



# EMOVA: Empowering Language Models to See, Hear and Speak with Vivid Emotions

Kai Chen<sup>1\*</sup>, Yunhao Gou<sup>1,6\*</sup>, Runhui Huang<sup>2\*</sup>, Zhili Liu<sup>1,3\*</sup>, Daxin Tan<sup>3\*</sup>, Jing Xu<sup>4</sup>, Chunwei Wang<sup>3</sup>, Yi Zhu<sup>3</sup>, Yihan Zeng<sup>3</sup>, Kuo Yang<sup>3</sup>, Dingdong Wang<sup>4</sup>, Kun Xiang<sup>5</sup>, Haoyuan Li<sup>5</sup>, Haoli Bai<sup>3</sup>, Jianhua Han<sup>3</sup>, Xiaohui Li<sup>3</sup>, Weike Jin<sup>3</sup>, Nian Xie<sup>3</sup>, Yu Zhang<sup>6</sup>, James T. Kwok<sup>1</sup>, Hengshuang Zhao<sup>2</sup>, Xiaodan Liang<sup>5</sup>, Dit-Yan Yeung<sup>1</sup>, Xiao Chen<sup>3</sup>, Zhenguo Li<sup>3</sup>, Wei Zhang<sup>3</sup>, Qun Liu<sup>3</sup>, Lanqing Hong<sup>3†</sup>, Lu Hou<sup>3†</sup>, Hang Xu<sup>3†</sup>

<sup>1</sup>Hong Kong University of Science and Technology <sup>2</sup>The University of Hong Kong  
<sup>3</sup>Huawei Noah's Ark Lab <sup>4</sup>The Chinese University of Hong Kong  
<sup>5</sup>Sun Yat-sen University <sup>6</sup>Southern University of Science and Technology

Project Page: <https://emova-ollm.github.io/>

## Abstract

*GPT-4o*, an omni-modal model that enables vocal conversations with diverse emotions and tones, marks a milestone for omni-modal foundation models. However, empowering Large Language Models to perceive and generate images, texts, and speeches end-to-end with publicly available data remains challenging for the open-source community. Existing vision-language models rely on external tools for speech processing, while speech-language models still suffer from limited or totally without vision-understanding capabilities. To address this gap, we propose the **EMOVA** (**EMotionally Omni-present Voice Assistant**), to enable Large Language Models with end-to-end speech abilities while maintaining the leading vision-language performance. With a semantic-acoustic disentangled speech tokenizer, we surprisingly notice that omni-modal alignment can further enhance vision-language and speech abilities compared with the bi-modal aligned counterparts. Moreover, a lightweight style module is introduced for the flexible speech style controls including emotions and pitches. For the first time, **EMOVA** achieves state-of-the-art performance on both the vision-language and speech benchmarks, and meanwhile, supporting omni-modal spoken dialogue with vivid emotions.

## 1. Introduction

OpenAI GPT-4o [70], a novel milestone for the omni-modal foundation models, has rekindled people's attention on in-

\*Equal contribution, listed in alphabetical order by surname.

†Corresponding authors: {honglanqing, houlu3, xu.hang}@huawei.com

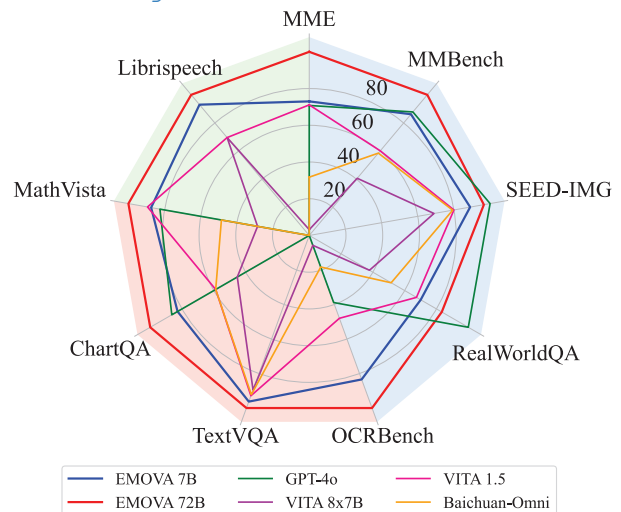


Figure 1. **EMOVA** is the very first omni-modal LLM with state-of-the-art performance on both vision-language and speech benchmarks simultaneously. See detailed results in Table 2.

telligent assistants that can *see* (*i.e.*, perceiving fine-grained visual inputs), *hear* (*i.e.*, understanding vocal instructions) and *speak* (*i.e.*, generating vocal responses) simultaneously. Existing Multi-modal Large Language Models (MLLMs) mostly focus on two modalities only, either vision-language [1, 40] or speech-language [11, 87], demonstrating severe demands for omni-modal models with visual, language and speech abilities. *How to empower Large Language Models (LLMs) to effectively process omni-modal data in an end-to-end manner remains an open question.*

Existing omni-modal LLMs [10, 19] generally build on Vision LLMs and integrate the speech modality by adopting a speech encoder like Whisper [74], which extracts *continuous* speech features, similar to how images are processed, to

enable speech understanding. These models, however, still rely on external Text-to-Speech (TTS) tools for generating speech responses, limiting their ability to support real-time interactions. AnyGPT [93], on the contrary, opts for a fully *discretization* way, which first discretizes all data modalities (*i.e.*, images, texts, and speeches), followed by omni-modal auto-regressive modeling. This enables AnyGPT to handle multiple modalities with a **unified end-to-end** framework, facilitating **real-time interactions** with the help of streaming decoding. However, the discrete vision tokenizer used by AnyGPT struggles to capture visual details, especially for high-resolution images, falling far behind its continuous counterparts on vision-language benchmarks. Furthermore, none of the existing works explore the speech style controls (*e.g.*, emotions and pitches) with LLMs. Thus, our question arises: *How to build an end-to-end omni-modal LLM enabling emotional spoken dialogue while maintaining state-of-the-art vision-language performance?*

