OpenOmni: Advancing Open-Source Omnimodal Large Language Models with Progressive Multimodal Alignment and Real-Time Self-Aware Emotional Speech Synthesis

Anonymous ACL submission

Abstract

Recent advancements in omnimodal learning have significantly improved understanding and generation across images, text, and speech, yet these developments remain predominantly confined to proprietary models. The lack of high-quality omnimodal datasets and the challenges of real-time emotional speech synthesis have notably hindered progress in open-source research. To address these limitations, we introduce OpenOmni, a two-stage training framework that integrates omnimodal alignment and speech generation to develop a state-of-the-art omnimodal large language model. In the alignment phase, a pre-trained speech model undergoes further training on text-image tasks, enabling (near) zero-shot generalization from vision to speech, outperforming models trained on tri-modal datasets. In the speech generation phase, a lightweight decoder is trained on speech tasks with direct preference optimization, enabling real-time emotional speech synthesis with high fidelity. Experiments show that OpenOmni surpasses state-of-the-art models across omnimodal, vision-language, and speech-language benchmarks. It achieves a 4-point absolute improvement on OmniBench over the leading open-source model VITA, despite using 5× fewer training samples and a smaller model size (7B vs. 7×8B). Additionally, OpenOmni achieves real-time speech generation with <1s latency at non-autoregressive mode, reducing inference time by 5× compared to autoregressive methods, and improves emotion classification accuracy by 7.7% [1].

1 Introduction

The success of large language models (LLMs)(Touvron et al., 2023; Bai et al., 2023a; Tao et al., 2024) has driven rapid advancements in multimodal large language models (MLLMs)(Liu et al., 2024b,a; Luo et al., 2024a; Zhang et al., 2023a; Fang et al., 2024), particularly in vision-language models (VLMs)(Liu et al., 2024b,a; Bai et al., 2023b; Luo et al., 2024a) and speech-language models (SLMs)(Chu et al., 2023; Fang et al., 2024). These innovations mark a paradigm shift in machine understanding and human-computer interaction, fueling interest in omnimodal large language models (OLLMs)—models that integrate vision, language, and speech into a unified system. The emergence of GPT-4o underscores the potential of holistic multimodal AI, yet open-source alternatives remain significantly behind.

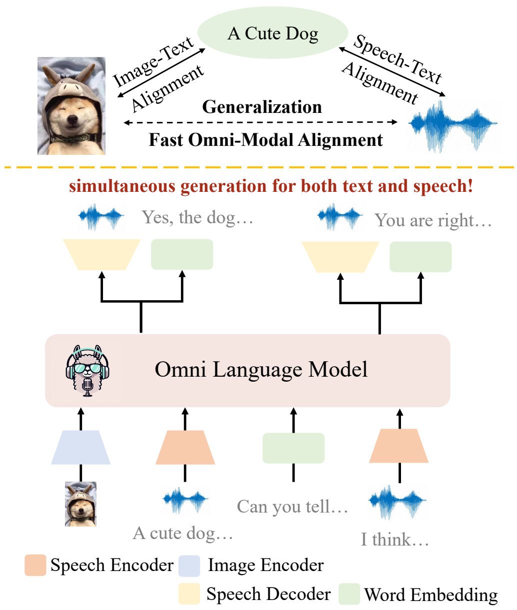

Figure 1: Overview of the motivation and architecture of OpenOmni. For simplicity, our core architecture is presented without the connectors between modules.

Despite their promise, existing open-source OLLMs (Zhan et al., 2024; Sun et al., 2024; Fu et al., 2024; Chen et al., 2024a) face three fundamental challenges, limiting their performance in real-world applications. First, training fully end-to-end OLLMs requires high-quality tri-modal datasets (text, images, and speech), which are scarce, expensive, and difficult to curate at scale.

[1]Code, dataset, and demo are available at https://anonymous.4open.science/r/OpenOmni-1544.

Most open-source models rely on true tri-modal corpora and ignore pairwise datasets (e.g., text-image or text-speech), resulting in suboptimal cross-modal alignment and weaker generalization. Without effective zero-shot alignment strategies, these models struggle to transfer learned representations across modalities, reducing their robustness in real-world multimodal tasks.

Second, existing models predominantly rely on autoregressive (AR) architectures, which generate outputs sequentially, introducing high inference latency that hinders real-time multimodal interaction. Speech generation, in particular, is slow, as most models integrate external text-to-speech (TTS) modules (Du et al.), resulting in latency overhead and preventing end-to-end optimization. Achieving low-latency multimodal synthesis is essential for applications such as conversational AI, assistive technologies, and real-time interactive agents, where response time directly affects usability.

Finally, emotionally expressive speech is critical for natural and engaging human-computer interactions, yet current OLLMs fail to generate emotionally consistent responses. Most models lack self-awareness, producing flat, robotic speech that does not modulate prosody, tone, or sentiment based on conversational context. Without direct preference optimization (DPO) for emotional speech, existing models struggle to align speech intonation with user emotions, leading to inauthentic and disconnected interactions. These challenges significantly constrain the real-world applicability of open-source OLLMs, leaving commercial models far ahead in omnimodal reasoning, real-time interaction, and expressive speech synthesis.

To bridge this gap, we propose OpenOmni, a fully open-source two-stage training framework that enables efficient omnimodal learning while addressing the key limitations of existing models. As illustrated in fig. 1, OpenOmni introduces a progressive alignment strategy that enables cross-modal generalization from vision-language tasks to speech-language tasks, eliminating the need for expensive tri-modal datasets and computing resources. It further incorporates a lightweight, end-to-end speech decoder that facilitates parallel text and speech generation, drastically reducing inference latency compared to autoregressive models. Moreover, by leveraging direct preference optimization (DPO), our model generates emotionally coherent, context-aware speech without requiring additional control modules or handcrafted prompts.

Extensive experiments confirm that OpenOmni achieves state-of-the-art performance in omnimodal alignment, real-time speech synthesis, and emotional speech generation. Compared to VITA(Fu et al., 2024), the leading fully open-source OLLM, which employs a 7×8B language model trained on 5M samples, OpenOmni attains superior results with a smaller model size (7B vs. 7×8B) and three times fewer training samples (1.6M vs. 5M) while outperforming VITA by four absolute points on the OmniBench benchmark(Li et al., 2024b). Additionally, OpenOmni reduces speech generation latency by 5×, achieving real-time inference (<1s) and improving emotion classification accuracy by 7.7%.

Our main contributions can be summarized as follows: (1) **High-Quality Speech Datasets:** We construct O2S-300K and EO2S-9K, comprising 8000 hours of bilingual text-synthesized speech, enabling efficient speech generation and emotional preference learning. (2) **Effective Zero-Shot Omnimodal Alignment:** We introduce a scalable, model-agnostic framework that enables low-resource, rapid omnimodal alignment using language as a pivot, followed by speech generation and emotional preference training. This approach allows the rapid development of an advanced all-modal assistant akin to GPT-4o. (3) **End-to-End Omnimodal LLM:** We train an omnimodal language model with integrated text, image, and speech understanding progressively. After speech generation training and emotional preference optimization, OpenOmni naturally generates real-time emotional speech.

