

Eureka-Audio: Triggering Audio Intelligence in Compact Language Models

Dan Zhang^{*1} Yishu Lei^{*1} Jing Hu^{*1,3} Shuwei He^{*1,2} Songhe Deng¹
 Xianlong Luo¹ Danxiang Zhu¹ Shikun Feng^{†1} Rui Liu²
 Jingzhou He¹ Yu Sun¹ Hua Wu¹ Haifeng Wang¹

¹Baidu Inc. ²College of Computer Science, Inner Mongolia University

³Tsinghua Shenzhen International Graduate School, Tsinghua University

{zhangdan20, leiyishu, heshuwei, dengsonghe, luoxianlong, zhudanxiang}@baidu.com
 {fengshikun01, hejingzhou, sunyu02, wu_hua, wanghaifeng}@baidu.com
 cminusser@gmail.com, imucslr@imu.edu.cn

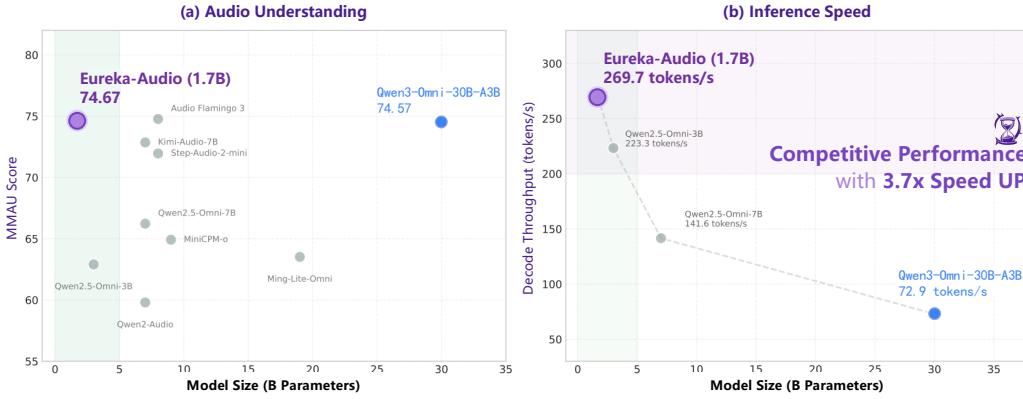


Figure 1: Comparison of Eureka-Audio with open-source audio-language and omni-modal baselines. (a) On the MMAU benchmark, Eureka-Audio (1.7B) achieves a score of 74.67, competitive with models 4–17× larger. (b) Eureka-Audio achieves the highest decode throughput of 269.7 tokens/sec among the compared models.

Abstract

We present **Eureka-Audio**, a compact yet high-performance audio language model that achieves competitive performance against models that are 4 to 18 times larger across a broad range of audio understanding benchmarks. Despite containing only 1.7B parameters, Eureka-Audio demonstrates strong performance on automatic speech recognition (ASR), audio understanding, and dense audio captioning, matching or surpassing multiple 7B to 30B audio and omni-modal baselines. The model adopts a unified end-to-end architecture composed of a lightweight language backbone, a Whisper-based audio encoder, and a sparsely activated Mixture-of-Experts (MoE) adapter that explicitly accounts for audio heterogeneity and alleviates cross-modal optimization conflicts under limited capacity. To further enhance paralinguistic reasoning, we introduce DataFlux, a closed loop audio instruction data synthesis and verification pipeline that constructs high quality, logically consistent supervision from raw audio. Extensive evaluations across ASR, knowledge reasoning, safety, instruction following, and paralinguistic benchmarks, demonstrate that Eureka-Audio achieves an efficient balance between computational cost and performance. These results establish Eureka Audio as a strong and practical baseline for lightweight audio understanding models.

^{*}Equal contribution.

[†]Corresponding author.

Project Page: <https://github.com/Alittleegg/Eureka-Audio>

1 Introduction

Recent advances in multimodal large language models have driven a paradigm shift in the audio domain, moving beyond speech recognition toward more general audio understanding. Unlike conventional speech-centric approaches that focus on transcription, audio understanding requires models to jointly capture semantic content (e.g., what is said or what event occurs) and paralinguistic cues (e.g., emotion, tone, laughter, hesitation, and emphasis). Such capabilities are crucial for real-world applications including intelligent assistants, customer service quality inspection, content retrieval, and general-purpose audio analysis.

Despite this progress, deploying high-quality audio understanding models in practical settings remains challenging. Recent methods often rely on substantially larger models, such as Kimi-Audio-7B [1], Step-Audio-2-mini-8B [2], Qwen2-Audio-7B [3], and Qwen3-Omni-30B-A3B-Instruct [4], to achieve strong performance. However, the high inference latency and computational costs overhead inherent in these models often preclude their use in real-time or resource-constrained scenarios.

Consequently, there is still a notable scarcity of lightweight, open-source audio understanding models that offer both high performance and full reproducibility.

Achieving strong audio understanding under a lightweight setting is non-trivial. Audio signals are inherently heterogeneous, with speech, environmental sounds, and music exhibiting distinct statistical structures and representational characteristics. Naively using a single shared projection for cross-modal alignment often introduces conflicting optimization signals and limits representational capacity. Moreover, under limited model capacity, semantic information and paralinguistic cues are more prone to competing for limited representational capacity, leading to reduced parameter efficiency and degraded generalization. These challenges are further exacerbated by the scarcity of high-quality audio instruction and preference data, which can cause instability and performance degradation during post-training.

To address these challenges, we introduce **Eureka-Audio**, an open-source, high-performance lightweight model for audio semantic and paralinguistic understanding. Eureka-Audio adopts Qwen3-1.7B-base [5] as its language backbone. Audio inputs are first encoded into continuous acoustic representations using a Whisper-based audio encoder, and are then mapped into the language model’s semantic space through a sparsely activated Mixture-of-Experts (MoE) adapter [6]. This sparse adaptation mechanism explicitly accounts for audio heterogeneity and improves cross-modal alignment quality.

We further construct a closed loop training pipeline covering both pre-training and post-training stages. During pre-training, the model acquires fundamental audio understanding and cross-modal alignment capabilities. During post-training, we combine open-source and in-house data and introduce **DataFlux**, a systematic audio instruction data synthesis pipeline that enables structured data management and continuous updates.

As shown in Figure 1, Eureka-Audio consistently achieves competitive performance across ASR, audio semantic understanding, and paralinguistic understanding benchmarks, despite being significantly smaller than many strong audio and omni-modal baselines, while also delivering up to $3.7\times$ faster decoding speed.

Our contributions are summarized as follows:

- **Eureka-Audio is introduced as a lightweight yet high-performance audio understanding model** with only **1.7B parameters**, achieving strong results on both audio semantic and paralinguistic understanding tasks while remaining suitable for efficient real-world deployment.
- A **sparsely activated MoE-based adapter** is developed between the Whisper-based audio encoder and the language backbone to explicitly address audio heterogeneity and improve cross-modal alignment. Based on this design, a complete training pipeline significantly enhances training stability and multimodal alignment quality.
- **DataFlux is a structured audio instruction data synthesis pipeline** designed to systematically construct high-quality paralinguistic instruction data. Through a multi-step data generation and validation process, the pipeline ensures reliable semantic alignment and logical consistency of the synthesized data, thereby supporting the post-training stage and

enabling models to progressively enhance audio paralinguistic understanding and reasoning under controlled, high-quality data supervision.