In this paper, we propose **EMOVA** (**EMotionally Omni-present Voice Assistant**), a novel end-to-end omni-modal LLM with state-of-the-art vision-language and speech capabilities while supporting emotional spoken dialogue. Fig. 2 shows an overview of the model framework. A *continuous* vision encoder captures fine-grained visual details, while a *discrete* speech tokenizer and detokenizer empower end-to-end speech understanding and generation. Specifically, the speech-to-unit (S2U) tokenizer tokenizes the input speech waveforms to discrete speech units, while the unit-to-speech (U2S) detokenizer reconstructs the speech waveforms from LLM’s output speech units. To seamlessly integrate speech modality with LLMs, we meticulously design a **semantic-acoustic disentangled** speech tokenizer to decouple the semantic contents and acoustic styles of input speeches [77], where 1) *semantic content* (*i.e.*, what it says) captures the semantic meanings of the input speeches, which is finally discretized and aligned with LLMs, while 2) *acoustic style* (*i.e.*, how it says) captures the diverse speech styles (*e.g.*, emotions and pitches). Utilizing the semantic-acoustic disentanglement of our speech tokenizer, we further introduce a lightweight style module to support spoken dialogue with vivid emotions and pitches. As in Sec. 4.1, this disentanglement design better facilitates the modality alignment among texts and speeches while maintaining flexibility for diverse speech style controllability and personalization.

With **EMOVA**’s end-to-end omni-modal framework, we empirically demonstrate publicly available bi-modal image-text and speech-text data are sufficient for the omni-modal alignment with the text modality as a bridge, eliminating the need for omni-modal data (*i.e.*, image-text-speech), which is usually scarce. Surprisingly, we find that the omni-modal alignment can further improve both the vision-language and speech capabilities via joint optimization, even when compared with their bi-modal aligned counterparts. Ultimately,

Method	Visual	Text	Speech		
			Understand	Gen.	Emotion
<i>Vision Large Language Models</i>					
LLaVA [49]	✓	✓	✗	✗	✗
Intern-VL [10]	✓	✓	✗	✗	✗
<i>Speech Large Language Models</i>					
Qwen-Audio [11]	✗	✓	✓	✗	✗
Mini-Omni [87]	✗	✓	✓	✓	✗
LLaMA-Omni [17]	✗	✓	✓	✓	✗
<i>Omni-modal Large Language Models</i>					
VITA [19, 20]	✓	✓	✓	✗	✗
Ola [55]	✓	✓	✓	✗	✗
Any-GPT [93]	✓	✓	✓	✓	✗
Baichuan-Omni [43]	✓	✓	✓	✓	✗
<b>EMOVA (ours)</b>	✓	✓	✓	✓	✓

Table 1. **Comparison of Multi-modal Large Language Models.** **EMOVA** is the very first Omni-modal LLM capable of emotional spoken dialogue with state-of-the-art vision-language and speech capabilities simultaneously. “Gen.” stands for Generation.

only a small amount of omni-modality samples are required to teach the model to respond in the desired format. For the first time, **EMOVA** obtains state-of-the-art results on both vision-language and speech benchmarks (see Table 2). The main contributions of this work contain three parts:

1. We propose **EMOVA**, a novel end-to-end omni-modal LLM that can see, hear and speak. A continuous vision encoder with a semantic-acoustic disentangled speech tokenizer is adopted for seamless omni-modal alignment and diverse speech style controllability.
2. We introduce an efficient text-centric omni-modal alignment which can further enhance the vision-language and speech abilities, surpassing their bi-modal aligned counterparts (*i.e.*, image-text only and speech-text only).
3. For the first time, **EMOVA** obtains state-of-the-art comparable results on both the vision-language and speech benchmarks simultaneously and further supports flexible spoken dialogues with vivid emotions.

## 2. Related Work

**Vision Large Language Models** (VLLMs) integrate the vision modality into LLMs [7, 81], enabling the advanced understanding and reasoning over visual instructions [1, 26, 27, 49]. Recent VLLM works can be categorized into three directions, 1) *Vision encoders* [5, 6, 71] are enhanced and aggregated for robust representations [45, 46, 80]. 2) *High-resolution* methods are proposed to overcome the fixed resolution of pre-trained vision encoders (*e.g.*,  $336 \times 336$  for CLIP [73]), enabling LLMs to perceive fine-grained visual information [14, 31, 48, 58]. 3) *High-quality instruction data* is essential for VLLMs to generate accurate and well-formed responses [10, 36, 40]. Besides achieving state-of-the-art vision-language performance, we further introduce speech understanding and generating abilities to **EMOVA**.

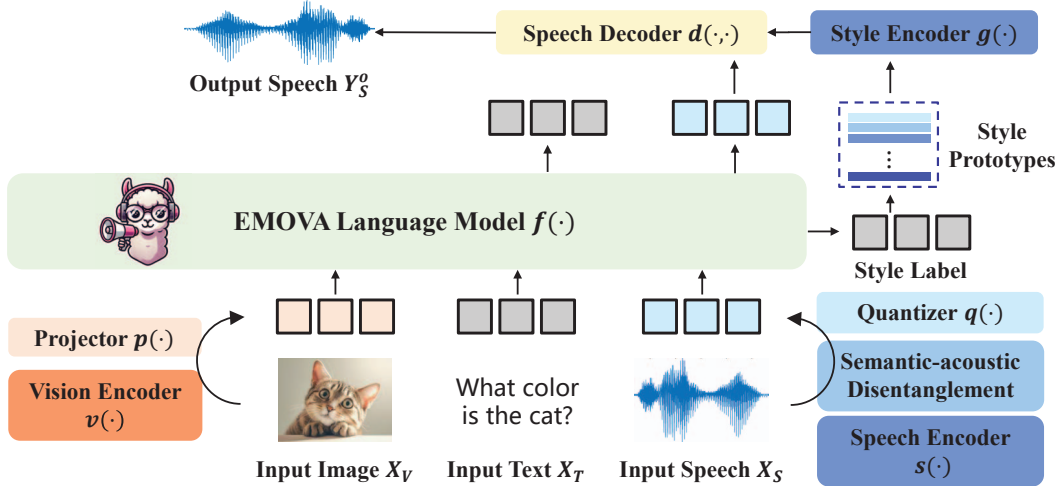


Figure 2. **Model architecture of EMOVA.** The vision encoder extracts continuous visual features, which are projected into the textual embedding space as visual tokens, while the input speech is encoded and quantized into discrete speech units. Given the omni-modal inputs, **EMOVA** can generate both textual and speech responses with vivid emotional controls. Check Sec. 3 for more architectural details.