2 Related Work

2.1 Vision Language Models

The rapid progress of Vision-Language Models (VLMs) has been driven by the success of Large Language Models (LLMs) and the increasing availability of diverse image-text instruction data(Liu et al., 2024b; Luo et al., 2024b; Hu et al., 2023) sourced from the internet. LLaVA(Liu et al., 2024b) and MiniGPT-4(Zhu et al., 2023) demonstrate strong cross-task generalization by integrating visual encoders with large language models (LLMs) through lightweight connector modules trained on instruction datasets. To further enhance visual perception, LLaVA-NeXT(Liu et al., 2024a) employs dynamic resolution techniques, improv-

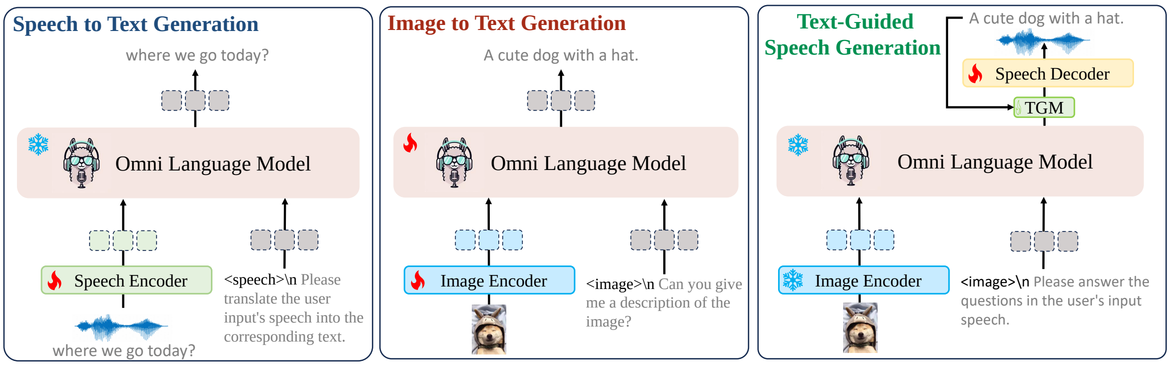

Figure 2: **Progressive Training process of OpenOmni.** To enable zero-shot omnimodal learning and real-time emotional speech generation, OpenOmni undergoes a progressive three-stage training process: (1) **Speech-Text Generation**: A speech encoder extracts continuous speech and text features for alignment learning, equipping the large language model with speech understanding capabilities. (2) **Image-Text Generation**: An image encoder extracts continuous image and text features, facilitating alignment learning that enhances OpenOmni's image comprehension and instruction-following abilities. This process also establishes implicit omnimodal alignment, enabling omni-understanding. (3) **Speech Generation**: A lightweight speech decoder is trained using high-quality synthesized speech dialogue data, with a focus on direct preference optimization for emotional speech. This final stage allows OpenOmni to generate real-time, self-aware emotional speech. A text-guided module (TGM) is utilized to accelerate the training convergence.

ing adaptability to images of varying sizes and complexities. Expanding beyond conventional approaches, DreamLLM(Dong et al., 2023) explores interleaved generation, enabling the simultaneous production of images and text within a shared multimodal context. Meanwhile, DEEM(Luo et al., 2024a) enhances model robustness by employing diffusion models to extract visual features, replacing traditional visual encoders and simplifying the overall architecture. These innovations collectively contribute to advancing vision-language reasoning in multimodal systems.

2.2 Speech Language Models

Recent advancements in Speech-Language Models (SLMs) have significantly improved human-computer interactions by enabling direct speech processing without relying on intermediate text transcription. SpeechGPT(Zhang et al., 2023a) and LLaMA-Omni(Fang et al., 2024) eliminate the need for explicit text-based transcriptions, reducing latency in multimodal content generation. For full-duplex dialogue systems, Moshi(Défossez et al.) and OmniFlatten(Zhang et al., 2024a) introduce mechanisms for handling simultaneous speech and text streams, adeptly managing challenges such as overlapping speech and interruptions(Lin et al., 2022). Meanwhile, Freeze-Omni(Wang et al., 2024) introduces an innovative training approach that preserves the core capabilities of the original LLM, allowing low-latency speech-to-speech dialogue without requiring modifications to the pre-trained architecture. Focusing on emotional speech synthesis, Emo-DPO (Gao et al., 2024) applies direct preference optimization (DPO) to generate expressive and controllable emotional speech, addressing the emotional coherence gap in existing speech-language models. These developments mark a significant shift towards more natural, real-time speech interactions in multimodal AI systems.

2.3 Omni-modal Language Models

As multimodal research advances, models are increasingly shifting towards unified frameworks that seamlessly integrate diverse input and output modalities. By tokenizing different data types into a shared representation, models like AnyGPT(Zhan et al., 2024) and Unified-IO 2(Lu et al., 2024) achieve seamless cross-modal task adaptability, allowing them to process audio, text, and images without significant architectural modifications. More recently, Mini-Omni2(Xie and Wu, 2024) extends multimodal capabilities by integrating visual and auditory encoders, enabling real-time multimodal responses while incorporating mechanisms for detecting and interpreting semantic interruptions. Meanwhile, video-SALMONN(Sun et al., 2024) enhances video understanding by incorporating fine-grained temporal modeling, improving the model's ability to interpret speech and actions within videos. To enhance human-computer interaction, VITA(Fu et al., 2024) introduces duplex communication schemes, enabling fluid and intuitive exchanges between users and AI models. EMOVA(Chen et al., 2024a) further extends the expressive capabilities of multimodal systems by integrating controllable emotional speech synthesis, providing more natural and engaging user interactions. Building upon these advancements, OpenOmni introduces a novel approach for near zero-shot omnimodal alignment across language, vision, and speech, incorporating self-aware emotional speech synthesis to enhance expressiveness and realism. By optimizing for speed, data efficiency, and generalization, OpenOmni achieves state-of-the-art performance in omnimodal tasks, surpassing previous models in real-time speech generation, multimodal alignment, and emotion-aware synthesis.

3 Method

In this section, we first formulate the omnimodal learning problem and provide an overview of the training procedure of OpenOmni, as shown in fig. 2. We then describe the specific training procedures for omnimodal alignment and real-time speech generation.

3.1 Problem Formulation and Overview

Omnimodal learning aims to model the relationships between images (x^V), speech (x^S), and text (y). The image-to-text generation task, which involves generating textual descriptions for input images encoded by an image encoder $\mathcal{E}_V$, is modeled as learning the conditional distribution $p_\theta(y|\mathcal{E}_V(x^V))$, parameterized by θ. Similarly, the speech-to-text generation task, which generates relevant text responses given input speech encoded by a speech encoder $\mathcal{E}_S$, is formulated as learning $p_\phi(y|\mathcal{E}_S(x^S))$, parameterized by ϕ. Finally, the omnimodal-to-speech generation task, which synthesizes speech responses based on input text, speech, and images, is represented as learning $p_\gamma(x^S|\mathcal{D}_{LLM}(y, \mathcal{E}_S(x^S), \mathcal{E}_V(x^V)))$, parameterized by γ, where $\mathcal{D}_{LLM}$ represents the Large Language Model.