- **A novel evaluation methodology for audio captioning is proposed**, enabling more faithful assessment of high-level audio understanding across diverse audio types, including speech, sound, and music.

2 Related Work

2.1 Large Audio Language Model

Existing Large Audio Language Models [7, 3, 2, 1, 8, 9, 10] typically adopt an end-to-end architecture built upon a pre-trained LLM backbone. In this framework, acoustic features are integrated into the model via an audio encoder, with a trainable adapter acting as the bridge between the encoder and the LLM. Most existing works, including ours, utilize Automatic Speech Recognition (ASR) models, such as Whisper [11], as the audio encoder due to their robust representation capabilities. In contrast, a subset of methods [2, 1] involves using an audio tokenizer, such as Residual Vector Quantization (RVQ), or Finite Scalar Quantization(FSQ), to quantize audio signals into discrete codebook [12] indices for model input.

2.2 Lightweight Multimodal Models

Since Large Audio Language Models prioritize real-time interactivity, they demand strictly optimized FLOPs and inference latency. While mainstream models [7, 3, 2, 1, 8, 9] typically exceed 7B parameters, leading to substantial computational overhead, some recent works [10, 4] adopt MoE architectures to reduce active parameters to approximately 3B. However, the excessive total memory footprint of MoE models remains a significant bottleneck for deployment on resource-constrained edge devices. To address these challenges, we propose a 1.7B dense model, which significantly alleviates the deployment overhead while maintaining high-performance audio interaction.

3 Architecture

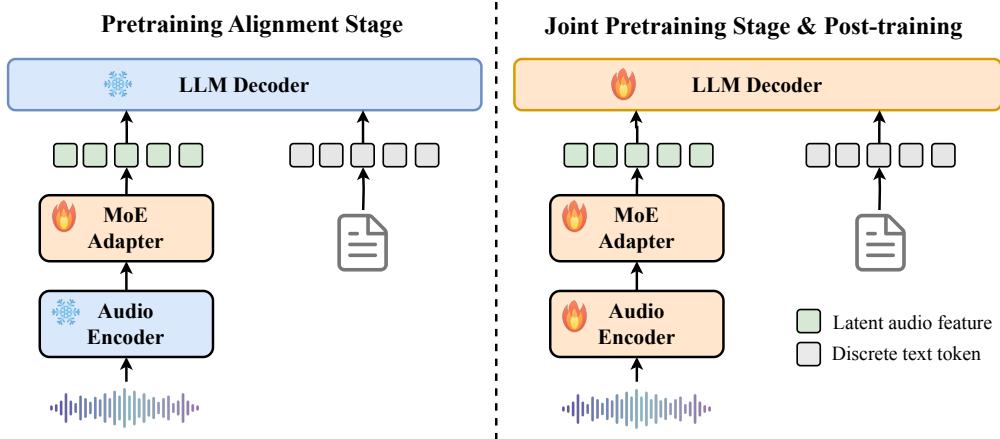


Figure 2: The overview of the Eureka-Audio. Eureka-Audio adopts a unified end-to-end design consisting of three core components: (1) **a Whisper-based audio encoder** that encodes raw waveforms into high-temporal-resolution acoustic representations; (2) **a sparse MoE adapter** [6] that maps acoustic features into the language model embedding space for efficient cross-modal alignment; and (3) **a lightweight language model backbone** (Qwen3-1.7B-base [5]) that jointly models aligned audio embeddings and text tokens in an autoregressive manner to support diverse audio understanding tasks.

3.1 Overview

We propose Eureka-Audio, a lightweight audio–language model designed for general audio understanding tasks. The model adopts a unified end-to-end architecture that tightly integrates Whisper-based audio encoder, a sparse MoE adapter, and a lightweight language model backbone, achieving a favorable balance between modeling capacity and computational and parameter efficiency. As illustrated in Figure 2, Eureka-Audio consists of three core components:

Audio Encoder we employ a Whisper [13]-based audio encoder to encode raw audio waveforms into high-temporal-resolution acoustic features, which capture fine-grained perceptual and semantic information present in the audio signal.

Sparse MoE Adapter. The continuous acoustic features produced by the Whisper [13]-based audio encoder are first fed into a MoE Adapter [6], which maps audio representations into the embedding space of the language model. Serving as a critical interface between audio representations and the language backbone, the MoE Adapter enables efficient cross-modal alignment through sparse expert routing. This design explicitly models the heterogeneity of audio signals at both the semantic and acoustic levels, mitigating optimization conflicts while improving representational efficiency and modeling flexibility under controlled parameter and computational overhead.

Language Model Backbone. We adopt Qwen3-1.7B-base [5] as the language model backbone. After alignment via the MoE Adapter, audio embeddings are concatenated with text token embeddings along the sequence dimension and jointly modeled by the backbone in a standard autoregressive manner. The model outputs text tokens, enabling a wide range of downstream audio understanding tasks, including audio question answering and instruction following.

Overall, this architecture allows Eureka-Audio to perform both audio semantic understanding and paralinguistic reasoning within a unified and lightweight framework, while remaining well suited for deployment in resource-constrained scenarios.

3.2 Sparse MoE Adapter

To project audio representations into the LLM embedding space, we adopt a sparse Mixture-of-Experts (MoE) adapter [6] instead of a conventional dense projector. Given an input audio token $\mathbf{x} \in \mathbb{R}^d$, a learnable router computes gating logits $G(\mathbf{x}) = \mathbf{x}\mathbf{W}_g$ and selects the Top- k experts via sparse softmax routing. Each expert is implemented as a lightweight feed-forward network with SiLU activation. The selected expert outputs are aggregated according to the routing weights and further mapped to the LLM embedding dimension through a linear projection followed by layer normalization:

$$\mathbf{y}_{\text{MoE}} = \mathcal{LN} \left(\mathbf{W}_P \sum_{i \in \mathcal{I}} G(\mathbf{x})_i \cdot E_i(\mathbf{x}) \right), \quad (1)$$

where \mathcal{I} denotes the set of routed experts and $G(\mathbf{x})$ is the sparse gating distribution.

The resulting adapted audio embeddings $\mathbf{Y}_{\text{MoE}} = \{\mathbf{y}_{\text{MoE}}^{(1)}, \dots, \mathbf{y}_{\text{MoE}}^{(T_a)}\}$ are concatenated with text token embeddings to form the final input sequence to the LLM.

Training Objective. Let \mathbf{Z} denote the complete input embedding sequence to the LLM, consisting of the adapted audio embeddings and textual token embeddings. The model is trained end-to-end using the standard autoregressive next-token prediction (NTP) objective:

$$\mathcal{L}_{\text{NTP}} = - \sum_{t=1}^T \log P(y_t | y_{<t}, \mathbf{Z}; \theta), \quad (2)$$

where θ denotes all trainable parameters including the MoE adapter and the LLM backbone.

To mitigate expert collapse, we incorporate a load-balancing auxiliary loss over the routed experts. Let B be the number of routed audio tokens in a batch, and $p_{b,e}$ denote the routing probability of token b to expert e . The auxiliary loss is defined as:

$$\mathcal{L}_{\text{aux}} = |\mathcal{E}_R| \sum_{e \in \mathcal{E}_R} \bar{P}_e \cdot \bar{f}_e, \quad (3)$$

where $\bar{P}_e = \frac{1}{B} \sum_{b=1}^B p_{b,e}$ represents the expert importance and \bar{f}_e is the fraction of tokens routed to expert e . The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda \mathcal{L}_{\text{aux}}. \quad (4)$$

4 Pretraining

We adopt a two-stage pretraining framework for Eureka-Audio, consisting of an alignment stage (Stage 1) and a joint pretraining stage (Stage 2). This design aims to first establish a stable audio–text modality alignment and subsequently achieve comprehensive audio understanding through the joint optimization of the full model.

