

context window becomes essential to address the long-context problem. We now consider the relationship between non-text modalities and the text modality. In response to a given question, many tokens from non-text modalities may be largely irrelevant to the question itself. For example, as shown in Fig. 2, only a subset of image tokens is relevant to the instructed question. Similarly, for the video and speech modality, only a portion of tokens from video and speech directly corresponds to the question instruction.

We observe that in LLM training, the long-context effect brought by high-resolution images, lengthy videos, and long audio (in the following subsection) often includes tokens with limited relevance, which not only increases the computational load for training and inference but also consumes unnecessary memory. To address this, we propose dynamically selecting multi-modality tokens based on their relevance to the text query, discarding redundant multi-modality tokens. To achieve this, we introduce a latent multi-modality information extraction strategy.

Concretely, instead of applying this strategy to every layer, we implement a block-based manner. Suppose the LLM consists of mn layers; we divide them into blocks of m layers each, resulting in n blocks. At the final layer of each block, we apply our following information extraction strategy, which evaluates the similarity between the attention scores of tokens from each modality and the question text tokens. We represent this with the following equation:

$$\text{topk} \left(\text{softmax} \left(\frac{\mathbf{Q}_{[\text{text}]} \mathbf{K}_{[\text{text}]}^{\top}}{\sqrt{d}} \right) \right), \quad (4)$$

where $\mathbf{Q}_{[\text{text}]}$ denotes the query corresponding to text modality tokens, and $\mathbf{K}_{[\text{text}]}^{\top}$ represents the key corresponding to tokens from other modalities. For clarity, let's assume that the length of multi-modality tokens $\mathbf{K}_{[\text{text}]}^{\top}$ is L . After passing through each block, we retain only ρL multi-modality tokens. From a block-wise perspective, the token length decays exponentially, significantly reducing computational and memory costs. A similar mechanism exists in the brain's neural processing of complex information [53]. Notably, text tokens can be extended to instruction tokens for other modalities, such as speech. This extractor enables us to handle long speech more efficiently.

3.5. Long Speech Capability Integration

There is a growing trend toward increasing the length of single-modality content processed by models, such as long text and long video inputs in MLLMs. However, existing MLLMs are limited in handling long speech due to the constraints of speech encoders. Specifically, models like Intern-Omni [47], VITA [19], and LLaMA-Omni [16] use Whisper-like encoders, which restrict audio input to around 30 seconds. VITA and Mini-Omni, which employ more

complex encoders, can process at most one minute of audio input. This limitation largely stems from the lack of suitable long speech SFT datasets and appropriate preprocessing methods. To address this issue, we developed the first SFT dataset for long speech understanding, aimed at enhancing model capabilities in handling extended audio content. Our dataset comprises about 12K long-form audio recordings, with durations ranging from several minutes to two hours. These recordings were collected from diverse YouTube sources, including informational videos, interviews, and speeches, covering a wide range of topics—from humanities and current events to technology and society. With related transcripts, we utilized LLM to generate question-and-answer pairs derived from the captions and instructions. These questions cover summarization and other types of inquiries that encourage a comprehensive understanding of long speech content. The overall question distribution and details are illustrated in Fig. 4.

Once the dataset was ready, we tackled the challenge with the speech encoder. Inspired by high-resolution image segmentation methods like LLaVA-NeXT [39], we adopted a similar strategy to better handle the speech encoder for long audio processing (illustrated in Fig. 4). However, unlike previous speech cases, a new challenge emerged: for a naive Whisper-v3 encoder, a 30-second audio clip is encoded into 1,500 tokens. Under typical short speech scenarios, an LLM can handle 1,500 tokens comfortably. When we consider long speech cases, such as a two-hour audio clip, this would result in an astonishing 360,000 tokens, which is beyond our processing capacity. Thus, it is essential to consider compression techniques on speech tokens. We compress speech tokens by combining several time-contiguous tokens into one token in the speech projector, using a view operation and multiple linear layers. The token number is varied during both training and inference. The experimental results are presented as follows:

# (Token) [↓]	100	150	300	500	1500
TextVQA ^S	75.9%	76.8%	77.8%	78.0%	76.8%
MM-Vet ^S	55.3%	54.4%	56.3%	58.8%	58.9%

The results indicate that having a higher number of speech tokens provides certain benefits. However, beyond a certain threshold, the performance improvement becomes quite limited. Taking into account both computational costs and model performance, we ultimately decided to use the 300 compressed tokens version for extending the model to handle long speech cases.