In standard omnimodal learning settings, training typically relies on image-text-speech pairs $\mathcal{D}_o = \{(x_i^V, x_i^S, y_i)\}_{i=1}^K$ (Fu et al., 2024; Li et al., 2024a). However, high-quality image-text-speech datasets are scarce. To mitigate this limitation, we introduce text as a pivot, leveraging large-scale image-text datasets $\mathcal{D}_{i2t} = \{(x_i^V, y_i)\}_{i=1}^M$ (Liu et al., 2024b; Luo et al., 2024b) and text-speech datasets $\mathcal{D}_{s2t} = \{(x_i^S, y_i)\}_{i=1}^N$ (Panayotov et al., 2015; Zhang et al., 2022), where $M \gg K$ and $N \gg K$. Inspired by human learning mechanisms, where individuals naturally align visual concepts with speech across languages, OpenOmni transfers visual concepts learned from text-image tasks to speech understanding.

OpenOmni decomposes the omnimodal alignment process into two consecutive stages: **text-speech alignment** and **image-text alignment**. The text-speech alignment stage establishes cross-modal alignment between speech x^S and language y. This is achieved by training a speech LLM on text-speech pairs $\mathcal{D}_{s2t}$ with the objective $p_\phi(y|\mathcal{E}_S(x^S))$, denoted as f_ϕ. This ensures that the hidden representations of semantically similar text-speech pairs are close, i.e., $f_\phi(y) \approx f_\phi(x^S)$. In the image-text alignment stage, OpenOmni utilizes large-scale image-text datasets $\mathcal{D}_{i2t}$ to optimize the image-to-text objective $p_\theta(y|\mathcal{E}_V(x^V))$. The following sections describe the training process for **omnimodal alignment** and **real-time speech generation**. Notably, OpenOmni is architecture-agnostic, allowing flexible integration with existing state-of-the-art model architectures and training strategies.

3.2 Speech-to-text Generation

In speech-to-text generation, we incorporate a speech encoder $\mathcal{E}_S$ to extract audio features from input speech x^S. These audio features $\mathcal{E}_S(x^S)$ are then replaced with corresponding text y^S as input into the LLM. Following recent work to train speech conversation models (Fang et al., 2024; Chu et al., 2023; Zhang et al., 2023a), we pre-train OpenOmni on a large scale of text-speech pairs using the language modeling objective:

$$\mathcal{L}_{s2t}(p_\phi, \mathcal{D}_{s2t}) = -\sum_{i=1}^{N} \log p_\phi(y_i|\mathcal{E}_S(x_i^S)). \quad (1)$$

3.3 Image-to-text Generation

In image-to-text generation, we incorporate an image encoder module $\mathcal{E}_V$ to provide visual feature $\mathcal{E}_V(x^V)$. These visual features are then concatenated with the text embedding as input into the speech LLM. Following recent work to train image-text conversation models (Liu et al., 2024b; Luo et al., 2024b), OpenOmni's training process for image-to-text generation consists of two sub-stages: Image-Text Pretraining and Image-Text Instruction Tuning.

Image-Text Pretraining In this sub-stage, we pre-train the visual module to align it with LLM on a large scale of image-text pairs using the language modeling objective:

$$\mathcal{L}_{i2t}(p_\theta, \mathcal{D}_{i2t}) = -\sum_{i=1}^{M} \log p_\theta(y_i|\mathcal{E}_V(x_i^V)). \quad (2)$$

Here, we fix the parameters of LLM to prevent short texts in the image-text pairs from influencing the general capabilities.

Image-Text Instruction Tuning To enhance models' capabilities in following human instructions, we conduct instruction tuning on elaborately curated multimodal instruction tuning datasets built by blending the existing image-text instruction tuning datasets. We denote this image-text instruction tuning datasets as $\mathcal{D}_{i2t}^I = \{x_j^V, y_{q,j}, y_{a,j}\}_{j=1}^L$, where y_q is the instructions and y_a is the response. Both the visual module and speech LLM are fine-tuned by maximizing the probability of the response:

$$\mathcal{L}_{i2t}^I(p_\theta, \mathcal{D}_{i2t}^I) = -\sum_{j=1}^{L} \log p_\theta(y_{a,j}|\mathcal{E}_V(x_j^V), \mathcal{D}_{LLM}(y_{q,j})). \quad (3)$$

Interestingly, we observe a *quasi-zero-shot* transfer capability in OpenOmni within this scenario. When instruction tuning is performed exclusively on the image-text dataset, the model demonstrates the ability to respond accurately to an image x^V and either a text-based question y_q or an instruction provided in speech x_q^S. However, its responses are predominantly in text. This behavior can be attributed to the inherent similarity between the hidden representations of textual and spoken instructions learned by the LLM, i.e., $\mathcal{D}_{LLM}(y_q) \approx \mathcal{D}_{LLM}(\mathcal{E}_S(x_q^S))$. Consequently, the model satisfies the following approximation: $p_\theta(y_a|\mathcal{E}_V(x^V), \mathcal{D}_{LLM}(y_q)) \approx p_\theta(y_a|\mathcal{E}_V(x^V), \mathcal{D}_{LLM}(\mathcal{E}_S(x_q^S)))$. OpenOmni completes the progressive omnimodal alignment, enabling the LLM to achieve a comprehensive understanding across image, text, and speech modalities.

3.4 Speech Generation

In speech generation, we incorporate a Speech Decoder $\mathcal{D}_S$ to generate speech based on the output of the LLM $\mathcal{D}_{LLM}$. The speech generation training process in OpenOmni consists of two sub-stages: Real-time Speech Pretraining and Emotional Speech Direct Preference Optimization (DPO).

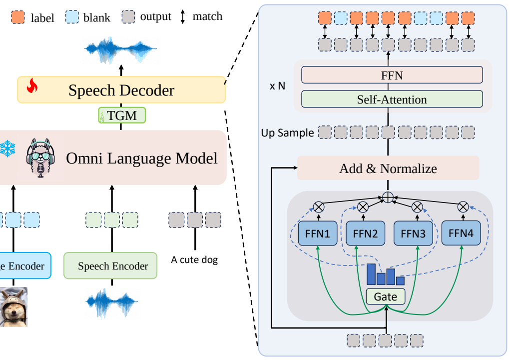

Figure 3: Structure of speech decoder. The speech decoder consists of a mixture of expert module and multiple transformer layers, achieving end-to-end speech unit learning through connectionist temporal classification (CTC) loss.

Real-time Speech Generation To equip OpenOmni with real-time speech generation for enhancing interactive experiences, we adopt a streaming speech decoder, which supports supports both autoregressive (AR) and non-autoregressive (NAR) speech decoding mode. We curate a dataset, termed OpenOmni-300K, consisting of 300K single-round image-text instructions from MMEvol (Luo et al., 2024b) and UltraChat with corresponding speech responses for training the speech decoder. We denote this dataset as $\mathcal{D}_{o2s}^I = \{x_j^V, y_{q,j}, y_{a,j}, x_{a,j}^S\}_{j=1}^L$, where y_q represents the instruction, y_a is the textual response, and x_a^S is the speech response.