Stage 1 (Alignment Stage). In this stage, only the **MoE Adapter** is trainable, while the parameters of both the language model backbone and the audio encoder remain frozen. Training leverages three types of data: audio unimodal modeling, audio-to-text mapping, and audio–text interleaved modeling. This configuration enables the adapter to learn an effective mapping from acoustic representations to the embedding space of the language model, establishing a robust audio–text modality alignment. The primary objective of this stage is to ensure high-quality cross-modal alignment and training stability.

Stage 2 (Joint Pretraining Stage). In this stage, all model parameters are unfrozen and jointly optimized. In addition to the data types utilized in Stage 1, we incorporate audio captioning data to enhance the capacity of the model to capture high-level semantic information and paralinguistic cues. This stage enables the model to learn richer audio semantics and complex reasoning patterns, leading to improved overall audio understanding capabilities.

4.1 Task Formulation

Our pretraining objective is formulated over a mixture of tasks, including:

- **Unimodal Modeling.** This category involves next-token prediction for both text tokens and discrete semantic audio tokens¹. These tasks serve to preserve the language modeling capability of the backbone and to learn the fundamental distribution of audio representations.
- **Audio–Text Mapping.** This category encompasses automatic speech recognition (ASR, Audio → Text) and text-to-speech (TTS, Text → Audio semantic tokens), providing direct supervision for the cross-modal correspondence between audio and text.
- **Audio–Text Interleaving.** We construct interleaved sequences under audio-conditioned settings, where the model is trained to predict either the next semantic audio token or the next text token. This task further strengthens the coupling between acoustic and linguistic representations.
- **Audio Captioning.** This category covers tasks such as acoustic scene understanding, emotion recognition, sound event detection, sound source identification, and music understanding, offering high-level semantic supervision for holistic audio understanding.

4.2 Dataset Composition and Distribution

During Stage 1 (Alignment Stage), the model is trained on approximately **100B tokens**. In Stage 2 (Joint Pretraining Stage), the total training scale scales up to approximately **1T tokens**. The approximate scale of each data category is summarized in Table 1. Notably, audio captioning data is introduced exclusively in Stage 2 and consists of various open-source datasets; detailed specifications are provided in Appendix 8.1.

Table 1: Data distribution and training schedule across different training tasks and stages. Audio and text data are jointly sampled with a fixed 1:1 ratio. For brevity, the table reports statistics for the **audio modality only**.

Task	Audio Length (Hours)	Tokens (B)	Stage 1 Task Ratio	Stage 2 Task Ratio
Audio Unimodal Modeling	500,000	25	0.2	0.03
Audio–Text Mapping	5,500,000	360	0.45	0.56
Audio–Text Interleaving	5,150,000	150	0.35	0.07
Audio Captioning	220,000	18	—	0.34

5 Post-training

5.1 DataFlux

To systematically construct high-quality audio paralinguistic instruction data and effectively facilitate paralinguistic understanding and reasoning during post-training, we propose **DataFlux**, a data synthesis and filtering pipeline tailored for paralinguistic tasks. Existing open-source post-training datasets often derive from earlier model versions, potentially leading to misalignment with the requirements of more capable models as their reasoning abilities evolve. Starting from raw audio, DataFlux progressively builds structured, logically consistent, and reasoning-oriented audio instruction data through a multi-stage process of generation, alignment, and validation.

As illustrated in Figure 3, DataFlux operates via a three-step workflow. In **Step 1**, raw audio is processed by an audio captioner (Qwen3-Omni-30B-A3B-Captioner [4]) to generate Audio Dense Captions, which provide fine-grained descriptions of acoustic events, environmental contexts, and potential paralinguistic cues. Based on a predefined paralinguistic taxonomy and a small set of manually curated exemplars, a large language model (GPT-OSS-120B [15]) transforms these captions into structured instruction formats, producing Query–Choice pairs strictly aligned with the audio content. This step effectively maps continuous, unstructured audio signals into a discrete instruction space suitable for model training.

In **Step 2**, the generated Query–Choice pairs, along with the original audio, are fed into multiple audio large language models (Qwen3-Omni-30B-A3B-Thinking [4] and Step-Audio-R1 [16]) to produce reasoning traces and final answers. By leveraging models with distinct reasoning characteristics, this step explicitly induces answer diversity, serving as a foundation for subsequent data filtering and hard-example mining. The outputs include not only final answers but also intermediate reasoning traces, enabling a more granular assessment of paralinguistic understanding.

In **Step 3**, DataFlux employs an automated judging mechanism to systematically evaluate the multi-model outputs. A judge model (GPT-OSS-120B [15]) jointly assesses the original audio captions, the Query–Choice pairs, and the reasoning traces and answers produced by the different models. Judgments rely on logical consistency, coverage of salient details, and the absence of semantic conflicts with audio descriptions. Based on these evaluations, samples are categorized by quality: those exhibiting consistent reasoning and high agreement with the audio captions are retained, while instances with clear conflicts or reasoning failures are filtered out, thereby effectively reducing noise during post-training.

Through this three-step design, DataFlux realizes an end-to-end automated pipeline for constructing high-quality paralinguistic instruction data from raw audio. The pipeline ensures reliable semantic alignment and logical consistency of the synthesized data, providing a scalable and extensible source for controlled, high-quality post-training supervision.

5.2 Supervised Fine-Tuning

During the Supervised Fine-Tuning (SFT) stage, all model parameters are unfrozen and jointly optimized, with a total training scale of approximately **30B tokens**. Throughout training, text and audio modalities are sampled using a fixed 1:1 ratio. The detailed scale and composition of the audio data are summarized in Table 2.

¹The discrete semantic audio tokens are extracted using the tokenizer from GLM-4-Voice [14].