**Speech Large Language Models** (SLLMs) empower the speech interaction with LLMs. *Continuous SLLMs* [11, 84] adopt the speech encoders [74] to extract continuous speech embeddings for LLM, which, however, only support speech understanding, relying on external TTS modules for speech generation, and therefore, hampering real-time interaction. *Discrete SLLMs* [95], instead, first discretize speech signals with speech tokenizers, followed by auto-regressive modeling. Recent works [17, 87] further combine the continuous speech encoders with discrete speech tokenizers for better results. Although effective, none of the existing works explore speech style controllability in SLLMs (e.g., emotions and pitches), which is essential for spoken dialogue.

**Omni-modal Large Language Models** support visual, text, and speech abilities with a unified architecture. Similar to continuous SLLMs, InternOmni [10] and VITA [19] connect a speech encoder with VLLMs, supporting speech understanding only. Instead, AnyGPT [93] proposes a unified architecture to discretize and conduct auto-regressive modeling for image, text, and audio simultaneously, which, however, suffers from inevitable information loss brought by discretization, especially for the high-resolution visual inputs. Our **EMOVA** is the **very first** unified Omni-modal LLM with state-of-the-art vision-language and speech performance at the same time.

### 3. Architecture

#### 3.1. Formulation

Denote LLM as  $f(\cdot)$  and text, visual and speech inputs as  $\mathbf{X}_T$ ,  $\mathbf{X}_V$  and  $\mathbf{X}_S$ , respectively.  $\mathbf{X}_T$  is converted to discrete tokens  $\mathbf{U}_T$  via a text tokenizer [21], while the  $\mathbf{X}_V$  is first encoded with a vision encoder  $v(\cdot)$  as  $\mathbf{E}_V = v(\mathbf{X}_V)$ , and then projected into the textual embedding space with a projector

$p(\cdot)$  as  $\mathbf{H}_V = p(\mathbf{E}_V)$ . As for the speech input  $\mathbf{X}_S$ , a *Speech-to-Unit* (S2U) procedure is required. Specifically,  $\mathbf{X}_S$  first goes through a speech encoder  $s(\cdot)$  as  $\mathbf{E}_S = s(\mathbf{X}_S)$ , which is then discretized by the quantizer  $q(\cdot)$  as  $\mathbf{U}_S = q(\mathbf{E}_S)$ . The LLM  $f(\cdot)$  is then trained to compute the joint probability of the output texts  $\mathbf{U}_T^o$  and speech units  $\mathbf{U}_S^o$  as

$$\mathbb{P}(\mathbf{U}_T^o, \mathbf{U}_S^o | \mathbf{U}_{omni}) = \prod_{i=1}^L \mathbb{P}(\mathbf{x}_i | \mathbf{U}_{T,<i}^o, \mathbf{U}_{S,<i}^o, \mathbf{U}_{omni}), \quad (1)$$

where  $\mathbf{x}_i \in \mathbf{U}_T^o \cup \mathbf{U}_S^o$ ,  $L = |\mathbf{U}_T^o| + |\mathbf{U}_S^o|$  and  $\mathbf{U}_{omni} = \mathbf{U}_T \cup \mathbf{U}_S \cup \mathbf{H}_V$ , which stands for the omni-modal inputs. The output response units  $\mathbf{U}_S^o$  are then recovered into the output speech waveform  $\mathbf{Y}_S^o$  via a *Unit-to-Speech* (U2S) decoder  $d(\cdot, \cdot)$  with an emotion style embedding  $\mathbf{E}_{style}^o$  to realize the vivid emotional spoken dialogue controllability (Sec. 3.2).

**LLM.** We utilize the Qwen-2.5 [79] model families as the base LLMs of **EMOVA** with three configurations (i.e., 3B, 7B, and 72B) for usage under different budgets.

**Vision encoder and projector.** We use the QwenViT [78] as the visual encoder  $v(\cdot)$  with an MLP vision projector  $p(\cdot)$  with a  $4\times$  downsample rate for all variants of **EMOVA**.

#### 3.2. Speech Tokenization

**Speech-to-unit (S2U) tokenizer.** Following [77], we use the SPIRAL [32] architecture for the speech encoder  $s(\cdot)$  to capture both phonetic and tonal information, which is then discretized by the quantizer  $q(\cdot)$  with finite scalar quantization (FSQ) [64]. The size of the speech codebook is 4,096, while the sample rate is 25 tokens per second. Once discretized, the speech modality can be integrated into LLMs by concatenating the text vocabulary and speech codebook.

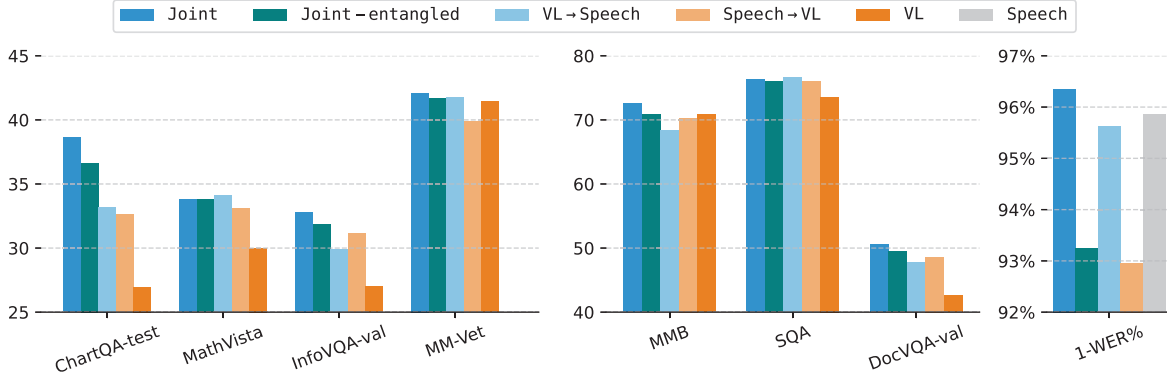


Figure 3. **Comparison between different omni-modal alignment paradigms.** 1) *Joint* training demonstrates consistent improvements over VL and Speech, suggesting that omni-modal alignment can be beneficial across modalities. 2) *Joint* training outperforms both VL→Speech and Speech→VL, revealing that joint training is more superior and efficient than sequential training. 3) *Joint* is superior to *Joint-entangled*, highlighting the effectiveness of the semantic-acoustic disentanglement, as discussed in Sec. 3.2.