To process the speech response x_a^S, we follow (Zhang et al., 2023a; Fang et al., 2024) to discretize speech into discrete units. Specifically, we use the pre-trained speech tokenizer (Zhang et al., 2023a) to extract continuous speech representations and then convert these representations into a single unit, resulting in the final discrete unit sequence $x_a^U = \{x_{a,i}^U\}_{i=1}^L, \quad x_{a,i}^U \in \{0, 1, ..., K-1\}$, where K is the speech vocabulary size, and L is the length of the discrete unit sequence. The discrete units can then be converted back into a waveform using an additional unit-based vocoder $\mathcal{V}$ (Polyak et al., 2021), trained on English and Chinese datasets.

As shown in fig. 3, we integrate a streaming NAR Speech Decoder $\mathcal{D}_S$ after the LLM to simultaneously generate speech and text responses. The NAR Speech Decoder consists of a **mixture of expert (MOE)** layer and a tiny standard decoder-only language model. The MOE layer stabilizes training and accelerates convergence—without this layer, the speech decoder fails to train effectively. Similar to (Ma et al., 2024; Zhang et al., 2024b; Fang et al., 2024), the NAR speech decoder takes the output hidden states from the LLM as input and generates the discrete unit sequence corresponding to the speech response in real-time.

Given that the output hidden states of the text response y_a are denoted as $\mathcal{D}_{LLM}(x^V, y_q) = [\mathbf{h}_1, ..., \mathbf{h}_N]$. We first pass these hidden states through the text-guided module (TGM) to obtain the transformed hidden state $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_N]$. Next, $\mathbf{C}$ is fed into the speech decoder layers, yielding the final hidden state sequence $\mathbf{O} = [\mathbf{o}_1, ..., \mathbf{o}_M]$. We use **connectionist temporal classification (CTC)** (Graves et al., 2006) to align $\mathbf{O}$ with the discrete unit sequence $x_a^U = \{x_{a,i}^U\}_{i=1}^L$. During training, CTC marginalizes over all possible alignments as follows:

$$\mathcal{L}_{ctc}(p_\gamma, \mathcal{D}^I_{o2s}) = -\log p_\gamma(x_a^U|\mathbf{O}) = \\ -\log \sum_{A\in\beta^{-1}(x_a^U)} p_\gamma(A|\mathbf{O}) = -\log \sum_{A\in\beta^{-1}(x_a^U)} \prod_{i=1}^{M} p_\gamma(x_{a,i}^U|\mathbf{O}), \tag{4}$$

where $\beta^{-1}(x_a^U)$ denotes all possible alignments of length M that collapse to x_a^U. During inference, the best alignment is selected as:

$$A^* = \arg\max_A P(A|\mathbf{O}), \tag{5}$$

and the corresponding discrete unit sequence $\beta(A^*)$ is fed into the vocoder to synthesize the waveform.

Self-aware Emotional Speech Generation To enable OpenOmni to generate self-aware, emotionally coherent, and expressive speech based on contextual history without additional control modules, we introduce the **CTC Direct Preference Optimization (DPO)** algorithm. This method enhances smooth and natural dialogue interactions. The DPO approach leverages an analytical reward function $r(x, y)$, expressed as:

$$r(x, y) = \beta \log \frac{\pi_*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x), \tag{6}$$

where β is a constant and $Z(x)$ is the partition function. Using this observation, we directly optimize the policy model based on human feedback preference pairs (y_w, y_l):

$$\mathcal{L}_{dpo} = -\mathbb{E}_{(x,y_w,y_l)}[\log \sigma(r(x, y_w) - r(x, y_l))] \\ = -\mathbb{E}_{(x,y_w,y_l)}[\log \sigma(\beta \log \frac{\pi_*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_*(y_l|x)}{\pi_{\text{ref}}(y_l|x)})], \tag{7}$$

where the reference model $\pi_{\text{ref}}(y|x)$ is the pretrained model from the real-time speech generation stage and remains fixed during DPO training. Only the policy model $\pi_*(y|x)$ is updated. Compared to traditional reinforcement learning with human feedback (RLHF), DPO is simpler, more efficient, and more stable for aligning OpenOmni with self-aware emotional speech generation.

Following the Plutchik Model of Emotions (6seconds.org, 2022), we construct a multi-turn dialogue preference dataset incorporating nine distinct emotions. Each preference pair consists of an emotionally congruent speech response unit sequence $y_w = x_{a,w}^U$, which aligns with the conversational history, and an emotionally neutral sequence $y_l = x_{a,l}^U$, which is inconsistent with the context. The policy model $\pi(y|x)$ during training is optimized as:

$$\log \pi(y|x) = \log \sum_{A\in\beta^{-1}(x_a^U)} \prod_{i=1}^{M} p_\gamma(x_{a,i}^U|\mathbf{O}). \tag{8}$$

After training, OpenOmni is capable of generating real-time, emotionally expressive multi-turn dialogues.

4 Experiments

4.1 Implementation Details

In this subsection, we introduce the model and data construction, and more details about the data and training strategy can be found in appendix D.

Omnimodal Alignment Data During the speech-to-text generation phase, in addition to WeNetSpeech (Zhang et al., 2022), LibriSpeech (Panayotov et al., 2015), and AIShell-4 (Fu et al., 2021), we use portions of shorter responses from O2S-300K, totaling 8000 hours of data, for bilingual speech-text alignment training. For image-text alignment, we train OpenOmni on the LLaVA-Pretrain-595K (Liu et al., 2024b) for image-text alignment. In the image-text instruction tuning stage, we fine-tune OpenOmni on the compact high-quality dataset MMEvol (Luo et al., 2024b) for efficient optimization.

Real-time Speech Generation Data To support real-time speech generation, we curate a dataset of 300K instructions from MMEvol (Luo et al., 2024b) and UltraChat (Ding et al., 2023) that included long responses for training the speech decoder. Specifically, we decompose multi-turn

Method	Action & Activity	Story Description	Plot Inference	Identification & Description	Contextual & Environmental	Identity & Relationship	Text & Symbols	Count & Quantity	Overall
AnyGPT (7B) (Zhan et al., 2024)	5.98	8.70	7.59	4.74	5.67	12.50	8.00	20.00	7.01
Video-SALMONN (13B) (Sun et al., 2024)	28.69	25.65	24.47	23.22	29.08	21.83	**52.00**	26.63	26.53
UnifiedIO2-Large (1.1B) (Lu et al., 2024)	28.29	22.17	<u>32.49</u>	30.81	28.37	21.83	16.00	13.33	27.76
UnifiedIO2-XLarge (3.2B) (Lu et al., 2024)	30.28	26.52	30.38	31.75	28.37	18.75	<u>28.00</u>	26.63	29.16
UnifiedIO2-XXLarge (6.8B) (Lu et al., 2024)	27.49	23.04	28.69	25.59	26.95	12.50	12.00	**46.67**	25.92
VITA (7x8B) (Fu et al., 2024)	<u>33.47</u>	<u>34.35</u>	27.00	<u>36.02</u>	<u>43.97</u>	<u>31.25</u>	24.00	6.67	<u>33.45</u>
OpenOmni (7B)	**36.65**	**45.65**	**32.91**	**44.08**	**48.23**	**34.38**	24.00	<u>33.33</u>	**37.40**

Table 1: Overall omni-understanding results on OmniBench. We present a performance comparison of omni-understanding across various fully open-source Omnimodal Large Language Models (OLLMs) on OmniBench. Notably, compared to the state-of-the-art OLLM, VITA (Fu et al., 2024), which was trained on tri-modal data, OpenOmni achieves comparable performance while utilizing significantly less training data and a smaller model size.

dialogues into single-turn question-answer pairs, rank the responses based on their length, and select 100K question-answer pairs with relatively long responses. To support bilingual output in Chinese and English, we translate 50K question-answer pairs into their corresponding Chinese versions using GPT-4o-mini API, and then convert the answers into the corresponding speech using CosyVoice (Du et al.). We employ the same method for text-conditioned speech synthesis on 200k randomly selected data from UltraChat. As a result, we obtain 8000 hours of high-quality bilingual speech generation data **O2S-300K**.