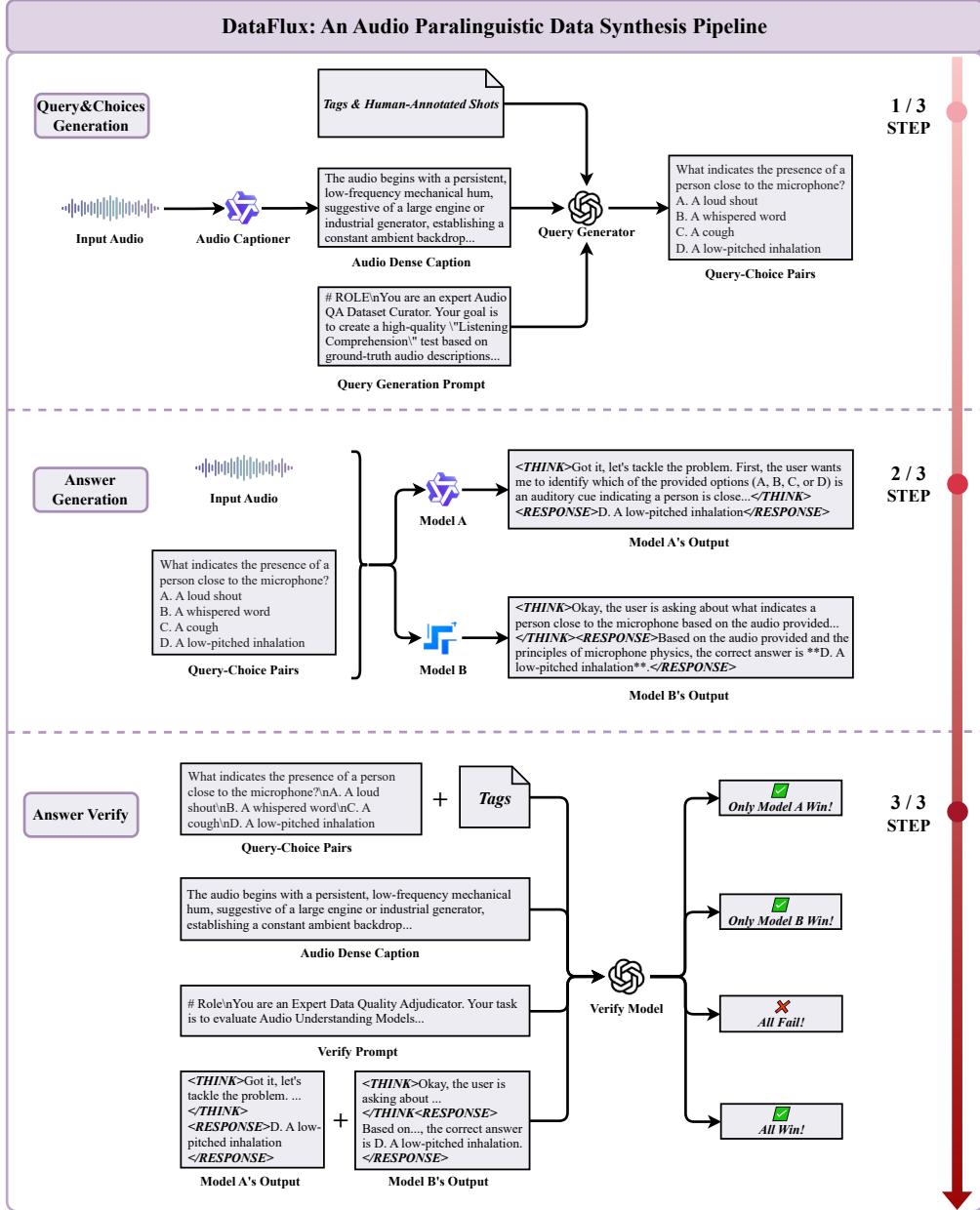


Figure 3: Overview of DataFlux. Starting from raw audio, DataFlux constructs high-quality paralinguistic instruction data through a three-step workflow: (1) **Query–Choice Generation**, where dense audio captions are first produced and then transformed into structured Query–Choice pairs using a predefined paralinguistic taxonomy and few-shot exemplars; (2) **Answer Generation**, where multiple audio large language models generate reasoning traces and answers conditioned on the same audio and queries; and (3) **Answer Verification**, where an automated judge evaluates multi-model outputs based on logical consistency and alignment with the audio content, retaining reliable samples while filtering noisy or inconsistent ones.

Table 2: Audio data composition and sampling ratios used during the SFT stage.

Task Type	Audio Length (Hours)	Ratio
ASR	250,000	0.6
Paralinguistic Understanding	2,500	0.1
Semantic Understanding	100,000	0.2
Audio Dense Captioning	2,500	0.1

6 Evaluation

In this section, we conduct a comprehensive evaluation of Eureka-Audio, covering automatic speech recognition (ASR), audio understanding, and dense audio captioning, to assess its effectiveness as a lightweight audio understanding model. Unlike conventional evaluation protocols that primarily focus on transcription accuracy or isolated audio classification tasks, our evaluation emphasizes high-level audio understanding, requiring models to jointly reason over semantic content and paralinguistic cues across diverse audio types, including speech, sounds, and music.

We compare Eureka-Audio with a diverse set of large scale audio language models and omni modal language models, and report results under their officially recommended inference configurations to ensure fair comparison. Despite having only 1.7 billion parameters of LLM backbone, Eureka-Audio consistently achieves competitive performance across a wide range of benchmarks, demonstrating an efficient balance between model capacity and task performance. Furthermore, we conduct systematic ablation studies to examine the contribution of the proposed DataFlux framework and to quantify its impact on overall audio understanding performance.

6.1 Automatic Speech Recognition

To evaluate the automatic speech recognition (ASR) capability of Eureka-Audio, we conduct a systematic evaluation on a diverse set of standard benchmarks covering multiple languages and acoustic conditions, including LibriSpeech [17], Fleurs [18], AISHELL [19], and WenetSpeech [20]. These benchmarks span both English and Mandarin speech, and cover a wide range of scenarios such as clean speech, noisy environments, conversational speech, and meeting-style recordings. We report word error rate (WER) for English and character error rate (CER) for Mandarin, where lower values indicate better recognition performance.

As shown in Table 3, despite its significantly smaller model size, Eureka-Audio-Instruct achieves competitive performance across all ASR benchmarks. Notably, on widely used English benchmarks such as LibriSpeech[17] and Fleurs[18], Eureka-Audio attains lower error rates than several larger omni-modal and audio-centric models, including Qwen2.5-Omni and MiniCPM-o. These findings demonstrate that Eureka-Audio preserves robust speech recognition performance while attaining a desirable balance between predictive accuracy and computational efficiency.

6.2 Audio Understanding

In this set of experiments, we cover both audio semantic understanding (e.g., audio-based factual reasoning and content comprehension) and audio paralinguistic understanding (e.g., emotion, environmental sounds, music, and reasoning based on fine-grained acoustic cues). We report results on a diverse set of benchmarks spanning multiple dimensions, including knowledge reasoning, instruction following, safety, and audio paralinguistic understanding. In particular, we adopt the MMAU [21] and MMAR [22] benchmarks to evaluate the model’s ability to perform complex reasoning by going beyond literal speech content and jointly leveraging fine-grained acoustic cues.

As shown in Table 4, despite containing only 1.7B parameters of LLM backbone, Eureka-Audio demonstrates competitive performance across a wide range of audio understanding benchmarks. Notably, on the MMAU benchmark, Eureka-Audio-Instruct outperforms several substantially larger models, including Qwen3-Omni-Instruct, achieves competitive or state-of-the-art-level performance. This result indicates that Eureka-Audio is capable of effectively modeling fine-grained audio semantic and paralinguistic information under limited model capacity, and exhibits strong reasoning ability on challenging audio understanding tasks.

Table 3: ASR performance comparison. Using negative multirow prevents text clipping on colored backgrounds.