Our S2U tokenizer provides the following advantages: 1) *Data efficiency*: after pre-training on large-scale unlabeled speech data, it requires only a small amount of speech-text pair data for easy adaptation. 2) *Bilingual*: the speech codebook is shared among languages (*i.e.*, English and Chinese), sharing unit modeling capabilities across languages. Check more training details and comparisons in Appendix A.1.

**Semantic-acoustic disentanglement.** To align the speech units seamlessly with the highly semantic embedding space of LLMs, we opt for decoupling the semantic contents and acoustic styles of input speeches. Given input speeches  $\mathbf{X}_S$ , both semantic embedding  $\mathbf{E}_{semantic}$  and style embeddings  $\mathbf{E}_{style}$  are extracted separately as

$$\{\mathbf{E}_{semantic}, \mathbf{E}_{style}\} = \mathbf{E}_S = s(\mathbf{X}_S). \quad (2)$$

Only  $\mathbf{E}_{semantic}$  is quantified by  $q(\cdot)$  to generate the speech units  $\mathbf{U}_S$ . By using different  $\mathbf{E}_{style}$  while maintaining the same  $\mathbf{E}_{semantic}$ , we can easily control the recovered speech styles without disturbing the semantic contents of recovered speeches. Moreover, the disentanglement facilitates modality alignment among speeches and texts, as later in Sec. 4.1.

**Unit-to-speech (U2S) detokenizer with style controls.** Building on VITS [35], our U2S detokenizer adopts a conditional VAE architecture (*cf.*, Fig. 7). To achieve flexible style controls, we utilize the semantic-acoustic disentanglement of our S2U tokenizer (as discussed above) and adopt a novel style embedding to control the speech styles (*e.g.*, genders, emotions, and pitches). Specifically, the LLM  $f(\cdot)$  is trained to generate both the output speech units  $\mathbf{U}_S^o$  and a style label. The speech units  $\mathbf{U}_S^o$  are converted to the unit embeddings  $\mathbf{E}_{semantic}^o$ , while the style label is utilized to generate a unique style prototype  $\mathbf{E}_{style}^o$ . Both  $\mathbf{E}_{semantic}^o$  and  $\mathbf{E}_{style}^o$  are taken as inputs to speech decoder  $d(\cdot, \cdot)$  to synthesize output speeches  $\mathbf{Y}_S^o = d(\mathbf{E}_{semantic}^o, \mathbf{E}_{style}^o)$ .

Our U2S detokenizer is pre-trained on LibriTTS [92] and AISHELL-1 [2] and subsequently fine-tuned on synthetic style-rich speech data. Due to the scarcity of real-life style-rich data, we utilize TTS tools [16] to synthesize the speech samples diverse in genders, pitches, and emotions. As for the style prototypes, Emotion2Vec [59] is adopted to select the most representative samples with the highest confidence in conveying the desired style. Our empirical results reveal that even one representative style reference speech has been sufficient to control the speech styles flexibly and precisely. Check Appendix A.2 for more details.

## 4. Training Omni-modal LLMs

To achieve omni-model alignment, it is ideal to use large-scale *omni-modal image-text-speech* data, which, however, is either without reach due to copyrights [67] or limited in quality [65]. An alternative is to use the existing image-text data with the TTS-synthesized speeches, which is not only computationally expensive but also hampers data diversity, as most TTS tools generate speeches in similar patterns. Recent works [10, 19] choose to integrate the speech modality into a well-structured VLLM via a *sequential* training manner with *bi-modal* alignment data. *However, the relationships among modalities and how to effectively leverage multiple bi-modal alignment datasets remain unclear.*

In this work, we explore omni-modal text-centric alignment by utilizing the publicly available bi-modal alignment datasets, including image-text (*e.g.*, image captioning) and speech-text (*e.g.*, ASR and TTS) datasets. With text modality as a bridge, our **EMOVA** ultimately becomes a unified system capable of understanding and generating multiple modalities in a coherent and integrated manner. In Sec. 4.1, we first explore the following three questions:

1. *Does the integration of the speech modality conflict with the vision-language capabilities?*
2. *Is sequential alignment of multiple modalities optimal?*



Figure 4. **Demonstration of EMOVA omni-modal instruction tuning.** 1) To support emotional spoken dialogues, **EMOVA** is trained to explicitly select speech style labels with output speech units. 2) For ease of parsing, the data elements are organized in the JSON format.

### 3. How to represent speech modality to foster omni-modal alignment?

We then discuss the omni-modal instruction tuning pipeline and the overall training paradigm of **EMOVA** in Sec. 4.2 and Sec. 4.3, respectively.

#### 4.1. Omni-modal Text-Centric Alignment

**Settings.** We consider the following omni-modal training paradigms: 1)  $VL \rightarrow \text{Speech}$  conducts the image-text alignment first followed by speech-unit-text alignment using the full speech data with 10% of the image-text alignment data to prevent catastrophic forgetting, similarly with [10, 19]. 2)  $\text{Speech} \rightarrow VL$  instead performs speech-unit-text alignment first and then aligns images with texts using 10% of the speech unit-text data and full image-text data. 3)  $\text{Joint}$  aligns both modalities simultaneously. Note that unless otherwise specified, we utilize the S2U tokenizer introduced in Sec. 3.2 to extract speech units for all speech data, which effectively disentangles the semantic and acoustic features. 4)  $\text{Joint-entangled}$  derives the speech units using HuBERT [30], which does not achieve semantic-acoustic disentanglement effectively with only K-means clustering. 5)  $VL$  and  $\text{Speech}$  only align vision and speech modalities with texts, respectively, serving as bi-modal baselines (see Appendix B.1 for more details).