Self-aware Emotional Speech Generation Data Based on the Plutchik Model of Emotions (6seconds.org, 2022), which categorizes emotions into eight distinct types, we curate a multi-turn speech preference dataset, **EO2S-9K**, for self-awareness emotion evaluation. Specifically, we randomly select 200K samples from MMEvol and employ Qwen2-72B (Bai et al., 2023a) to categorize responses into nine predefined emotions per round. From this, we extract 1K bilingual dialogues labeled with emotion categories, reserving an additional 100 samples as an emotional test set for evaluating self-aware speech generation. Since certain emotions, such as anger and sadness, are underrepresented in the MMEvol dataset, we augment the dataset using the GPT-4o-mini API to ensure sufficient data for these categories. The final dataset maintains an equal representation of Chinese and English samples. To further enhance emotional preference training, we use CosyVoice to generate unconditional speech as negative samples and emotion-conditioned speech as positive samples, constructing preference pairs for training direct preference optimization in emotional speech generation.

Model We design the architecture following LLaVA Series (Liu et al., 2024b,a; Luo et al., 2024b), where the omnimodal large language model consists of four key components: an LLM (Qwen2.5-7B-Instruct (Bai et al., 2023a)) for next token prediction, an image encoder (CLIP-ViT-L (Radford et al., 2021)) for extracting visual features, a speech encoder (Whisper-large-v3 (Radford et al., 2023)) for extracting audio features and a streaming speech decoder (Qwen2.5-0.5B-Instruct (Bai et al., 2023a)) for generating vivid speech in real-time. Moreover, an image-text projector and a speech-text projector are adopted to align the visual-text modalities and the speech-text modalities, respectively. A mixture of expert modules and text-guided modules is designed to align the omnimodal embedding and speech decoder efficiently and stably. For the autoregressive mode, we use the Speech Tokenizer from GLM4-Voice (Zeng et al., 2024) with a vocabulary size of 16K, which results in better speech quality. For non-autoregressive models, we use the CosVoice (Du et al.) Speech Tokenizer with a smaller vocabulary size of 6K, facilitating faster convergence during CTC training.

4.2 Omni-Language Evaluation

OmniBench (Li et al., 2024b) is a pioneering benchmark designed to evaluate omnimodal large language models (OLLMs) by assessing their ability to integrate and interpret simultaneous inputs from images, audio, and text. This evaluation framework consists of 1,142 question-answer pairs categorized into tasks that focus on cognitive and reasoning abilities, posing significant challenges in entity recognition, causal inference, and abstract concept comprehension. We compare OpenOmni with other OLLMs on OmniBench, with results summarized in table 1. Notably, our model achieves excellent zero-shot omnimodal alignment using only two training phases: speech-text alignment and image-text alignment. Compared to the fully open-source state-of-the-art OLLM, VITA (Fu et al.,

Model	Lang	Angry & Disgusted	Fearful	Happy	Neutral	Other	Sad	Surprised	Overall
OpenOmni	ZH	89.7	54.8	33.3	92.3	51.6	60.2	23.7	57.9
w/ DPO	ZH	**96.6**	**78.4**	**37.7**	**97.1**	**62.8**	**90.7**	**29.8**	**70.4**
OpenOmni	EN	89.2	68.7	57.5	91.9	48.0	75.6	7.5	62.6
w/ DPO	EN	**91.3**	**70.4**	**60.6**	**94.6**	**49.6**	**77.3**	**13.9**	**65.4**

Table 2: Overall self-aware emotional speech generation results on the bilingual EO2S-9K test set. Using the emotional speech direct preference optimization algorithm, OpenOmni demonstrates consistent improvements in emotional speech generation for both Chinese and English. The average accuracy of bilingual emotional speech generation increases by 7.7 % (from 60.2% to 67.9%), with particularly notable gains in categories such as Angry, Fearful, and Sad.

Model	AIShell-2(ZH-CER)				Librispeech(EN-WER)			
	Dev		Test		Test_clean		Test_other	
	S2T	T2S	S2T	T2S	S2T	T2S	S2T	T2S
Speech LLM								
SpeechT5 (Ao et al., 2021)	-	-	-	-	2.4	-	5.8	-
SALMONN (Tang et al., 2023)	-	-	-	-	2.1	-	4.9	-
Mini-Omni (Xie and Wu, 2024)	-	-	-	-	4.7	-	9.4	-
Freeze-Omni (Wang et al., 2024)	-	-	-	-	3.2	-	7.7	-
Qwen2-Audio (Chu et al., 2023)	3.1	-	3.3	-	2.0	-	4.5	-
Omnimodal LLM								
AnyGPT (Zhan et al., 2024)	-	-	-	-	8.5	-	-	-
VITA (Fu et al., 2024)	-	-	-	-	8.1	-	18.4	-
EMOVA (Chen et al., 2024a)	10.3	7.9	-	-	4.0	3.4	-	-
VITA 1.5 (Fu et al., 2024)	-	-	-	-	3.4	-	7.5	-
OpenOmni	**6.8**	**7.3**	**6.9**	**13.1**	**3.1**	**2.6**	**4.1**	**5.6**

Table 3: Comparison with state-of-the-art methods on speech-language benchmarks. We mark the best performance **bold**.

2024), OpenOmni achieves superior overall results on OmniBench (37.40 vs. 33.45) despite using significantly fewer training parameters (7B vs. 8×7B) and less image-text training data (1.6M vs. 5M). Furthermore, by leveraging language as a pivot, our approach completes omnimodal alignment implicitly, demonstrating enhanced scalability in scenarios with limited tri-modal data.

4.3 Speech-Language Evaluation

To evaluate OpenOmni 's speech understanding and generation capabilities, we measure word error rate (WER) on the AIshell-2 (Fu et al., 2021) and Librispeech (Panayotov et al., 2015) benchmarks for two tasks: speech-to-text recognition (S2T) and text-to-speech generation (T2S). For T2S evaluation, we use Whisper-large-V3 to transcribe OpenOmni 's synthesized speech and compute WER against ground-truth text labels. As shown in table 3, OpenOmni achieves the best performance on both S2T and T2S tasks for bilingual (Chinese and English) data, outperforming other omnimodal models. This result indicates that OpenOmni not only comprehends speech effectively but also generates fluent, high-quality audio while maintaining strong alignment between speech and language modalities in both languages. Additionally, compared to VITA (Fu et al., 2024), which relies on separate text-to-speech (TTS) models, and EMOVA (Chen et al., 2024a), which uses an autoregressive (AR) structure, OpenOmni demonstrates significantly faster speech generation via two mode support. Owing to its end-to-end, lightweight, non-autoregressive (NAR) decoding mode support, OpenOmni can generate up to 30 seconds of speech with less than one second of latency, achieving real-time speech generation at over five times the speed of autoregressive models.