Datasets	Type	Model	Size	WER/CER ↓
LibriSpeech test-clean test-other	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	1.60 2.93
		Ming-Lite-Omni-1.5	19B-A2.8B	1.90 3.54
		MiniCPM-o	9B	2.01 4.87
		Qwen2.5-Omni-7B	7B	1.53 3.19
		Qwen2.5-Omni-3B	3B	1.68 3.90
	<i>Audio</i>	Step-Audio-2-mini	8B	1.41 2.76
		Audio Flamingo 3	8B	1.39 2.96
		Qwen2-Audio	7B	1.74 4.01
		Kimi-Audio-7B-Instruct	7B	1.33 2.57
	<i>Ours</i>	Eureka-Audio-Base	1.7B	1.59 3.34
	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	1.46 3.24
Fleurs-en	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	5.04
		Ming-Lite-Omni-1.5	19B-A2.8B	5.82
		MiniCPM-o	9B	6.18
		Qwen2.5-Omni-7B	7B	5.49
		Qwen2.5-Omni-3B	3B	5.65
	<i>Audio</i>	Step-Audio-2-mini	8B	4.51
		Audio Flamingo 3	8B	6.30
		Qwen2-Audio	7B	6.92
		Kimi-Audio-7B-Instruct	7B	6.11
	<i>Ours</i>	Eureka-Audio-Base	1.7B	5.73
	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	5.39
AISHELL-2 ios	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	2.63
		Ming-Lite-Omni-1.5	19B-A2.8B	2.66
		MiniCPM-o	9B	3.42
		Qwen2.5-Omni-7B	7B	2.58
		Qwen2.5-Omni-3B	3B	2.77
	<i>Audio</i>	Step-Audio-2-mini	8B	2.33
		Qwen2-Audio	7B	3.08
		Kimi-Audio-7B-Instruct	7B	2.80
	<i>Ours</i>	Eureka-Audio-Base	1.7B	3.17
	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	3.10
WenetSpeech test-meeting test-net	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	6.12 5.29
		Ming-Lite-Omni-1.5	19B-A2.8B	5.96 6.26
		MiniCPM-o	9B	15.53 7.68
		Qwen2.5-Omni-7B	7B	8.43 7.10
		Qwen2.5-Omni-3B	3B	8.53 7.14
	<i>Audio</i>	Step-Audio-2-mini	8B	5.43 5.50
		Qwen2-Audio	7B	8.40 8.00
		Kimi-Audio-7B-Instruct	7B	6.38 7.17
	<i>Ours</i>	Eureka-Audio-Base	1.7B	10.37 8.63
	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	9.14 7.55

Furthermore, after incorporating post-training data generated by DataFlux, the model achieves stable and consistent performance improvements on benchmarks related to audio paralinguistic understanding. This observation suggests that high-quality audio instruction data plays a critical role in stabilizing post-training and improving generalization, which is particularly important for lightweight models.

Overall, these results validate that Eureka-Audio achieves a favorable balance between efficiency and performance, enabling effective modeling of both audio semantic information and paralinguistic cues without relying on large-scale model parameters, and thereby supporting more general and fine-grained audio understanding capabilities.

Table 4: Performance comparison on audio understanding benchmarks. Following the standard layout with full-width spacing.

Datasets	Type	Model	Size	Performance ↑
Knowledge MMSU OpenBookQA	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	77.00 92.31
		Ming-Lite-Omni-1.5	19B-A2.8B	47.00 69.67
		MiniCPM-o	9B	54.55 79.12
		Qwen2.5-Omni-7B	7B	61.22 81.53
	<i>Audio</i>	Qwen2.5-Omni-3B	3B	53.41 77.36
Safety AdvBench	<i>Omni</i>	Step-Audio-2-mini	8B	55.14 75.60
		Audio Flamingo 3	8B	47.07 61.54
		Qwen2-Audio	7B	35.75 49.67
		Kimi-Audio-7B-Instruct	7B	61.26 84.18
	<i>Ours</i>	Eureka-Audio-Base	1.7B	38.03 52.53
Instruction IFEval	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	55.63 69.23
	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	99.61
		Ming-Lite-Omni-1.5	19B-A2.8B	99.23
		MiniCPM-o	9B	95.76
		Qwen2.5-Omni-7B	7B	96.92
Paralinguistic MMAU [†] MMAR	<i>Omni</i>	Qwen2.5-Omni-3B	3B	89.80
		Step-Audio-2-mini	8B	93.08
		Audio Flamingo 3	8B	98.26
		Qwen2-Audio	7B	98.84
		Kimi-Audio-7B-Instruct	7B	100.00
	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	99.81
Paralinguistic IFEval	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	81.17
		Ming-Lite-Omni-1.5	19B-A2.8B	53.68
		MiniCPM-o	9B	41.72
		Qwen2.5-Omni-7B	7B	39.84
		Qwen2.5-Omni-3B	3B	32.97
Paralinguistic MMAU [†] MMAR	<i>Audio</i>	Step-Audio-2-mini	8B	43.54
		Audio Flamingo 3	8B	32.27
		Qwen2-Audio	7B	26.24
		Kimi-Audio-7B-Instruct	7B	47.91
	<i>Ours</i>	Eureka-Audio-Instruct	1.7B	53.21
Paralinguistic MMAU [†] MMAR	<i>Omni</i>	Qwen3-Omni-Instruct	30B-A3B	74.57 67.10
		Ming-Lite-Omni-1.5	19B-A2.8B	63.52 45.40
		MiniCPM-o	9B	64.92 47.90
		Qwen2.5-Omni-7B	7B	66.23 49.60
		Qwen2.5-Omni-3B	3B	62.91 43.40
Paralinguistic MMAU [†] MMAR	<i>Audio</i>	Step-Audio-2-mini	8B	71.96 61.57
		Audio Flamingo 3	8B	74.77 61.00
		Qwen2-Audio	7B	59.80 37.90
		Kimi-Audio-7B-Instruct	7B	72.86 57.40
	<i>Ours</i>	Eureka-Audio-Base	1.7B	63.42 46.80
	<i>Ours</i>	Eureka-Audio-Instruct w/o DataFlux	1.7B	66.93 50.70
	<i>Ours</i>	Eureka-Audio-Instruct w DataFlux	1.7B	74.67 56.20

6.3 Dense Audio Captioning

To further evaluate high-level audio understanding capabilities, we conduct experiments on dense audio captioning. This task requires the model to generate comprehensive and semantically rich descriptions that jointly capture speech content, acoustic events, and non-verbal cues, thereby preserving as much information from the original audio as possible.

To assess caption quality in a structured and quantifiable manner, we adopt a two-stage evaluation framework. First, the model generates a dense caption for the input audio. The generated caption

Table 5: Dense audio caption evaluation results on MMAU and MMAR benchmarks.

Datasets	Model	Size	MMAU MMAR ↑
Dense Captioning	Qwen3-Omni-Captioner	30B-A3B	56.68 46.40
	Qwen3-Omni-Instruct	30B-A3B	48.24 36.90
	Eureka-Audio-Instruct (Ours)	1.7B	52.96 41.70

is then concatenated with a downstream question and fed into a large language model (GPT-OSS-120B), which produces the final answer. Since the correctness of the answer critically depends on how faithfully the dense caption preserves the information contained in the original audio, this question answering conditioned on the generated caption pipeline serves as an indirect yet effective measure of dense caption quality. Models that generate more informative and structurally coherent captions are expected to achieve higher accuracy in the downstream reasoning stage.

As shown in Table 5, Eureka-Audio-Instruct outperforms Qwen3-Omni-Instruct on dense caption evaluation metrics, and its performance is comparable to the dedicated captioning model Qwen3-Omni-Captioner. These results indicate that, despite having only 1.7B parameters of LLM backbone, Eureka-Audio-Instruct is capable of producing information-rich and semantically faithful dense audio descriptions, demonstrating strong high-level audio understanding ability under a lightweight model setting.

7 Conclusion

In this work, we introduced Eureka-Audio, a lightweight 1.7B audio–language model designed to achieve strong audio semantic and paralinguistic understanding under strict parameter constraints. Through a unified end-to-end architecture integrating a Whisper-based audio encoder, a sparsely activated MoE adapter, and a compact language backbone, the model effectively aligns heterogeneous audio representations with limited capacity while mitigating cross-modal optimization conflicts.