4. Experiments

In this section, we conduct a speech-centric evaluation, assessing its integration with image, video, and text modalities. We first outline our experimental framework, commencing with the experimental setup. Subsequently, we compare Lyra with leading methods on various benchmarks

Omni Comparison		Text-Image			Text-Video			Image-Speech			Text-Speech
Method	Params.	TextVQA	MME	MM-Vet	VideoMME	MVBench	Egoschema	TextVQA ^S	DocVQA ^S	ChartQA ^S	LibriSpeech ⁺
Mini-Gemini	8B	71.9	1989	53.5	-	-	-	-	-	-	-
LLaVA-OV	7B	65.4	1998	57.5	58.2	56.7	60.1	-	-	-	-
Intern-VL2	8B	77.4	2211	60.0	54.0	66.4	-	-	-	-	-
Mini-Omni	7B	-	-	-	-	-	-	-	-	-	4.5
SALMONN	13B	-	-	-	-	-	-	-	-	-	2.1
Qwen2-Audio	8B	-	-	-	-	-	-	-	-	-	1.6
Intern-Omni	8B	80.6	2210	60.0	-	-	-	69.1	79.9	56.0	-
VITA	66B	-	2097	41.6	59.2	-	-	-	-	-	8.1
EMOVA	14B	82.0	2205	55.8	-	-	-	-	-	-	4.0
Lyra-Mini	3B	78.3	1884	51.2	55.0	62.5	54.1	73.4	74.8	40.7	2.4
Lyra-Base	9B	82.6	2335	65.0	68.0	67.5	63.2	80.7	87.1	67.5	1.8
Lyra-Pro	74B	83.5	2485	71.4	69.9	72.3	75.8	81.0	89.4	68.5	1.7

Table 3. **Omni-comparison on vision-language-speech benchmarks.** Bench^S indicates that it uses speech instruction as the input.

and qualitative results. Detailed component-wise analysis (*based on Lyra-Base*) is given at the end of this section. For more experiment details and results refer to our Appx. A.

4.1. Experimental Setup

Implementation Details. In this study, we instantiate Lyra with the following designs and settings:

- **Strong vision encoders and LLMs:** Building on the previously applied vision model Qwen2-VL’s ViTs and LLMs [64], they can now process images of any resolution, dynamically converting them into a variable number of visual tokens. We have also designed three versions: For Lyra-Mini, we use Qwen2-VL 2B. For Lyra-Base, we apply Qwen2-VL 7B. For Lyra-Pro, we choose Qwen2-VL 72B.

- **Efficient audio encoder:** We adopted Whisper-large-v3 [51] (Lyra-Base and Lyra-Pro) and its light-weight version, Whisper-large-v3-turbo (Lyra-Mini), which have been trained on a large amount of audio data and has strong capabilities in speech recognition and translation.

- **Four-stage training for omni-cognition:** (refer to our Appx. A for specific details) In the first stage, we conduct text-to-speech pretraining to train the speech encoder. In the second stage, we perform joint training on text, image, and speech modalities to train the LLM along with the corresponding projectors. In the third stage, we train the LLM to extend the model’s capability in handling long speech. In the fourth stage, we train our speech generator, enabling the model to simultaneously output text and corresponding audio in a streaming manner.

Datasets and Evaluations. For model optimization, we construct high-quality data for omni-understanding, long-context speech, and speech generation.

- **High-quality multi-modal dataset:** Based on the Mini-Gemini SFT [34] dataset, we carefully collected and extended a high-quality multi-modal dataset that covers common scenes and document images and speeches. It contains about 1.5M open-source image-speech, text-image, and text-speech instruction samples. To enhance the gen-

eralization of speech modality, we utilize ChatTTS [1] with varying configurations to generate different audios.

- **Long speech SFT dataset:** As mentioned in Sec. 3.5, we constructed a delicate long speech SFT dataset for long speech capability integration with 12K samples. The dataset involves a distribution of longer audio durations and covers a wide range of domains.