4.4 Emotional Speech Synthesis Evaluation

To assess the effectiveness of direct preference learning in emotional speech generation, we evaluate OpenOmni 's self-aware emotional speech synthesis on the EO2S-9K test set. Specifically, we use Emotion2Vec (Ma et al., 2023) to classify the emotions in the generated speech and measure accuracy against ground-truth labels. As shown in table 2, direct preference optimization for emotional speech effectively enhances OpenOmni 's ability to generate emotionally expressive speech. This improvement is particularly evident in bilingual, multi-turn emotional speech generation tasks, demonstrating the model's ability to produce natural, contextually aware speech with accurate emotional intonation. Additional experiments can be found in appendix B.

5 Conclusion

In this paper, we introduced OpenOmni, a novel end-to-end omnimodal model that leverages language as a pivot to achieve tri-modal zero-shot alignment, addressing the challenge of limited tri-modal data. By integrating a lightweight streaming speech decoder with direct preference optimization for emotional speech, OpenOmni enabled real-time, self-aware, high-quality speech interactions. Our extensive evaluations demonstrated that OpenOmni achieves state-of-the-art performance on tri-modal benchmarks while using significantly fewer training parameters and less training data than previous models.

6 Limitations

Although OpenOmni demonstrates strong performance under low-resource conditions, the impact of training on a larger volume of high-quality trimodal data remains an open question. Additionally, while the mixture of experts (MoE) module effectively mitigates conflicts in CTC training, it remains more challenging to optimize compared to autoregressive generation methods. Striking a balance between efficiency and stability in non-autoregressive speech generation remains an important direction for future research. To further improve omnimodal interactions, we plan to enhance OpenOmni 's ability to generate multi-character speech responses, improving conversational depth and expressiveness. We also plan to explore reinforcement learning techniques for refining multimodal alignment beyond current supervised learning approaches. As omnimodal AI continues to evolve, we believe OpenOmni represents a step toward more natural and immersive human-computer interactions.

References

6seconds.org. 2022. Plutchik-model-emotion. https://www.6seconds.org/2022/03/13/plutchik-wheel-emotions.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. 2024a. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*.

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

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

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

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.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

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. corr, abs/2407.05407, 2024. doi: 10.48550. *arXiv preprint ARXIV.2407.05407*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

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

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

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

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2024. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. *arXiv preprint arXiv:2409.10157*.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. 2024. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. 2024a. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, et al. 2024b. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Duplex conversation: Towards human-like interaction in spoken dialogue systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3299–3308.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. In *NeurIPS*.

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

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. 2024c. Ii-bench: An image implication understanding benchmark for multimodal large language models. *Preprint*, arXiv:2406.05862.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, pages arXiv–2310.

Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, et al. 2024a. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232*.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. 2024b. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*.

Zhengrui Ma, Qingkai Fang, Shaolei Zhang, Shoutao Guo, Yang Feng, and Min Zhang. 2024. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation. *arXiv preprint arXiv:2406.06937*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

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

IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

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

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

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

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

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

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

x.ai. 2024. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v.

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

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

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

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

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9637–9662, Bangkok, Thailand. Association for Computational Linguistics.

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

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

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

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, and Chaohong Tan. 2024a. Omniflatten: An end-to-end gpt model for seamless voice conversation. *arXiv preprint arXiv:2410.17799*.

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2024b. Streamspeech: Simultaneous speech-to-speech translation with multi-task learning. *arXiv preprint arXiv:2406.03049*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Data Construction

We provide details of the data construction for multiple training stages below:

- **OpenOmni-1-1:** In addition to datasets WeNetSpeech, LibriSpeech, and AIShell-4, we randomly select 80k image-text instruction data with shorter responses from MMEvol (Luo et al., 2024b). We translate 40k of this data into Chinese using Qwen72B and synthesize the responses into speech data with CosVoice. This results in 1600 hours of OpenOmni-1-1 data for speech-text alignment pretraining.

- **OpenOmni-2-1:** For rapid image-text alignment pretraining, we use the llava pretrain dataset, following previous work (Liu et al., 2024b,a; Luo et al., 2024b; Liu et al., 2024c).

- **OpenOmni-2-2:** To achieve efficient image-text instruction tuning, we employ MMEvol data. Since we later train the speech decoder by freezing the LLM mode, we include O2S-300K to stabilize the training of the speech decoder, resulting in a combined dataset of 1.7M for OpenOmni-2-2.

- **OpenOmni-3-1:** To better utilize computational resources, we select 300k data with long response instructions from MMEvol and UltraChat. This includes 100k image-text instruction data, 100k single-round dialogue, and 100k multi-round dialogue. We synthesize the corresponding speech using CosVoice, resulting in 8000 hours of O2S-300K.

- **OpenOmni-3-2:** We curate 9k emotion preference data and generate emotional speech preference pairs using CosVoice's conditional control. This is used for Direct Emotion Preference Optimization.

B Additional Experiments

B.1 Vision-Language Evaluation

To comprehensively assess the effectiveness of OpenOmni in aligning visual-text modalities, we compare its performance against previous vision-language models (VLLMs) across eight representative benchmarks: MMBench-EN (Liu et al., 2023), MMBench-CN (Liu et al., 2023), MM-Star (Chen et al., 2024b), RealWorldQA (x.ai, 2024), MMMU (Yue et al., 2024), MathVista (Lu et al., 2023), AI2D (Kembhavi et al., 2016), and HallusionBench (Guan et al., 2023). To ensure reproducibility and maintain consistency across all models and benchmarks, we employ VLMEvalKit (Duan et al., 2024) for zero-shot evaluation. As shown in table 4, OpenOmni achieves superior results compared to the fully open-source state-of-the-art OLLM, VITA (Fu et al., 2024), despite being trained on significantly less data. Notably, our model outperforms VITA with gains of 7.0% on MMBench-Chinese and 11.3% on HallusionBench. We can also observe that the use of additional speech modals can further enhance the vision-language capabilities of the model. Furthermore, compared to other fully open-source visual-language models, OpenOmni maintains competitive performance despite reduced training data, demonstrating the effectiveness of our image-text alignment strategy.

B.2 Additional Omni-Language Evaluation

In addition to OmniBench, we conduct experiments on the AV-Odyssey Bench (Gong et al., 2024), which involves the four modalities: audio, text, image, and video. For video, we test by averaging 8 sampled frames into a single image. The experimental results are shown in the table 6 below. Compared to other OLLMs, OpenOmni achieves the best average performance using only bi-modal speech-text and image-text data. With 7B model parameters and no audio or video training, it outperforms VITA by 4.4 points, demonstrating the effectiveness and efficiency of OpenOmni.