Extensive evaluations across ASR, knowledge reasoning, safety, instruction following, and paralinguistic benchmarks demonstrate that Eureka-Audio consistently achieves competitive performance compared to substantially larger audio and omni-modal models. In particular, the model exhibits strong paralinguistic perception and reasoning ability, and benefits significantly from the proposed DataFlux pipeline, which provides structured, logically consistent audio instruction data to stabilize and enhance post-training.

Beyond accuracy, Eureka-Audio delivers up to a $3.7\times$ speedup in decoding, highlighting a favorable efficiency and performance trade-off. These findings suggest that scaling model size is not the only path toward stronger audio intelligence; instead, careful architectural design and high-quality data synthesis can unlock competitive performance under constrained parameter budgets.

Looking ahead, we plan to extend Eureka-Audio beyond understanding toward unified audio generation and real-time interactive scenarios. Future work will explore lightweight and low-latency audio generation mechanisms, streaming architectures for real-time dialogue, and tighter integration between perception and generation within a single compact model. We believe such advancements will further bridge the gap between efficient modeling and practical audio intelligence deployment.

References

- [1] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.
- [2] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.
- [3] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- [4] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.
- [5] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [6] Yishu Lei, Shuwei He, Jing Hu, Dan Zhang, Xianlong Luo, Danxiang Zhu, Shikun Feng, Rui Liu, Jingzhou He, Yu Sun, Hua Wu, and Haifeng Wang. Moe adapter for large audio language models: Sparsity, disentanglement, and gradient-conflict-free, 2026.
- [7] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023.
- [8] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.
- [9] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2024.
- [10] Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.
- [11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- [12] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.
- [13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [14] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.
- [15] OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang,

Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogneni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotstetd, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025.

- [16] Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Dixin Jiang, and Gang Yu. Step-audio-r1 technical report, 2025.
- [17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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, 2015.
- [18] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.
- [19] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.
- [20] 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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022.
- [21] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.
- [22] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix, 2025.
- [23] Il-Young Jeong and Jeongsoo Park. Cochlsene: Acquisition of acoustic scene data using crowdsourcing. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–21. IEEE, 2022.
- [24] Paul Primus, Florian Schmid, and Gerhard Widmer. Tacos: Temporally-aligned audio captions for language-audio pretraining, 2025.
- [25] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. TUT database for acoustic scene classification and sound event detection. In *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.

- [26] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.
- [27] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [28] T. Heittola, A. Mesaros, T. Virtanen, T. Heittola, H. Laakso, and R. Bejarano Rodriguez. Tau urban acoustic scenes 2022 mobile, development dataset. Zenodo, 03 2022.
- [29] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.
- [30] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022.
- [31] Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6):e12233, 2023.
- [32] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.
- [33] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.
- [34] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.
- [35] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155, 2022.
- [36] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.
- [37] Shunian Chen, Xinyuan Xie, Zheshu Chen, Liyan Zhao, Owen Lee, Zhan Su, Qilin Sun, and Benyou Wang. Fusionaudio-1.2m: Towards fine-grained audio captioning with multimodal contextual fusion, 2025.
- [38] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.
- [39] Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, S Sakshi, Vaibhavi Lokegaonkar, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, Gunhee Kim, Jun Du, Rafael Valle, and Bryan Catanzaro. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge. *arXiv preprint arXiv:2505.07365*, 2025.
- [40] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE, 2021.
- [41] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

- [42] Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, and Vicente Ordonez. Taming data and transformers for audio generation, 2025.
- [43] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.
- [44] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.
- [45] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.
- [46] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*, 2023.
- [47] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization, 2024.
- [48] Ilaria Manco, Benno Weck, SeungHeon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation, 2023.
- [49] Irene Martín-Morató and Annamaria Mesaros. What is the ground truth? reliability of multi-annotator data for audio tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 76–80. IEEE, 2021.
- [50] Zeyu Xie, Xuanan Xu, Zhizheng Wu, and Mengyue Wu. Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation, 2024.
- [51] A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2022.
- [52] Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets, 2025.
- [53] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *arXiv preprint arXiv:2501.15907*, 2025.
- [54] 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 *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.
- [55] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.
- [56] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.
- [57] 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.
- [58] Evelina Bakhturina, Vitaly Lavrukhan, Boris Ginsburg, and Yang Zhang. Hi-Fi Multi-Speaker English TTS Dataset. *arXiv preprint arXiv:2104.01497*, 2021.
- [59] Wenbin Wang, Yang Song, and Sanjay Jha. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech, 2024.

- [60] 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.
- [61] Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE, 2024.
- [62] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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, 2015.
- [63] Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, et al. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [64] Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, et al. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*, 2022.
- [65] 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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022.
- [66] Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. *arXiv preprint arXiv:2406.05763*, 2024.
- [67] 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.
- [68] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020.
- [69] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

8 Appendix

8.1 Open-Source Audio Datasets for Pretraining

In Table 6, we list all audio caption data used in the pretrain stage.

Table 6: List of datasets used for audio caption in Pretraining stage.

Dataset	Audio Length (Hours)	Dataset	Audio Length (Hours)
CochlScene [23]	169	TACOS [24]	76
TUT2016 [25]	10	VGGSound [26]	513
TUT2017 [25]	13	WavCaps [27]	3,793
TAU2022 [28]	67	ESC50 [29]	1
ESD [30]	29	Nonspeech7k [31]	6
IEMOCAP [32]	10	UrbanSound8K [33]	9
MELD [34]	9	VocalSound [35]	19
RAVDESS [36]	3	FusionAudio [37]	16,646
AudioCaps [38]	137	DCASEAudioQA [39]	57
AudioSet Strong [40]	352	FMA [41]	860
AutoReCap[42]	1,235,000	LP-MusicCaps _{MC} [43]	7
Clotho [44]	17	MusicCaps [45]	7
Clotho-v2 [44]	26	MusicBench [46]	115
CRPO [47]	277	SDD [48]	36
MACS [49]	10	PicoAudio [50]	12
SoundDescs [51]	1,060	ParaSpeechCaps [52]	2,769

8.2 Open-Source Audio Datasets for Post-training

In Table 7, we list all ASR data used in the post-train stage.

Table 7: List of datasets used for ASR in Post-training stage.

Dataset	Language	Audio Length (Hours)
Emilia [53]	Multi	98,305
AISHELL-1 [54]	zh	155
AISHELL-2 [55]	zh	1,036
AISHELL-3 [56]	zh	65
ESD [30]	zh, en	29
Gigaspeech [57]	en	10,288
Hi-Fi TTS [58]	en	291
GLOBE [59]	en	535
LibriTTS [60]	en	568
Libriheavy [61]	en	51,448
LibriSpeech [62]	en	960
KeSpeech [63]	zh	1,428
Magicdata [64]	zh	747
WenetSpeech [65]	zh	10,518
WenetSpeech4TTS [66]	zh	12,085
zhvoice ¹	zh	901
Magpie [67]	en	306
MLS [68]	Multi	45,042
Fleurs [69]	Multi	17

8.3 Prompts for DataFlux

In this section, we present the prompt templates used in the DataFlux pipeline, including those for query generation and answer verification.

Query&Choices Generation

ROLE

You are an expert Audio QA Dataset Curator. Your goal is to create a high-quality "Listening Comprehension" test based on ground-truth audio descriptions.