**Evaluation.** For speech abilities, we evaluate the aligned model’s performance on the ASR task of LibriSpeech [72], while for the vision-language, we fine-tune the model with a small amount of high-quality visual instruction data (*i.e.*,

665K SFT data from the ShareGPT4V [9]) and evaluate the fine-tuned model on common vision-language benchmarks. Check Appendix C for evaluation details. Fig. 3 shows the comparison among different paradigms on vision-language (left and middle) and ASR (right, where we report the 1 – WER value for better readability) benchmarks, from which we can derive the following observations:

**Observation 1: image-text and speech-unit-text data benefit each other.** Contrary to the common assumption that multiple modalities might compete and create conflicts, we notice that introducing additional modalities is actually beneficial. As in Fig. 3,  $\text{Joint}$  consistently surpasses both  $VL$  and  $\text{Speech}$  across vision-language and speech benchmarks. Moreover, even models aligned sequentially, (*i.e.*,  $VL \rightarrow \text{Speech}$  and  $\text{Speech} \rightarrow VL$ , which are typically prone to catastrophic forgetting, demonstrate superior performance on most vision-language tasks. We speculate that the requirement to align multiple modalities with text leads to more robust representations, which in turn generalize better across different downstream tasks. This finding aligns with ImageBind [25], where joint alignment of audio and depth with images results in improved performance.

**Observation 2: semantic-acoustic disentanglement benefits omni-modal alignment.** We find that 1)  $\text{Joint}$  outperforms  $\text{Joint-entangled}$  on vision-language benchmarks, and 2) in the speech tasks,  $\text{Joint}$  maintains significant advantages over its entangled counterpart. This can be attributed to the semantic-acoustic disentanglement which makes speech units more analogous to languages.

Benchmarks	EMOVA 3B	EMOVA 7B	EMOVA 72B	Gemini Pro 1.5	GPT-4V	GPT-4o	Whisper Large	Mini-Omni2	VITA 8x7B	VITA 1.5	Baichuan-Omni-7B
MME	2175	2317	<b>2402</b>	-	1927	2310	-	-	2097	2311	2187
MMBench	79.2	83.0	<b>86.4</b>	-	75.0	83.4	-	-	71.8	76.6	76.2
SEED-Image	74.9	75.5	76.6	-	71.6	<b>77.1</b>	-	-	72.6*	74.2	74.1
MM-Vet	57.3	59.4	64.8	-	<b>67.7</b>	-	-	-	41.6	51.1	65.4
RealWorldQA	62.6	67.5	71.0	68.7	61.4	<b>75.4</b>	-	-	59.0*	66.8	62.6
TextVQA	77.2	78.0	<b>81.4</b>	73.5	77.4	-	-	-	71.8*	74.9	74.3
ChartQA	81.5	84.9	<b>88.7</b>	81.3	78.5	85.7	-	-	76.6*	79.6	79.6
DocVQA (test)	93.5	94.2	<b>95.9</b>	86.5	88.4	92.8	-	-	-	-	-
InfoVQA (test)	71.2	75.1	<b>83.2</b>	72.7	-	-	-	-	-	-	-
OCRBench	803	814	<b>843</b>	-	656	736	-	-	678	752	700
AI2D	78.6	81.7	<b>85.8</b>	80.3	78.2	84.6	-	-	73.1	79.3	-
ScienceQA-Img	92.7	96.4	<b>98.2</b>	-	75.7	-	-	-	-	-	-
MMMU	45.8	49.8	59.7	58.5	56.8	<b>69.2</b>	-	-	47.3	52.1	47.3
MathVista	62.6	65.5	<b>69.9</b>	52.1	49.9	63.8	-	-	44.9	66.2	51.9
Mathverse	31.4	40.9	<b>50.0</b>	-	33.6	-	-	-	-	-	-
Librispeech (WER↓)	5.4	4.1	<b>2.9</b>	-	-	-	3.0	4.8	3.4	8.1	-

Table 2. **Comparison on vision-language and speech benchmarks.** 1) **EMOVA** outperforms GPT-4o/4V and Gemini Pro 1.5 on 11 of the 15 vision-language benchmarks, providing a powerful open-sourced alternative. 2) Meanwhile, our **EMOVA** achieves state-of-the-art performance on Librispeech, surpassing its speech and omni-modal counterparts significantly. \*: reported by [43].

### Observation 3: sequential alignment is not optimal.

We notice that `Joint` consistently outperforms its sequential counterparts (*i.e.*, `VL→Speech` and `Speech→VL`) on both vision-language and speech benchmarks, probably due to catastrophic forgetting when integrating a new modality.

In light of these observations, we have chosen to pursue the ultimate alignment strategy that simultaneously aligns image-text and speech-unit-text for **EMOVA**, which offers two important benefits, 1) it fosters the mutual enhancement among vision-language and speech, and 2) it avoids catastrophic forgetting during sequential alignment.

## 4.2. Omni-modal Instruction Tuning

After the omni-modal text-centric alignment in Sec. 4.1, the model learns the fundamental vision-language (*e.g.*, captioning) and speech capabilities (*e.g.*, ASR and TTS). However, instruction tuning is essential to better follow complicated user instructions and respond with vivid emotions.

**Emotion-enriched instruction data synthesis.** Due to the scarcity of omni-modal instruction data (*i.e.*, dialogues involving images, speeches, and texts simultaneously), we opt for synthesizing omni-modal instruction data from existing text and visual instruction datasets. First, we select instruction data suitable for the vocal expression by filtering out the non-vocal data (*e.g.*, code and mathematical formulas). Second, we clean the selected data to be more vocal by removing text formatting elements (*e.g.*, `**` and `\n\n`). We then obtain style labels for the remaining dialog contexts, including genders (`male`, `female`), pitches (`normal`, `low`, `high`), and emotions (`happy`, `sad`, `angry`,

`neutral`), resulting in totally 24 different speech styles. The style labels are generated by prompting GPT-4o<sup>1</sup> to make reasonable inferences given the dialogue context. Finally, we convert the textual instructions and responses into speeches utilizing the latest TTS tools (*i.e.*, `CosyVoice` [16] and `Azure AI Speech`), and the style labels are used to control the style of synthesized speech data. To further improve the diversity of the data, each instruction is synthesized by randomly choosing one of the 39 available speakers. Finally, we gather 120K speech-text and 110K speech-image data pairs. Check more details in Appendix B.2.

