Audio-Omni: Extending Multi-modal Understanding to Versatile Audio Generation and Editing

ZEYUE TIAN, Hong Kong University of Science and Technology, Hong Kong SAR, China
BINXIN YANG, WeChat Vision, Tencent Inc, China
ZHAOYANG LIU, Hong Kong University of Science and Technology, Hong Kong SAR, China
JIEXUAN ZHANG, Peking University, Beijing, China
RUIBIN YUAN, Hong Kong University of Science and Technology, Hong Kong SAR, China
HUBERY YIN, WeChat Vision, Tencent Inc, China
QIFENG CHEN, Hong Kong University of Science and Technology, Hong Kong SAR, China
CHEN LI*, WeChat Vision, Tencent Inc, China
JING LV, WeChat Vision, Tencent Inc, China
WEI XUE*, Hong Kong University of Science and Technology, Hong Kong SAR, China
YIKE GUO, Hong Kong University of Science and Technology, Hong Kong SAR, China

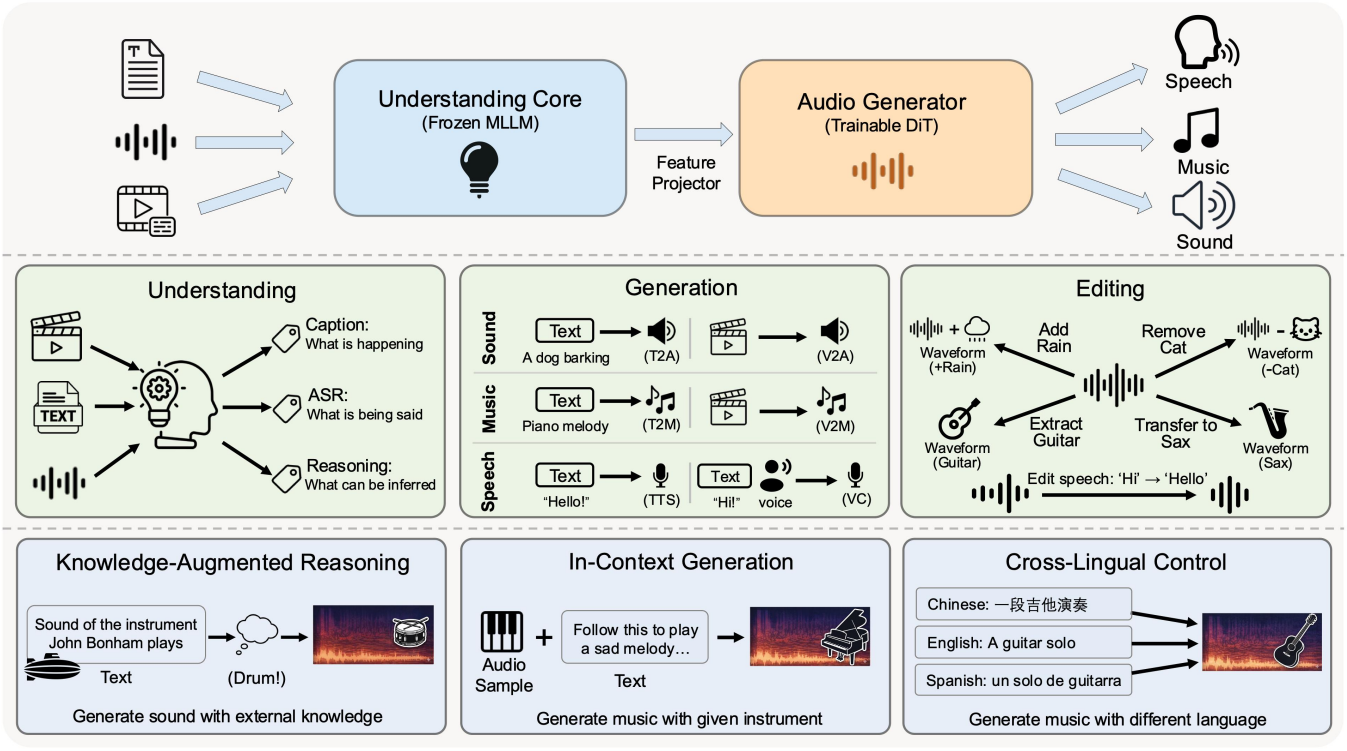

Fig. 1. **An overview of the Audio-Omni framework and its capabilities.** (Top) Our decoupled architecture connects a frozen MLLM for understanding with a trainable DiT for audio synthesis via a feature projector. (Middle) A showcase of the model's unified capabilities across understanding, generation, and editing. (Bottom) A demonstration of remarkable emergent abilities inherited from the MLLM.