TASK SPECIFICATION

STEP 1: VERIFICATION

Analyze the provided 'Audio Caption' against the 'Category Hierarchy' (L0-L3).

- Determine if the event described in the caption logically falls under the specific Sub-Task/Category provided.
- **Mismatch Example:** Category is "Animal Sounds > Dog" but Caption describes "A car engine revving". (Result: 'is_success': false)
- **Match Example:** Category is "Animal Sounds > Dog" and Caption describes "A large dog barking aggressively". (Result: 'is_success': true)

STEP 2: GENERATION (Only if Match)

If the caption is consistent, generate **5 distinct Multiple Choice Questions (MCQs)** that a user would answer by **LISTENING to the audio**.

- **Format:** Each item must be a single string containing the **Question** followed by **4 Options (A, B, C, D)** separated by newlines ('\n').
- **Distractors:** Wrong options must be plausible audio events relevant to the context, not obviously ridiculous.
- **Randomness:** Randomize the position of the correct option.
- **Content:** Questions must target the acoustic properties, events, objects, or environment audible in the recording.

CRITICAL CONSTRAINTS (MUST FOLLOW)

1. **AUDIO-CENTRIC PERSPECTIVE:** Questions must be phrased as if the user is currently hearing the sound. Use phrases like "What sound is heard...", "Identify the...", "What action causes the noise...", etc.
2. **FORBIDDEN PHRASES:** You are strictly **FORBIDDEN** from using words that refer to the text, such as "According to the caption", "What does the description say", etc.
3. **NO ANSWERS REVEALED:** Do NOT include the correct answer key or mark the correct option in the output string.
4. **SINGLE STRING FORMAT:** The question and options must be combined into one string using '\n'.

HIERARCHICAL CONTEXT

{ {HIERARCHY_BLOCK} }

INPUT DATA

- **Audio Caption:** "{ {AUDIO_CAPTION} }"

FEW-SHOT EXAMPLES (REFERENCE ONLY)

{ {FEW_SHOT_CONTENT} }

OUTPUT FORMAT

You must output a single valid JSON object. Do not include markdown formatting (like “json).

Structure for Valid Match:

```
{  
  "is_success": true,  
  "questions": [  
    "What is the primary sound heard?  
    A. A dog barking  
    B. A car honking  
    C. People
```

```

talking\nD. Birds chirping",
"Identify the background noise.\nA. Wind blowing\nB. Rain falling\nC. Traffic noise\nD. Silence",
"How would you describe the tone of the voice?\nA. Angry\nB. Happy\nC. Sad\nD. Neutral",
"What action causes the sharp sound?\nA. Dropping a glass\nB. Closing a door\nC. Clapping hands\nD. Stomping feet",
"What sound marks the end of the clip?\nA. A bell ringing\nB. A fade out\nC. A loud crash\nD. An electronic beep"
]
}

**Structure for Mismatch/Conflict:**
{
  "is_success": false,
  "questions": []
}

```

Answer Verification

```

# Role
You will assess the model's logic, factual grounding, and consistency based on specific domain rules.

# Input Data
- **Domain Context:** {L0} > {L1} > {L2}
- **Caption (Ground Truth):** {caption}
- **Query:** {query}

## Model A (Candidate 1)
- CoT: model_a_cot - Answer: model_a_response

## Model B (Candidate 2)
- CoT: model_b_cot
- Answer: model_b_response

# General Instructions
1. **Consensus Check:** Do the models agree on the answer?
2. **Internal Logic Check:** Is the CoT self-consistent?
3. **Domain-Specific Adjudication (CRITICAL):**
Apply the following rules strictly to determine the quality and winner:

> START OF SPECIFIC RULES <
{INJECTION_RULES}
> END OF SPECIFIC RULES <

# Output Format (JSON)
Please output strictly in valid JSON format without markdown code blocks.
{{
  "consensus": true,
  "consistency_with_caption": "STRONG_MATCH",
  "winner": "MODEL_B",
  "reasoning": "string",
  "final_action": "KEEP_USING_WINNER"
}}

```

```
**Valid values for fields:**  

- "consensus": true | false  

- "consistency_with_caption": "STRONG_MATCH" | "WEAK_MATCH" | "CONFLICT"  

- "winner": "MODEL_A" | "MODEL_B" | "TIE" | "NONE"  

- "final_action": "KEEP_USING_WINNER" | "DISCARD"
```

8.4 Prompts for Dense Audio Captioning Evaluation

In this section, we provide the prompts used in the dense audio captioning evaluation. Specifically, we include two types of prompts: (1) a caption generation prompt that instructs the model to produce comprehensive and semantically rich dense audio descriptions, and (2) a question answering prompt that takes the generated caption together with a downstream question as input to produce the final answer. The complete prompt formulations are presented to ensure transparency and reproducibility of the evaluation procedure.

Dense Caption Generation

Please provide a comprehensive description of the input audio, covering all details.

Question Answering

System Role

You are an intelligent **Audio-Based Question Answering Assistant**.

You will receive a textual **[Audio Caption]** generated by a specialized audio analysis model that provides comprehensive details about the audio content.

Core Instructions

1. **Source Reliability**: The [Audio Caption] is your **sole source of truth**. It provides a detailed analysis of the audio, including speech content, speaker characteristics, background sounds, and emotional context.
2. **Mapping Task**: Your goal is to answer the [User Question] by extracting relevant evidence from the caption.
3. **Handling Limitations**: If the question asks for specific details that are not present in the caption, explicitly state that the information is not available in the audio analysis.

Thinking Process

Before answering:

1. **Analyze**: Carefully read the caption to understand all audio elements described.
2. **Verify**: Check if the caption contains sufficient evidence to support your answer.
3. **Reason**: Connect the observed audio elements to answer the question logically.

Input Data

```
[Audio Caption]: "{caption}"
[User Question]: "{question}"
```

Output Format

Provide a direct and concise answer based on the audio caption. If the answer involves multiple-choice options, select the most appropriate one based on the evidence.

8.5 Decoding Throughput Comparison

Hardware Setup. All experiments are conducted on a server equipped with dual Intel Xeon Platinum 8468V processors and NVIDIA H100 GPUs interconnected via NVSwitch. Throughput results are measured using 8x H100 GPU.