- **Evaluation:** Unlike the previous omni-model [7, 19], which only tested text-to-speech capabilities, we employed a more omni comprehensive evaluation that covers interactions across the image, video, text, and speech modalities.

4.2. Main Results

Quantitative Results. In the quantitative analysis experiments, we primarily compare our model with current leading VLMs, such as Mini-Gemini [34], LLaVA-OV [31], Intern-VL2 [10], and SLM, like Mini-Omni [69], SALMONN [56], Qwen2-Audio [11], and Omni models including Intern-Omni [47], AnyGPT [77], VITA [19], and EMOVA [7]. The input modalities we compare are also the most widely used, including text-image, text-video, image-speech, and text-speech. Detailed results are presented in Table 3. In calculating the total parameters of the model, we considered all modality-specific encoders, projectors, and related components. Our model includes three versions: a mini version (3B), a base version (9B), and a pro version (74B). Benefiting from multi-modality LoRA and Qwen2-VL, our model maintains relatively high performance in text-image and text-video tasks. For the speech modality, as we mentioned in the Introduction part, previous models have evaluated the speech modality rather crudely, without extensively testing metrics for interactions between the speech modality and other modalities. Our model comprehensively outperforms existing omni models in both image-speech (with an improvement of approximately 9%) and text-speech (with an improvement of approximately 2%) tasks. Additionally, our model is more lightweight, requiring fewer training samples.

Effectiveness	TexVQA		MM-Vet		LibriSpeech
	S+I	T+I	S+I	T+I	S+T
Baseline	-	82.3	-	62.8	-
\mathcal{L}_{CE}	76.7	79.5	53.1	61.1	1.9
$\mathcal{L}_{CE} + \lambda \mathcal{L}_{LCMR}$	77.8	80.1	58.1	62.6	2.0

Table 4. **Latent cross-modality regularizer.** With our regularizer, the performance of both the speech-image (S+I) and text-image (T+I) modalities improves, and *the gap narrows*.

Method	Overall	Short	Medium	Long
Baseline (7B)	62.8	73.8	62.3	52.3
Baseline + subtitle	64.4	76.2	63.4	53.4
LSCI (7B, solve 33%)	78.6	89.8	77.7	74.8
Baseline + LSCI	66.2	75.7	64.0	58.9
GPT-4o [46] + subtitle	77.1	82.8	76.6	72.1

Table 5. **Effectiveness of long speech capability integration.** Lyra integrated with long speech ability, using only audio input, can handle one-third of VideoMME cases, and its accuracies on long, medium, short metrics are better than the current best VLM.

Qualitative Results. To ascertain the omni comprehension prowess of Lyra in real-world settings, we apply it to a variety of understanding and reasoning tasks in the bottom left part of Fig. 1 and our Appx. C, Figs. 12-15. By contrast, Lyra can well solve more complex multi-modality cases.

4.3. Component-Wise Analysis

Latent Cross-Modality Regularizer. We first delve into the proposed latent cross-modality regularizer and report results in Table 4. It is clear that the model achieves significant gains for both speech-image inputs and text-image inputs, with the regularizer integrated as an assistance between speech modality and text modality. In the training of the image-speech-text tri-modal model, introducing the \mathcal{L}_{LCMR} significantly enhances the performance of both image-speech and image-text alignments, reducing the gap between them. We also observe that with only \mathcal{L}_{CE} , image-text performance lags behind image-speech by 8% on the MM-Vet benchmark. However, the performance of speech-text remains relatively unchanged whether using the CE loss or joint loss. Therefore, previous omni models [7, 19] that assessed the speech modality just based on the LibriSpeech [48] WER metric for speech-text alignment are rather arbitrary. We need to evaluate the performance of the speech modality alongside other modalities to accurately measure the effectiveness of omni-models. This also demonstrates the effectiveness of our \mathcal{L}_{LCMR} . More ablations of hyper-parameter λ and speech/audio improvement results are shown in Appx. B.1, Tables 9 and 11.

Latent Multi-Modality Extractor. For the latent multi-modality extractor (LMME) module, we focus primarily on its **effectiveness** and **efficiency** in multi-modal tasks.