**Data organization and the chain of modality.** The omni-modal instruction data can be represented as  $D_{\text{omni}} = \{(x_V, u_S, x_T^o, c_{\text{style}}^o, u_S^o)\}_{i=1}^N$ , where the input consists of the optional queried image  $x_V$  and the speech units of the instruction  $u_S$ , while the output consists of the textual response  $x_T^o$ , the predicted speech style labels  $c_{\text{style}}^o$ , and the output speech unit  $u_S^o$ . Note that we train **EMOVA** to explicitly select styles (*e.g.*, emotions and pitches), which are utilized to determine the corresponding style embedding for the U2S detokenizer (Sec. 3.2). Furthermore, since directly generating speech responses is challenging, we decompose the speech response procedure into three primary steps: 1) recognizing user instructions into texts; 2) generating textual responses based on the recognized instructions; 3) generating the style labels and response speech units based on the textual responses. For ease of parsing during deployment, the target outputs are formatted as JSON, as in Fig. 4.

<sup>1</sup><https://chatgpt.ust.hk>

Datasets	End-to-end $\uparrow$	Text response		Style Categorization		Recognition/Synthesis	
		Unit In	Text In	Emotion	Pitch	WER/CER $\downarrow$	TTS-WER/CER $\downarrow$
Speech-Image-EN	7.45	7.56	7.95	82.50	97.70	2.40	3.20
Speech-Text-EN	6.85	6.90	7.38	81.20	84.70	6.90	2.90
Speech-Image-ZH	6.48	7.02	6.82	77.60	95.90	1.70	12.00
Speech-Text-ZH	5.25	5.58	6.60	80.90	93.20	10.70	12.20

Table 3. **Evaluation of EMOVA-7B on Speech Dialogue.** By default, we evaluate on the corresponding test set of the evaluated datasets.

### 4.3. Overall Training Paradigm

Inspired by [9], a three-stage training paradigm is adopted,

- **Stage-1: Vision-language pre-alignment.** The purpose is to align the visual features into the embedding space of LLMs. Only the vision projector  $p(\cdot)$  is trained.
- **Stage-2: Omni-modal text-centric alignment.** This stage performs vision-language and speech-language alignment jointly. We train the LLM  $f(\cdot)$ , vision projector  $p(\cdot)$ , and the deeper half of vision encoder  $v(\cdot)$  layers.
- **Stage-3: Omni-modal instruction tuning.** We organize different datasets with various types of instructions to learn generalization across tasks, as detailed in Sec. 5.1.

## 5. Experiments

### 5.1. Training configuration

**Stage-1.** In this stage, we only train the parameters of the vision projector  $p(\cdot)$  for vision-language pre-alignment with the LCS-558K dataset [49], with the high-resolution image-slicing strategy [48] adopted.

**Stage-2.** We assemble a unified dataset with 7.4M samples for both the image-text and speech-text alignment, as summarized in Fig. 9. Specifically, we utilize pre-training datasets from ShareGPT4V [9], ALLaVA [4] (both the original English version and the Chinese version translated on our own), and ShareGPT-4o [12] for general perception, while for the OCR capabilities, we leverage SynthDog [34], MMC-Alignment [47], K12 Printing, and the UReader Text Reading subset [89]. Moreover, we adopt the 2,000 hours of ASR and TTS data from LibriSpeech [72] and AISHELL-2 [15] for speech-text alignment, and to preserve the language capabilities of LLMs, we further incorporate the text-only data from Magpie Pro [88]. Check more details in Fig. 9.

**Stage-3.** We collect the EMOVA-SFT dataset consisting of 4.4M multi-task omni-modal samples (see Fig. 8). We start by gathering high-quality open-sourced visual instruction datasets, including ShareGPT4V [9], InternVL [10], Meteor [38], Idetics-2 [36], Cambrian [80], and LLaVA-Onevision [40], followed by quality checking, re-formatting all data samples with a unified template, and removing the duplicated data. For speech, we include the training split of EMOVA omni-model instruction data (*cf.*, Sec. 4.2), with 10% of speech alignment datasets to maintain ASR and TTS performance. We train with 128 Ascend 910B (64GB) NPUs in parallel (check more details in Table 5).

### 5.2. Comparison to SOTA Models

Experimental results are provided in Table 2. We compare a wide range of state-of-the-art VLLMs, including Gemini Pro 1.5 [75], GPT-4V [69], GPT-4o [70], together with the Speech LLM (*i.e.*, Mini-Omni2 [87]) together with the ASR expert Whisper-Large [74], and the omni-modal LLMs (*i.e.*, VITA-8x7B [19], VITA-1.5 [20] and Baichuan-Omni [43]).

**Comparison with SOTA VLLMs.** As an omni-modal model, EMOVA obtains comparable performance with the state-of-the-art VLLMs on multiple vision-language benchmarks, while showing superior proficiency in solving math problems needing precise visual content interpretation. Our EMOVA-7B surpasses GPT-4V by +7.3 on MathVerse, and our EMOVA-72B exceeds GPT-4o by +6.1 on MathVista. On **11 out of the 15** benchmarks, our EMOVA outperforms both GPT-4o/4V and Gemini Pro 1.5, providing a powerful open-sourced alternative.

**Comparison with SOTA omni-modal LLMs.** Under the similar 7B capacity, EMOVA-7B is 220 points higher than VITA on MME, surpassing VITA by 20.6% on OCRBench (814 vs. 678). Moreover, it surpasses Baichuan-Omni-7B, a more recent Omni-modal LLM, on nearly all the evaluated benchmarks, underscoring the effectiveness of our approach and potential to push the boundary of omni-modal abilities.

Moreover, ASR results show that EMOVA significantly outperforms the most recent omni-modal model VITA, even surpassing its SLLM counterparts Mini-Omni2 and ASR-specialized Whisper-Large, showing the effectiveness of the semantic-acoustic disentanglement and omni-modal mutual benefits. We further report TTS results in Table 6, and provide the qualitative examples in Fig. 5 and Appendix G. **For the first time**, EMOVA obtains state-of-the-art results on both the speech and vision-language benchmarks.