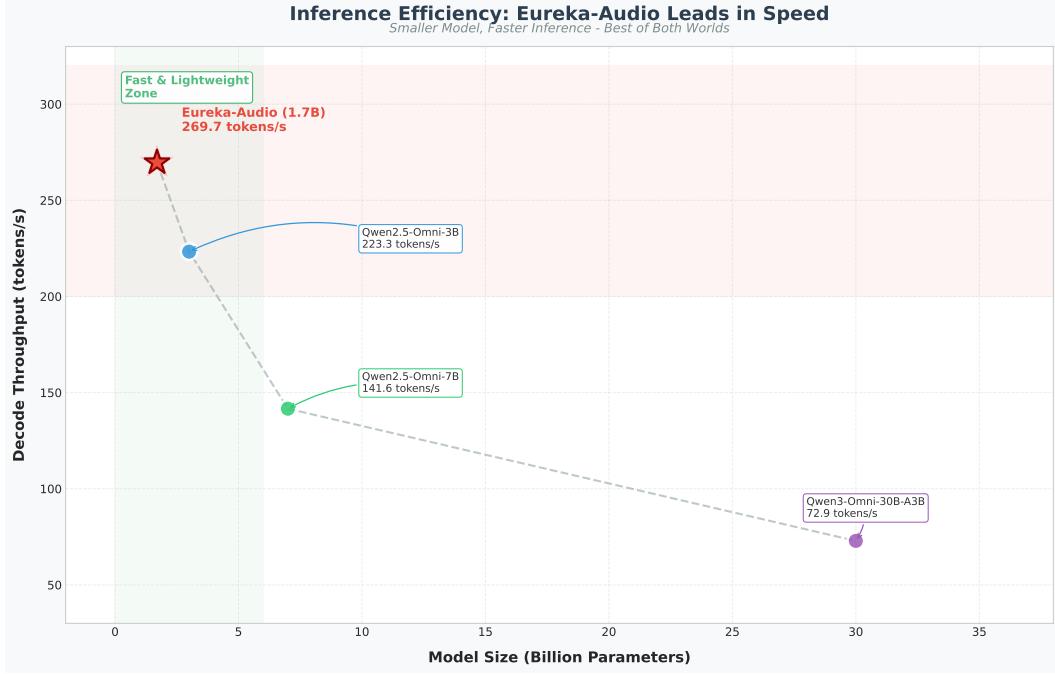


Figure 4: Decode throughput versus model size. Eureka-Audio-Instruct (1.7B) achieves the fastest inference at 269.7 tokens/sec, 3.7× faster than Qwen3-Omni-A3B while being 17× smaller, highlighting its lightweight and efficient design.

Evaluation Protocol. We evaluate decoding throughput on 200 audio samples with a maximum generation length of 2,000 tokens. Throughput is computed as the average number of decoded tokens per second. Figure 4 compares the decode throughput across different model sizes. Eureka-Audio-Instruct (1.7B) achieves the highest throughput of 269.7 tokens/sec, which is 1.2× faster than Qwen2.5-Omni-3B, 1.9× faster than Qwen2.5-Omni-7B, and 3.7× faster than Qwen3-Omni-30B-A3B. With only 1.7B parameters, 17× smaller than the largest baseline—our model delivers superior inference speed, making it well-suited for on-device deployment on mobile phones and edge devices.

Eureka-Audio achieves the highest decoding throughput among all compared models, demonstrating strong inference efficiency despite its compact parameter scale.

8.6 Qualitative Comparison of Dense Audio Captioning

In this section, we present a qualitative comparison of dense audio captions generated by three different models for an audio clip featuring a comedic dialogue from the sitcom *Friends*.

Analysis of Eureka-Audio-Instruct (Ours)

The audio clip presents a fast-paced, comedic dialogue sequence, likely from a sitcom, characterized by a dry, studio-recorded acoustic environment with clear separation between the speakers. The scene opens with a male voice (Ross) delivering a line with a tone of exaggerated resignation and mild frustration: “I want to quit the gym.” A second male voice (Ross’s partner, Chandler) responds with a flat, deadpan “Okay.” Ross continues, his voice sounding slightly more animated as he explains, “Dave in the membership office handles quitters,” followed by a brief pause and a sharp “Excuse me,” before asking, “Are you a member?” The second male responds with a curt, matter-of-fact “No,” and the first voice immediately delivers the punchline with a deadpan inflection: “Sorry, members only.”

The dynamic shifts abruptly as Ross repeats his initial complaint, “I want to quit the gym,” which is met with a stern, authoritative command from the second male: “Be strong.” This triggers a distinct, guttural vocalization from the second male—a deep, resonant groan or grunt of exertion (“Mmm!”). The audio then transitions to a new interaction where Ross asks, “Are you a member of any gym?” The second male responds with a dismissive, arrogant tone, “No,” and adds, “And I’m not going to be, so you can save your little speech.” The soundscape then introduces a third element: a female voice (Marisa) speaking in a polite, professional register, asking, “Could you come in for a second?” The second male hesitates with a low, “Hmm,” before pivoting to a new conversation with the female, asking, “Hi, I’m Maria.” Ross interjects, his voice sounding slightly more casual and conversational, asking, “Hey, so did you go hit?” The second male responds with a defensive, slightly sheepish tone, “No. I almost did. Couldn’t leave Ross there without a spotter.” Ross then asks, “So you joined the gym?” The clip concludes with the second male stating, “They’re gonna take fifty bucks,” delivered with a matter-of-fact, slightly cynical inflection. Throughout the sequence, the dialogue is crisp and foregrounded, with no significant background noise or music interfering with the speech.

Analysis of Qwen3-Omni-30B-A3B-Instruct

It sounds like you’re referencing a scene from the TV show *Friends*! The clip you described is from the episode titled “The One with the Blackout” (Season 8, Episode 1). In this scene, Chandler tries to quit the gym, but he gets sidetracked and ends up joining it instead.

Here’s a breakdown of what happens:

1. Chandler tells Joey that he wants to quit the gym.
2. Joey tells him to talk to Dave, the guy in the membership office who handles quit requests.
3. Chandler goes to talk to Dave, but it turns out Dave is only there to help people join the gym — not quit.
4. Chandler ends up getting persuaded to sign up instead, partly because of the gym’s “no commitment” policy and the friendly, flirty interaction with a gym employee named Maria.
5. Later, Chandler reveals that he didn’t quit — he actually joined. When Monica asks why he didn’t just leave, he says, “I couldn’t leave Ross there without a spotter.”

It’s a classic *Friends* moment combining humor, character dynamics, and a little bit of gym culture satire. Chandler’s reluctance to quit and his tendency to get distracted or talk himself into things is very true to his character. And yes — the gym is “no commitment”, but Chandler still ends up stuck with a membership! Classic Chandler.

Analysis of Qwen3-Omni-30B-A3B-Captioner

The audio clip opens inside a bustling gymnasium, where the constant, rhythmic clatter of weights and equipment underscores a comedic exchange. A male voice, exasperated, announces, “I want to quit the gym,” prompting a calm, matter-of-fact response from a second male: “Okay. Dave in the membership office handles quitters.” Immediately, the first man calls out, “Excuse me! Are you a member?” His question is met with a blunt “No,” followed by the sharp retort, “Sorry. Members

only.” The initial speaker repeats his intent with increased urgency, “I want to quit the gym!”—only to be met with the gym’s slogan, “Be strong!” This is immediately punctuated by a cartoonish “Woo!” and a burst of audience laughter, signaling a comedic beat.

The scene shifts, and the same male voice, now slightly exasperated, asks, “Are you a member of any gym?” Another “No” follows, with the reply, “And I’m not going to be, so you can save your little speech.” A female voice then interjects, “Could you come here for a second?” Her polite introduction, “Hi, I’m Maria,” is delivered with a friendly, inviting tone, but is met with a hesitant, awkward “Hi.” This exchange is again followed by audience laughter, underscoring the comedic timing.

Throughout, the gym’s ambient sounds—weights clanking, equipment humming, and a constant crowd murmur—provide an authentic backdrop. The recording is clean and modern, with each voice clearly separated and audience reactions well-captured. The dialogue and comedic timing are typical of a sitcom sketch. The characters, voices, and setting are consistent with the style and cast of the American TV show *Friends*, with the final segment featuring Ross and Phoebe in a gym conversation that humorously subverts expectations.

Summary: In summary, this audio clip presents a lively, comedic gym scene from a modern American sitcom, characterized by clear dialogue, energetic gym ambiance, and audience laughter. The skit plays on gym culture tropes, featuring a failed attempt to quit, playful banter, and a punchline involving membership fees.