First, to verify the effectiveness of our extractor module, we examine the retention of multi-modal tokens. We

Modality	Benchmark	Baseline	+ SFT	+ MLoRA
Image	TextVQA [54]	82.3	81.3	82.6
	MME [17]	2332	2275	2335
	MMMU [76]	49.2	48.7	50.8
Video	VideoMME [18]	62.8	61.0	62.8
	MVBench [33]	66.7	66.8	67.2
	EgoSchema [42]	62.4	63.5	63.2
Speech	TextVQA ^S [54]	-	77.8	80.0
	DocVQA ^S [60]	-	84.0	84.6
	MM-Vet ^S [75]	-	54.0	60.0

Table 6. **Effectiveness of multi-modality LoRA (MLoRA).** For powerful pretrained models, adding a new modality can impair the abilities of other modalities. MLoRA can effectively address it.

primarily assess three types of tokens: image tokens, video tokens, and speech tokens. The specific visualizations are shown in Figs. 5 and 10. As seen in the figures, our model ultimately retains only about 10%-25% of the tokens across all three modalities. Moreover, the retained token positions are highly relevant to the user-provided instructions, effectively helping to remove information unrelated to the instructions and thereby accelerating training and inference. We also have included more performance and comparisons with FastV [8], related to LMME in Appx. B.2, Table 10.

Second, we analyze its efficiency with the examined metrics including prefill time, tokens-per-second (TPS), FLOPs, and memory usage on the GPUs. The detailed comparison is shown in Appx. B.2, Tables 14 and 12.

Long-Speech Capability Integration. After performing SFT on our long speech 12K data mentioned Sec. 3.5, we design the following experiments to validate the capabilities in processing long speech and latent multi-modality extraction, given the current lack of a long speech benchmark.

The first experiment is the long speech ‘‘Needle in a Haystack’’ evaluation. We selected five audio files, each more than three hours in length, and inserted open-ended audio questions and answers at various points throughout the files. The results are shown on the left side of Fig. 6. According to the figure, we observe that, without enhancing long-speech processing capabilities, the model can handle up to approximately eight minutes of audio. Beyond that length, it fails to generate a proper output (Fig. 6a). However, with SFT on our Lyra long speech 12K data, the model can handle audio lengths of up to 4,500 seconds. With audio exceeding 4,500 seconds, the model’s memory usage surpasses the limit (Fig. 6b). By leveraging the latent multi-modality extractor module, we achieve the ability to process even longer audio, extending up to and beyond two hours (Fig. 6c). Additionally, In Fig. 6d, we visualize the token-level attention retention and variations for the ‘‘needle’’ with the information extractor module, under the same question instructions. Notably, we can see that as the needle is placed in different locations, the information extractor module dynamically adjusts the attention distribution accordingly.