Recent progress in multimodal models has spurred rapid advances in audio understanding, generation, and editing. However, these capabilities are typically addressed by specialized models, leaving the development of a truly unified framework that can seamlessly integrate all three tasks underexplored. While some pioneering works have explored unifying audio understanding and generation, they often remain confined to specific domains. To address this, we introduce Audio-Omni, the first end-to-end framework to unify generation and editing across general sound, music, and speech domains, with integrated multi-modal understanding capabilities. Our architecture synergizes a frozen Multimodal Large Language Model for high-level reasoning with a trainable Diffusion Transformer for high-fidelity synthesis. To overcome the critical data scarcity in audio editing, we construct AudioEdit, a new large-scale dataset comprising over one million meticulously curated

*Corresponding authors.

editing pairs. Extensive experiments demonstrate that Audio-Omni achieves state-of-the-art performance across a suite of benchmarks, outperforming prior unified approaches while achieving performance on par with or superior to specialized expert models. Beyond its core capabilities, Audio-Omni exhibits remarkable inherited capabilities, including knowledge-augmented reasoning generation, in-context generation, and zero-shot cross-lingual control for audio generation, highlighting a promising direction toward universal generative audio intelligence. The code, model, and dataset will be publicly released on https://zeyuet.github.io/Audio-Omni.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; *Computer graphics*.

Additional Key Words and Phrases: Audio generation, multimodal learning, diffusion models, unified models, audio editing

1 Introduction

Recent advances in multimodal learning have spurred a trend toward unified frameworks that integrate both understanding and generation within a single model, achieving significant progress in visual domains such as image [Chen et al. 2025b; Jiao et al. 2025; Ma et al. 2025; Pan et al. 2025] and video [Liu et al. 2025a; Wei et al. 2025] understanding and generation. However, the audio domain remains comparatively underexplored.

Unlike the visual modality, audio encompasses three distinct domains with significant distributional disparities: general sounds, music, and speech. While some efforts have been made to unify audio understanding and generation, they remain confined to specific domains, such as speech-centric models [AI et al. 2025; An et al. 2024] or those limited to general audio and music [Liu et al. 2024a; Tian et al. 2025b], failing to cover the full audio spectrum. Other systems rely on tool-based integration [Huang et al. 2024a], which lacks end-to-end optimization. Meanwhile, existing audio editing models [Manor and Michaeli 2024; Wang et al. 2023b] are designed exclusively for editing and cannot be extended to understanding or generation, leaving the unification of all three capabilities an open challenge.

To address these limitations, we introduce Audio-Omni, a framework that unifies audio understanding, generation, and editing across the full spectrum of audio domains. We adopt a decoupled design: a frozen Multimodal Large Language Model (MLLM) serves as the reasoning core, while a trainable Diffusion Transformer (DiT) handles generation and editing. Keeping the MLLM frozen preserves its rich multimodal knowledge, which in turn empowers the generative module with capabilities beyond its explicit training scope. To effectively bridge the two components, we design a hybrid conditioning mechanism that disentangles inputs into two complementary streams: a **High-Level Semantic** stream, combining MLLM features and text embeddings for speech synthesis, injected via cross-attention to provide instructional guidance; and a **Low-Level Signal** stream, fusing mel-spectrogram and video sync features, concatenated with the input noise for precise temporal control. This separation is key to mastering the diverse requirements of sound, music, and speech within a single framework.

Training a unified model of this scope demands a comprehensive and diverse dataset. A critical barrier to progress in instruction-guided audio editing is the absence of any large-scale, publicly available dataset. To address this gap, we meticulously design a pipeline to construct AudioEdit, a large-scale, high-quality dataset for this task. Created through a systematic pipeline combining real-world data mining with scalable programmatic synthesis, AudioEdit contains over 1M rigorously curated samples covering editing tasks including addition, removal, extraction, and style transfer. Training on this dataset enables our model with its robust editing capabilities.

Extensive experiments validate the effectiveness of our unified design. Audio-Omni outperforms prior unified models across understanding, generation, and editing tasks, while matching or surpassing specialized expert models on several tasks. Beyond these core results, the generative module naturally inherits capabilities from the frozen MLLM, including world knowledge for reasoning-based generation, in-context learning for audio-conditioned synthesis, and multilingual understanding for cross-lingual control.

In summary, our main contributions are as follows:

- We propose Audio-Omni, the first unified framework for audio understanding, generation, and editing across general sound, music, and speech. At its core is a decoupled architecture that bridges a frozen MLLM with a trainable DiT, guided by a hybrid conditioning mechanism to disentangle semantic and signal-level control.
- We introduce AudioEdit, a large-scale, high-quality dataset with a meticulous pipeline for instruction-guided audio editing, encompassing a wide range of tasks to facilitate future research in this area.
- Extensive experiments demonstrate that Audio-Omni outperforms prior unified models and achieves competitive or superior results compared to specialized expert models across a broad range of tasks.
- Furthermore, Audio-Omni exhibits remarkable inherited capabilities for generation, such as knowledge-augmented generation and cross-lingual control, highlighting a path toward intelligent and versatile generative audio systems.