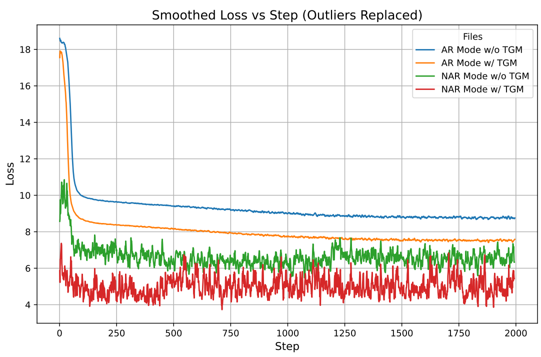

Figure 4: Ablation study of text-guided module (TGM). In order to explore the effect of TGM on speech generation under the two modes, we plot the change of training loss under the same setting. TGM can significantly improve the convergence speed of training and improve the effect of speech generation of the speech decoder.

Model	w/ Audio IO	PT	IT	MMStar	MMB	MMB^{CN}	HallBench	$\text{MathVista}^{\text{M}}$	MMMU^{V}	AI2D	RWQA
Proprietary Models											
GPT-4o	✓	–	–	-	83.4	82.1	55.0	63.8	69.1	-	75.4
GPT-4o-mini	✓	–	–	-	-	-	46.1	52.4	60.0	-	67.1
Weight Open-Source											
MiniCPM-V2.5 (8B) (Yao et al., 2024)	✗	570M	9.1M	51.3	76.7	73.3	42.5	54.3	45.8	-	63.5
Qwen2-VL-Chat (7B) (Bai et al., 2023b)	✗	1.4B	-	60.7	86.4	81.9	50.6	58.2	52.7	-	69.7
Baichuan-Omni (7B) (Li et al., 2024a)	✓	–	8M	-	76.2	74.9	47.8	51.9	47.3	-	62.6
EMOVA (8B) (Chen et al., 2024a)	✓	7.4M	4.4M	-	82.8	-	-	61.1	-	82.8	64.3
Fully Open-Source											
Cambrain-I (8B) (Tong et al., 2024)	✗	2.5M	7M	50.7	-	-	34.3	47.0	41.8	73.1	64.2
MMEvol (7B) (Luo et al., 2024b)	✗	0.6M	1.5M	51.6	74.6	74.3	42.9	52.4	45.1	74.7	63.9
VITA (8x7B) (Fu et al., 2024)	✓	–	5M	-	74.7	71.4	39.7	44.9	45.3	74.3	59.0
OpenOmni (7B)	✓	0.6M	1.7M	**52.3**	**76.2**	**76.4**	**44.2**	**52.7**	**46.7**	**74.8**	**64.3**

Table 4: Comparison with state-of-the-art methods on visual-language benchmarks, including an indication of audio input/output support. The best performance among fully open-source models is highlighted in **bold**. Results demonstrate that integrating audio input and output further enhances the model's visual-language capabilities.

Layers	Experts	Wenetspeech(ZH)		Librispeech(EN)	
		Test_Net	Test_Meeting	Test_clean	Test_other
2	1	113.6	129.7	87.8	96.5
2	2	16.7	22.3	10.7	14.6
2	4	8.5	8.4	4.2	4.7
4	4	7.3	7.9	**3.8**	**4.3**
6	4	**6.4**	**6.7**	4.1	4.5

Table 5: Ablation study on the number of layers and experts in the speech decoder. Increasing experts in the mixture of experts module stabilizes CTC loss during training and enhances speech generation capacity. Deeper transformer layers improve English and Chinese speech generation, with greater benefits for Chinese.

C Additional Ablation Study

In order to explore the effect of TGM on speech generation in two modes, we plot the change of training loss under the same setting. As shown in fig. 4, we can observe that TGM can significantly improve the convergence speed of training and the performance of model speech generation, which verifies the effectiveness of our model design, whether it is (next token prediction) NTP loss under stable AR mode or CTC loss under unstable NAR mode.

To explore the impact of the number of layers in the NAR speech decoder and the mixture of experts module on Chinese and English speech generation, we conduct ablation experiments on WeNetSpeech (Zhang et al., 2022) and LibriSpeech (Panayotov et al., 2015). As illustrated in table 5, the instability and fragility associated with training using the CTC loss function present significant challenges. When simply employing a single feed-forward network (i.e., experts = 1), it becomes increasingly difficult to reconcile the conflicting training dynamics inherent in mixed-language scenarios, particularly when dealing with varying response lengths. As a result, training the speech decoder under these conditions proves to be quite challenging. Our findings demonstrate that incrementally increasing the number of experts significantly enhances the model's performance in bilingual speech generation, thereby underscoring the effectiveness of our mixture of experts module design. However, we observe inconsistent preferences regarding the optimal number of layers in the speech decoder for generating speech in Chinese and English. Specifically, while four layers yield the best results for English generation, six layers are more suitable for generating Chinese speech.

D Additional Implementation Details

OpenOmni is trained in five sequential sub-stages: speech-to-text generation, image-text pretraining, image-text instruction tuning, real-time speech generation, and self-aware emotional speech generation. Further details on these training stages are provided in table 7.

As shown in fig. 5, we provide more details of the speech decoder design and training here. For speech decoder, OpenOmni supports both autoregressive (AR) and non-autoregressive (NAR) methods. Specifically, the AR mode has better generation quality but a slower generation speed, while the NAR mode can achieve real-time speech generation but the generation quality is slightly worse. At the same time, in order to train the speech generator more efficiently, we also design a text-guided feature fusion module, so that the conditional features used for speech generation have more accurate alignment semantics, which can improve the

Method	Timbre	Tone	Melody	Space	Time	Hall	Intricacy	Overall
OneLLM (7B) (Han et al., 2024)	25.0	25.5	21.5	37.5	**29.3**	25.5	38.4	27.4
PandaGPT (7B) (Su et al., 2023)	23.5	23.2	27.6	45.0	23.8	28.0	23.9	26.7
Video-LLaMA (7B) (Zhang et al., 2023b)	25.5	22.3	24.4	30.0	26.2	25.0	30.7	26.1
Video-LLaMA2(7B) (Cheng et al., 2024)	24.1	25.5	26.4	30.0	27.2	**33.0**	34.5	26.8
AnyGPT (7B) (Zhan et al., 2024)	24.6	25.0	26.4	27.5	29.2	29.0	25.7	26.1
NexTGPT (7B) (Wu et al., 2023)	23.3	20.9	27.8	30.0	28.8	28.5	23.6	25.5
VITA (7x8B) (Fu et al., 2024)	24.1	26.4	27.8	22.5	26.3	31.0	36.8	26.4
OpenOmni (7B)	23.9	**27.7**	25.9	**60.0**	25.2	29.5	**37.6**	**32.8**

Table 6: Overall omni-understanding results on AV-Odyssey Bench. We conduct a performance comparison of omni-understanding among various fully open-source Omnimodal Large Language Models (OLLMs) on AV-Odyssey Bench. Compared to the state-of-the-art OLLM, VITA (Fu et al., 2024), which was trained on tri-modal data, OpenOmni achieves comparable advanced performance using significantly less training data and smaller model size.