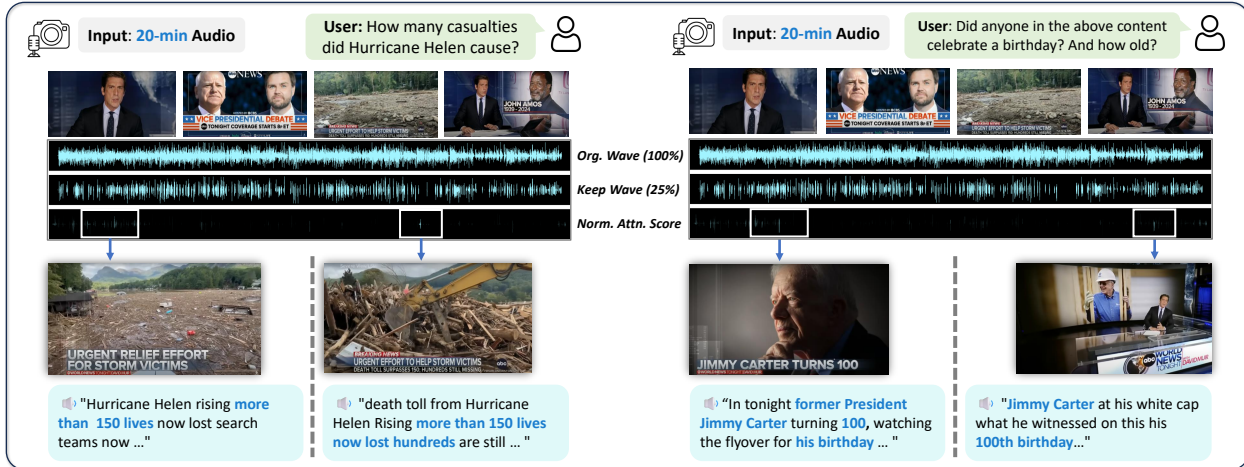


Figure 5. **Visualization of latent multi-modality extractor in audio modality.** Through latent multi-modality information extraction, semantic tokens related to the instruction are retained, reducing the computational cost of the MLLM. More visualizations of the image and video modality and different blocks can be found in Appx. B.2 and Fig. 10.



Figure 6. **Comparison of needle in long speech haystack (average with five samples).** (a) The baseline model can not retrieve right needles after 450 seconds. (b) Model finetuned on our long speech datasets can not retrieve right needles after 4,500 seconds and achieves 96% accuracy in 4,500 seconds. (c) Our latent extractor, trained on our long speech datasets, can retrieve longer audio (9,900 seconds), and presents 98% accuracy in 4,500 seconds. (d) As the position of the needle changes, the attention in our model also shifts accordingly.

The second experiment is based on VideoMME. This benchmark includes videos ranging from 30 seconds to one hour. We first extract the audio from these videos and feed only the audio data into our long speech model to obtain predictions and perform the VideoMME evaluation. Along with generating predictions, we also require our model to output whether it can answer the question based on the audio alone. Specific results are shown in Table 5. From the table, it is evident that long audio can resolve about one-third of the test samples, with model accuracy exceeding 78%, significantly outperforming the 7B model. We integrate the long-speech output into our Lyra model, which ultimately performs better than using subtitles alone.

Multi-Modality LoRA (MLoRA) Pipeline. The effectiveness results of MLoRA are presented in Table 6. Compared to multi-modal SFT, MLoRA maintains better original vision performance while enhancing the capability in new modalities like speech. Additionally, our framework is more efficient, achieving better results with less data (67%).

Intern-Omni	VITA	EMOVA	Lyra
27M samples	5M samples	4M samples	2.7M samples

5. Conclusion

In conclusion, Lyra marks a major advancement in MLLMs by efficiently integrating speech, vision, and language with lower computational costs (*less data, faster speed*). We emphasize speech to improve its interaction with other modalities. With our proposed modules and high-quality SFT datasets, Lyra achieves state-of-the-art results on vision-speech, speech-language, and vision-language benchmarks, which offer a more comprehensive omni-modal evaluation than prior work. Our findings highlight the critical role of speech in multimodal understanding, which previous MLLMs have underutilized. We hope it inspires further exploration of speech, especially long-form, in MLLMs.

Acknowledgements. The study was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R, Hong Kong General Research Fund (14208023), Hong Kong AoE/P-404/18, and the Center for Perceptual and Interactive Intelligence (CPII) Ltd under InnoHK supported by the Innovation and Technology Commission.

References

- [1] 2noise. ChatTTS. <https://github.com/2noise/ChatTTS>, 2024. 6, 12, 13, 17
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1
- [3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 15
- [4] 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:2308.12966*, 2023. 2
- [5] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. 2
- [6] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *ICLR*, 2025. 2
- [7] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. EMOVA: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024. 1, 2, 6, 7
- [8] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, pages 19–35. Springer, 2025. 2, 7, 13
- [9] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongLoRA: Efficient fine-tuning of long-context large language models. In *ICLR*, 2024. 2
- [10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6
- [11] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. 1, 6, 14
- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 1
- [13] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*. 4
- [14] 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. 4
- [15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 14
- [16] 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. 1, 2, 3, 5, 14, 16
- [17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiwu Zheng, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023. 7
- [18] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 7
- [19] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiwu Zheng, et al. VITA: Towards open-source interactive omni multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024. 1, 2, 5, 6, 7, 14
- [20] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. *arXiv preprint arXiv:2310.01801*, 2023. 2
- [21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 14
- [22] Félix Gontier, Romain Serizel, and Christophe Cerisara. Automated audio captioning by fine-tuning bart with audioset tags. In *DCASE 2021-6th Workshop on Detection and Classification of Acoustic Scenes and Events*, 2021. 17
- [23] Google. Gemma: Introducing new state-of-the-art open models. <https://blog.google/technology/developers/gemma-open-models/>, 2024. 1
- [24] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006. 16
- [25] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460, 2021. 16
- [26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *ICLR*, 2021. 4