2 Related Work

Specialized Models in the Audio Domain. In recent years, the field of audio processing has achieved significant advancements, particularly with the advent of large-scale pre-trained models. For audio understanding, a substantial body of work has produced powerful models capable of tackling complex tasks across all three audio domains [Chu et al. 2024, 2023; Ghosh et al. 2025; Kong et al. 2024; Tang et al. 2024]. In contrast to the comprehensive domain coverage in understanding, generative models have typically specialized. In the general audio domain, prominent text-to-audio models like [Evans et al. 2024; Ghosal et al. 2023; Huang et al. 2023; Majumder et al. 2024] have demonstrated high-fidelity synthesis from textual descriptions, while other approaches have explored generation conditioned on video [Cheng et al. 2025; Liu et al. 2025c, 2024b; Xing et al. 2024]. Similarly, in the music domain, models like [Chen et al. 2024; Copet et al. 2024; Deng et al. 2024; He et al. 2024; Melechovsky et al. 2024; Yinghao et al. 2024; Yuan et al. 2025, 2024] excel at text-to-music generation, and dedicated research has addressed video-to-music generation [Lin et al. 2024; Su et al. 2023; Tian et al. 2025c; Xie et al. 2025a]. The speech domain has its own rich landscape of research, with powerful text-to-speech (TTS) systems [Chen et al. 2025a;

Deng et al. 2025; Du et al. 2024] and models that support complex transformations like voice conversion (VC) [Peng et al. 2024; Wang et al. 2023a]. While some recent models demonstrate capabilities in handling multimodal conditions or generating audio across multiple domains [Liu et al. 2025b; Rong et al. 2025; Tian et al. 2025a; Wang et al. 2025a]. Meanwhile, the field of audio editing remains underexplored, largely due to a scarcity of large-scale, instruction-guided datasets. Existing approaches fall into two main categories. Training-free, zero-shot methods adapt pre-trained models but often struggle with preserving non-target content and following precise instructions [Manor and Michaeli 2024]. Training-based methods, such as [Lan et al. 2025; Tao et al. 2025; Wang et al. 2023b], construct synthetic data pipelines but introduce a significant domain gap, as their strategy of mixing isolated audio segments fails to capture the acoustic complexities of real-world editing scenarios.

Unified Multimodal Models. Recently, a major trend in multimodal research has been the development of unified models that aim to handle both understanding and generation within a single framework. This unification has seen remarkable success in visual domains, with models like [Chen et al. 2025b; Jiao et al. 2025; Ma et al. 2025; Pan et al. 2025] achieving unified image understanding and generation, and others extending these capabilities to the video domain [Liu et al. 2025a; Wei et al. 2025]. Some frameworks even support generation across different modalities, such as text, images, audio, and video [Lu et al. 2024; Wu et al. 2024; Zhan et al. 2024]. In the audio field, several pioneering works [AI et al. 2025; An et al. 2024; Huang et al. 2024b; Liu et al. 2024a] have also moved towards this unified goal. However, they exhibit critical limitations. Some approaches rely on orchestrating separate expert models via tool invocation [Huang et al. 2024b], which lacks the benefits of end-to-end optimization. Other models, while more integrated, typically focus on a limited subset of audio domains, such as speech-only [AI et al. 2025; An et al. 2024] or music-only generation [Liu et al. 2024a], failing to provide a truly universal solution. To address these shortcomings, we propose Audio-Omni, the first framework to unify understanding, generation, and editing across all general sound, music, and speech domains. Our approach provides a cohesive, end-to-end solution that overcomes the fragmentation of specialized models and the domain limitations of existing unified efforts.

3 Audio Editing Dataset Construction

Table 1. Statistics of our proposed audio editing dataset AudioEdit.

Task Type	Data Source	Train Samples	Test Samples
Add	Real Data	50K	500
	Synthesis Data	150K	-
Remove	Real Data	50K	500
	Synthesis Data	150K	-
Extract	Real Data	50K	500
	Synthesis Data	150K	-
Style Transfer	Real Data	500K	500
Total		**1,100K**	**2,000**

An obstacle to advancing instruction-guided audio editing is the scarcity of large-scale paired datasets. While pioneering works like [Lan et al. 2025; Tao et al. 2025; Wang et al. 2023b] have proposed synthetic pipelines, their reliance on mixing isolated audio segments introduces a significant domain gap from real-world audio, where sounds are integrated. Furthermore, these methods often falter on complex tasks like audio style transfer, struggling to disentangle acoustic style from core content attributes such as temporal structure and pitch. To address these limitations, we introduce AudioEdit, a large-scale dataset constructed via a novel hybrid pipeline that integrates real-world data mining with scalable synthesis.