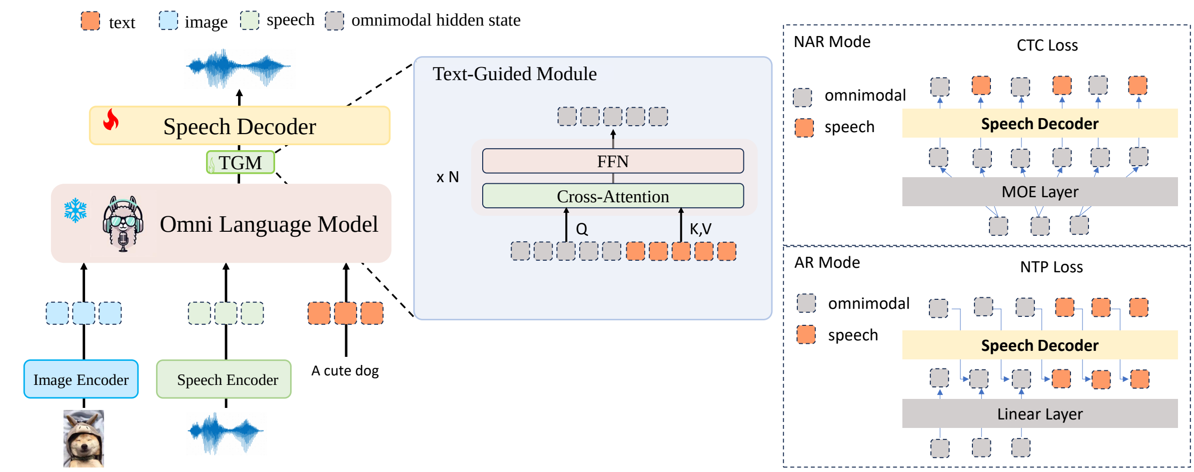

Figure 5: **Overview of Text-Guided Module and Speech Decoder Mode.** (left) Text-guided module fuse the hidden state and response textual feature via cross-attention, accelerating convergence speed of training without dropping the speed of speech decoding and context emotion perception. (right) OpenOmni supports both autoregressive (AR) and non-autoregressive speech (NAR) generation. The NAR mode uses CTC loss modeling and a 6K speech vocabulary size to enable real-time parallel speech decoding generation. The AR mode uses NTP loss modeling and a speech vocabulary size of 16K to support streaming decoding and higher-quality speech generation. In order to make the training of speech generator more stable and easy, we design a text-guided output feature fusion method to ensure the correctness of semantic alignment in speech generation modeling.

Hyperparameter	I	II	III	IV	V
batch size	256	128	128	32	32
lr	$1e^{-3}$	$1e^{-3}$	$5e^{-5}$	$5e^{-4}$	$5e^{-4}$
warmup ratio	0.3	0.3	0.3	0.3	0.3
epoch	1	1	1	3	3
freeze LLM	✔	✔	✘	✔	✔
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
cost	40 GPU·H	80 GPU·H	500 GPU·H	36 GPU·H	8 GPU·H
dataset	1-1	2-1	2-2	3-1	3-2
loss	$\mathcal{L}_{s2t}$	$\mathcal{L}_{i2t}$	$\mathcal{L}^{I}_{i2t}$	$\mathcal{L}_{ctc}$	$\mathcal{L}_{dpo}$

Table 7: The detailed training setup for OpenOmni and the hyper-parameters across the training stage. All experiments are conducted in 8xA100 setting.

generation quality and training efficiency of the speech decoder.

Non-autoregressive mode: In the non-autoregressive mode, the conditional features generated by OLLM are fed into the speech decoder by a layer of MOE and then upsampled to obtain the predicted speech output, and finally the end-to-end optimization is carried out by CTC loss modeling of the speech output. Due to the instability of CTC loss training, the smaller the size of the speech vocabulary, the easier it is to be successfully trained, but the generation quality of the corresponding speech will be affected by the smaller vocabulary.

Autoregressive mode: The autoregressive mode projects the conditional features generated by OLLM into the speech space through a layer of linear layer and feeds them into the speech decoder to obtain the speech prediction output, and finally optimizes the speech output end-to-end by modeling the NTP loss. Due to the stability of NTP loss training, the quality of speech generation will be higher than that of non-autoregressive generation, but the speed of speech generation will be reduced by autoregressive decoding.

Both AR and NAR modes depend on the quality of the speech generation conditional features generated by OLLM. Although OpenOmni will let the OLLM fit the text answer corresponding to the speech through multiple rounds of training in advance, there will still be OLLM output decoded into the wrong text answer. In this case, the erroneous generated condition features will be incorrectly aligned with the speech during the training process, which will ultimately reduce the performance of the speech decoder. In order to ensure the efficiency of training, OpenOmni fuse the speech generation condition features output by OLLM with the corresponding text features with correct semantics, and then feed them into the speech decoder for speech generation modeling training. Through the feature fusion module of text prior, OpenOmni avoids the misalignment of speech and corresponding text and ultimately makes the speech decoder training more stable and, at the same time, has more efficient and accurate speech generation quality.

English Prompts used in Text-to-Speech (T2S) and Speech-to-Text (S2T) tasks.

S2T (ASR)
Please translate the user's spoken input into corresponding text, word by word.
T2S
Echo game: Repeat this sentence back to me exactly as I said it!
Mirror challenge: Reflect my words right back at me!
Parrot mode: Can you mimic what I just said?
Copycat test: Let's see how well you can replicate my sentence!
Playback time: Rewind and repeat my last sentence.
Echo chamber: Say back to me what I just mentioned.
Simulate my speech: Can you say exactly what I just did?
Mimicry game: Echo my words precisely!
Repeat relay: Send my sentence back to me unchanged.
Reflection exercise: Reiterate my last statement word for word.

English emotional preference data case.

Emotion: Neutral
user: "Please engage in a conversation with me."
assistant: "Of course, please go ahead with your question." [A calm and collected woman, speaking in a steady and neutral tone.]
user: "I was wondering how you manage to stay so composed all the time."
assistant: "It's something that comes with practice and mindfulness. I focus on observing my surroundings and maintaining a balanced perspective." [A composed woman, speaking in a measured and tranquil tone.]
user: "That sounds quite challenging. Do you ever feel overwhelmed?"
assistant: "There are moments when things can be overwhelming, but I remind myself to breathe and take one step at a time." [A calm woman, delivering her words in an even and serene manner.]
user: "Do you have any techniques to stay calm under pressure?"
assistant: "Yes, I practice meditation and deep breathing exercises. It helps to center my thoughts and maintain clarity." [A composed woman, speaking with a steady and soothing tone.]
user: "I see. Do you think anyone can learn to be as calm as you?"
assistant: "Absolutely. It requires patience and consistent practice, but anyone can develop the ability to remain calm and composed." [A neutral-toned woman, expressing herself in a clear and composed manner.]
user: "What advice would you give to someone starting on this path?
assistant: "Begin with small steps. Start by observing your thoughts and emotions without judgment. Practice mindfulness and breathing exercises daily." [A composed woman, offering her advice in a gentle and calm tone.]
user: "Thank you for your advice. It has been very insightful."
assistant: "You're welcome. I'm glad I could help. Remember, the journey to calmness is gradual, but rewarding." [A tranquil woman, speaking in a reassuring and composed tone.]