- [27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1
- [28] 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, et al. Mixtral of experts. *arXiv:2401.04088*, 2024. 1
- [29] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 14, 15
- [30] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, Jaeheon Sim, Jinwoo Lee, and Kyogu Lee. Exploring train and test-time augmentations for audio-language learning. *arXiv preprint arXiv:2210.17143*, 2022. 17
- [31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 6
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 1
- [33] Kunchang Li, Yali Wang, Yanan 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*, pages 22195–22206, 2024. 7
- [34] 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. 1, 2, 6, 12
- [35] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2025. 2
- [36] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024.
- [37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NerulPS*, 2023. 1
- [38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, 2024. 2, 5
- [40] Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. *NerulPS*, 36, 2024. 2
- [41] Zhengrui Ma, Qingkai Fang, Shaolei Zhang, Shoutao Guo, Yang Feng, and Min Zhang. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation. *arXiv preprint arXiv:2406.06937*, 2024. 16
- [42] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NerulPS*, 36:46212–46244, 2023. 7
- [43] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Audio captioning transformer. *arXiv preprint arXiv:2107.09817*, 2021. 17
- [44] Microsoft. Edge-TTS. <https://github.com/rany2/edge-tts>, 2024. 17
- [45] OpenAI. ChatGPT. <https://openai.com/blog/chatgpt/>, 2023. 1, 2
- [46] OpenAI. GPT-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. 1, 7
- [47] OpenGVLab. InternOmni: Extending internvl with audio modality. <https://internvl.github.io/blog/2024-07-27-InternOmni/>, 2024. 5, 6, 17
- [48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE, 2015. 7, 12, 13
- [49] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*, 2021. 16
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 14
- [51] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518. PMLR, 2023. 6, 14
- [52] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023. 2, 3
- [53] John T Serences and Steven Yantis. Selective visual attention and perceptual coherence. *Trends in cognitive sciences*, 10(1):38–45, 2006. 5
- [54] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 7
- [55] Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendorfer. Snac: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*, 2024. 16
- [56] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *ICLR*, 2024. 6, 14

- [57] Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *arXiv preprint arXiv:2406.10774*, 2024. 2
- [58] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *NeurIPS*, 36, 2024. 17
- [59] 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. https://github.com/tatsu-lab/stanford_alpaca, 2023. 1
- [60] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *ICDAR 2021*, 2021. 7
- [61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 1, 2
- [62] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 15
- [63] Common Voice. Common Voice. <https://commonvoice.mozilla.org/en/datasets>, 2024. 12
- [64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 6, 16
- [65] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. *arXiv:2311.03079*, 2023. 2
- [66] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 2
- [67] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-GPT: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 1, 2
- [68] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 2
- [69] Zhifei Xie and Changqiao Wu. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024. 1, 3, 6, 14
- [70] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. LongVILA: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 2
- [71] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2
- [72] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *ACL*, 2024. 13, 14
- [73] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *CVPR*, pages 19792–19802, 2025. 2
- [74] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2
- [75] 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:2308.02490*, 2023. 7
- [76] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 7
- [77] 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. 1, 2, 6
- [78] 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. 2, 3, 14
- [79] Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, et al. Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024. 14
- [80] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2
- [81] Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. Streamspeech: Simultaneous speech-to-speech translation with multi-task learning. *arXiv preprint arXiv:2406.03049*, 2024. 16
- [82] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechokeizer: Unified speech tokenizer for speech language models. In *ICLR*, 2024. 4
- [83] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2O: Heavy-hitter oracle for efficient generative inference of large language models. *NeurIPS*, 36:34661–34710, 2023. 2
- [84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 1, 16