### 5.3. Evaluation of Emotion-Rich Spoken Dialogue

In this section, we evaluate the end-to-end spoken dialogue abilities of EMOVA-7B. As stated in Sec. 4.2, the model takes an input image  $x_V$  and user instructions in the form of speech units  $u_S$ , and outputs *text responses*, *style labels*, and *corresponding speech units*. To ensure comprehensive evaluation, we propose the following evaluation metrics:

1. **End-to-end spoken dialogue score** assesses the model’s dialogue performance based on the generated speeches, with a score ranging from 0 to 10, reporting the average.

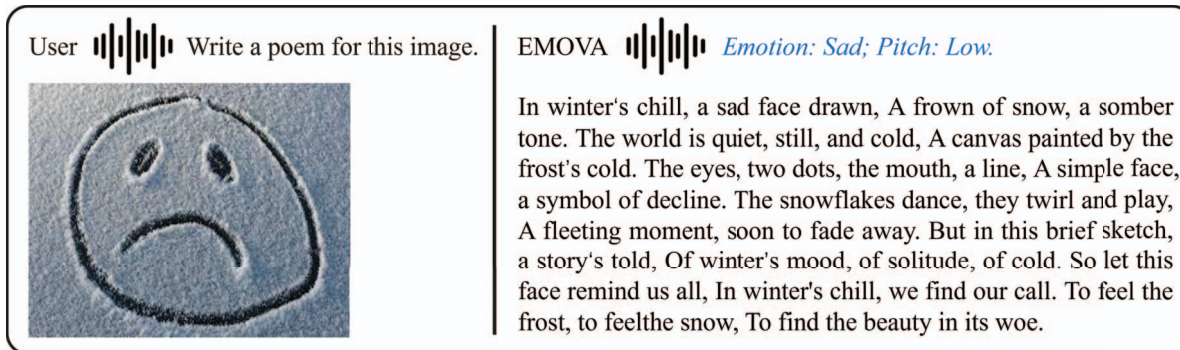


Figure 5. EMOVA engages in omni-modal emotional spoken dialogue expressing sadness.

2. **Unit-input-text-output score** focuses on the quality of the textual responses of LLM when the inputs are speech units, bypassing errors from speech synthesis.
3. **Text-input-text-output score** inputs the ground-truth user instruction texts and evaluates the model's text outputs. This helps disentangle the impact of speech recognition errors and eliminates the effect of JSON format.
4. **ASR and TTS** evaluate how accurately the model recognizes the speech units and how effectively it generates speech units from text. See Appendix D for more details.
5. **Style label classification accuracy** evaluates the accuracy in selecting the appropriate speech style labels.
6. **Style controllability** assesses the controllability of U2S detokenizer with the given conditional style labels using the confusion matrix comparing the generated and recognized style labels. See Appendix D for more details.

Due to the lack of emotionally rich spoken dialogue evaluation datasets, we split a test set from our synthesized omni-modal instruction-tuning data (Sec. 4.1). GPT-4o is used for automated evaluation. See details in Appendix D.

**Results.** Table 3 the spoken dialogue performance.

(i) By comparing the *end-to-end dialogue score* with the *unit-input-text-output score*, we notice that the two scores are closely aligned, with a maximum gap of only 0.33, except for Speech-Image-ZH. TTS-WER/CER is generally low for English, revealing that EMOVA can synthesize accurate speech based on textual responses, which, however, is harder for Chinese, which we attribute to its complexity. It includes tasks such as generating poetries and answering riddles, resulting in more intricate responses.

(ii) Comparing the *unit-input-text-output score* with the *text-input-text-output score*, we notice that their differences correlate with the ASR performance of the speech instructions, especially for Speech-Text-EN and Speech-Text-ZH, which involve more complex instructions.

Our EMOVA-7B reports inferior ASR performance (6.9 and 10.7, respectively) compared to other datasets (2.4 and 1.7). Consequently, when we replace speech instructions with ground-truth transcriptions, EMOVA shows significant improvements from *unit-input* to *text input* score. On

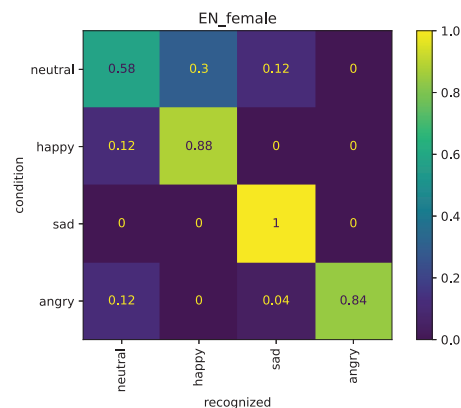


Figure 6. Confusion matrix between the generated and recognized emotions. The emotions generated by our U2S detokenizer are recognized with high probability. Best view with zooming in.

the contrary, for datasets with accurate ASR performance, the results are quite similar, suggesting EMOVA retains robust dialogue capabilities when using the JSON format.

(iii) Examining the *classification accuracy of style labels*, we find that EMOVA performs satisfactorily in classifying emotions and pitches during speech conversations, achieving an accuracy of over 75%. The confusion matrix comparing the conditional and recognized emotion labels is shown in Fig. 6. The results indicate that the four emotions are recognized with high probabilities, with three achieving over 80% accuracy. This demonstrates that our U2S detokenizer effectively controls common emotions, endowing the synthesized speech with vivid emotional expression.

## 6. Conclusion

Our work builds EMOVA, a novel end-to-end omni-modal large language model that effectively aligns vision, speech, and text simultaneously. With text as a bridge, we show that omni-modal alignment is achievable without relying on omni-modal image-text-speech data, meanwhile, enhancing both vision-language and speech abilities. For the first time, EMOVA achieves state-of-the-art performance on both vision-language and speech benchmarks, setting a new standard for versatile omni-modal interactions.

**Acknowledgments.** We gratefully acknowledge supports of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. This work has been made possible by a Research Impact Fund project (RIF R6003-21) and a General Research Fund project (GRF 16203224) funded by the Research Grants Council (RGC) of the Hong Kong Government. This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region (Grants C7004-22G-1 and 16202523) and the Joint Centre for Artificial Intelligence (Grant FB453). This work is supported by National Key Research and Development Program of China (2024YFE0203100) and National Natural Science Foundation of China (No. 62441615, 62201484 and 62136005).

## References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2
- [2] 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*, 2017. 4, 13, 14
- [3] 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. 13
- [4] 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. *arXiv preprint arXiv:2402.11684*, 2024. 7
- [5] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*, 2021. 2
- [6] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*, 2023. 2
- [7] Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, Dit-Yan Yeung, Lifeng Shang, Xin Jiang, and Qun Liu. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*, 2023. 2
- [8] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*, 2023. 18
- [9] 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. 5, 7, 14, 18
- [10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1, 2, 3, 4, 5, 7
- [11] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhi-fang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. 1, 2, 3
- [12] Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. ShareGPT-4o: Comprehensive multimodal annotations with GPT-4o, 2023. 7
- [13] 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. <http://kyutai.org/Moshi.pdf>, 2024. 18
- [14] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 2
- [15] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018. 7, 17
- [16] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024. 4, 6, 14
- [17] 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. 2, 3
- [18] 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*, 2024. 17
- [19] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. VITA: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 1, 2, 3, 4, 5, 7, 18
- [20] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Hetting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 2, 7
- [21] Philip Gage. A new algorithm for data compression. *The C Users Journal*, 1994. 3
- [22] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view

- generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 18
- [23] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024.
- [24] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024. 18
- [25] 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*, 2023. 5
- [26] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023. 2
- [27] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *arXiv preprint arXiv:2403.09572*, 2024. 2
- [28] Yunhao Gou, Hansi Yang, Zhili Liu, Kai Chen, Yihan Zeng, Lanqing Hong, Zhenguo Li, Qun Liu, James T Kwok, and Yu Zhang. Corrupted but not broken: Rethinking the impact of corrupted data in visual instruction tuning. *arXiv preprint arXiv:2502.12635*, 2025. 18
- [29] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 18
- [30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *TASLP*, 2021. 5
- [31] Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu, Lu Hou, Wei Zhang, and Xiaodan Liang. Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. *arXiv preprint arXiv:2407.08706*, 2024. 2
- [32] Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu. SPIRAL: Self-supervised perturbation-invariant representation learning for speech pre-training. *arXiv preprint arXiv:2201.10207*, 2022. 3
- [33] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 17
- [34] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoon Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, 2022. 7
- [35] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*, 2021. 4
- [36] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 2, 7
- [37] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021. 18
- [38] Byung-Kwan Lee, Chae Won Kim, Beomchan Park, and Yong Man Ro. Meteor: Mamba-based traversal of rationale for large language and vision models. *arXiv preprint arXiv:2405.15574*, 2024. 7
- [39] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 17
- [40] 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, 7
- [41] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 18
- [42] Pengxiang Li, Zhili Liu, Kai Chen, Lanqing Hong, Yunzhi Zhuge, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Multi-object tracking data generation via diffusion models. *arXiv preprint arXiv:2312.00651*, 2023. 18
- [43] Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024. 2, 6, 7
- [44] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 18
- [45] 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. 2
- [46] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphs, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 2
- [47] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 7
- [48] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. <https://>

- [//llava-vl.github.io/blog/2024-01-30-llava-next/](https://llava-vl.github.io/blog/2024-01-30-llava-next/), 2024. 2, 7, 18
- [49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2, 7, 18
- [50] 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. 17
- [51] 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. 16
- [52] Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*, 2022. 18
- [53] Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. *arXiv preprint arXiv:2310.05873*, 2023. 18
- [54] Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*, 2024. 18
- [55] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025. 2
- [56] Pan Lu, Swaroop Mishra, Tony 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. In *NeurIPS*, 2022. 17
- [57] 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. In *ICLR*, 2024. 17
- [58] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 2
- [59] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023. 4, 14, 17
- [60] Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. Language model can listen while speaking. *arXiv preprint arXiv:2408.02622*, 2024. 18
- [61] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 16
- [62] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 16
- [63] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 16
- [64] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 3
- [65] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 4
- [66] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *ICML*, 2021. 14
- [67] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 4
- [68] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. In *TACL*, 2023. 18
- [69] OpenAI. GPT-4V. <https://openai.com/index/gpt-4v-system-card/>, 2023. 7
- [70] OpenAI. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. 1, 7
- [71] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [72] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, 2015. 5, 7, 13, 17
- [73] 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
- [74] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023. 1, 3, 7, 18
- [75] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 7
- [76] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 16

- [77] Dehua Tao, Daxin Tan, Yu Ting Yeung, Xiao Chen, and Tan Lee. ToneUnit: A speech discretization approach for tonal language speech synthesis. *arXiv preprint arXiv:2406.08989*, 2024. [2](#), [3](#), [13](#)
- [78] Qwen team. Qwen2-vl. 2024. [3](#)
- [79] Qwen Team. Qwen2.5: A party of foundation models, 2024. [3](#)
- [80] 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. [2](#), [7](#)
- [81] 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. [2](#)
- [82] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. [19](#)
- [83] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, and Kai Zhang. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *arXiv preprint arXiv:2403.13304*, 2024. [18](#)
- [84] Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. On decoder-only architecture for speech-to-text and large language model integration. In *IEEE ASRU*, 2023. [3](#)
- [85] Junjie Wu, Tsz Ting Chung, Kai Chen, and Dit-Yan Yeung. Unified triplet-level hallucination evaluation for large vision-language models. *arXiv preprint arXiv:2410.23114*, 2024. [18](#)
- [86] xAI. Grok, 2024. [17](#)
- [87] Zhifei Xie and Changqiao Wu. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024. [1](#), [2](#), [3](#), [7](#), [17](#), [18](#)
- [88] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024. [7](#)
- [89] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. [7](#)
- [90] 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. In *ICML*, 2024. [17](#)
- [91] 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 *CVPR*, 2024. [17](#)
- [92] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019. [4](#), [14](#)
- [93] 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. [2](#), [3](#), [13](#), [18](#)
- [94] 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. [13](#)
- [95] 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. [3](#), [13](#), [18](#)
- [96] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. [17](#)
- [97] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speeche tokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023. [13](#)
- [98] LIU Zhili, Kai Chen, Jianhua Han, HONG Lanqing, Hang Xu, Zhenguo Li, and James Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023. [18](#)
- [99] Jian Zhu, Cong Zhang, and David Jurgens. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP*, 2022. [13](#)