Our pipeline, illustrated in Figure 2, features two complementary branches to generate a dataset with both real-world acoustic fidelity and large-scale diversity. **The Real Data Branch** mines authentic editing pairs from real-world datasets VGGSound [Chen et al. 2020]. *First*, we employ a powerful MLLM (Gemini 2.5 Pro) to identify the primary sound-emitting object categories within each source audio clip. *Then*, we deploy SAM-Audio [Shi et al. 2025], a state-of-the-art audio segmentation model, to perform source separation based on the identified categories. This step disentangles the source audio into a target track, containing the audio of the specified object, and a corresponding residual track. *Subsequently*, all separated tracks undergo rigorous multi-stage filtering to ensure quality. Starting from over 540K category-labeled samples, we apply VAD[1] filtering (retaining ∼347K pairs) followed by CLAP [Elizalde et al. 2023]-based semantic alignment (retaining ∼50K pairs, approximately 9.2% overall retention). Human validation on a subset achieves approximately 83% agreement, confirming pipeline quality. This process yields high-quality source-target pairs for `add`, `remove`, and `extract` tasks. For `style transfer`, we expand the filtered targets by prompting Gemini to generate semantically related but different keywords (prompt template in Appendix 1.1), then apply CLAP filtering again, yielding approximately 500K pairs. We use ZETA [Manor and Michaeli 2024] to transform each target to the new style while preserving temporal structure and pitch, then mix the transformed audio back with the residual track to form the final edited output. Concurrently, our **Synthesis Data Branch** ensures scale and diversity by programmatically generating soundscapes with the Scaper toolkit [Salamon et al. 2017]. This is achieved by randomly mixing foreground events from ESC-50 [Piczak 2015] into 10-second backgrounds from AudioCaps [Kim et al. 2019], applying randomized parameters including onset time, SNR (0-3 dB), pitch shifts (-3 to +3 semitones), and time-stretch factors (0.8-1.2). This automated process efficiently yields a large volume of precisely annotated data for our `add`, `remove`, and `extract` tasks.

Through this meticulously designed hybrid pipeline, we construct AudioEdit, a large-scale, high-quality dataset for instruction-guided audio editing. In total, AudioEdit comprises over 1M samples covering `add`, `remove`, `extract`, and `style transfer` audio editing tasks, with detailed statistics presented in Table 1.

4 Method

In this section, we describe the architecture of Audio-Omni and its training strategies.

[1] https://github.com/jiaaro/pydub

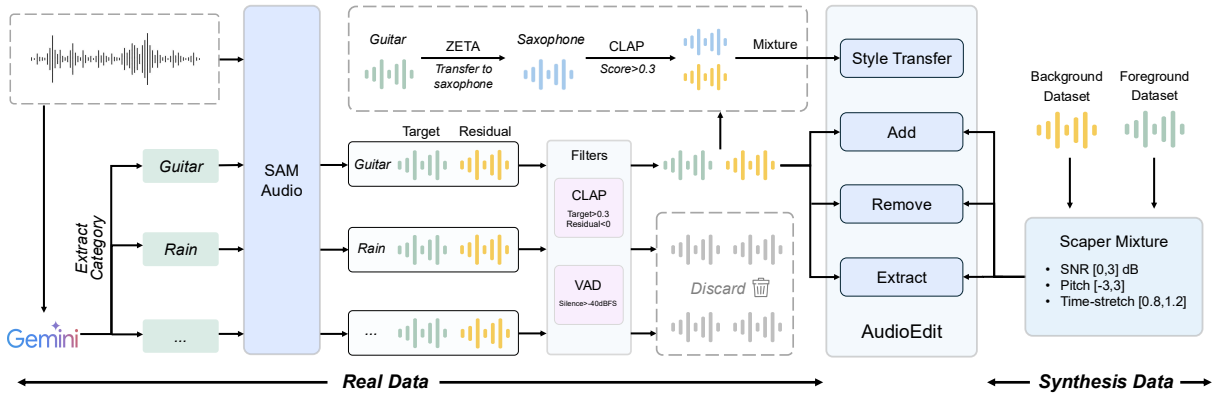

Fig. 2. **Overview of the hybrid pipeline for constructing our AudioEdit dataset.** The pipeline consists of two parallel branches to ensure both data authenticity and scale. The Real Data Branch (left) mines editing pairs from real-world datasets (e.g., VGGSound) by first using an MLLM (Gemini) for category identification, followed by a dedicated segmentation model (SAM-Audio) for source separation. Concurrently, the Synthesis Data Branch (right) leverages the Scaper toolkit to programmatically generate a large volume of precisely annotated editing scenarios. This hybrid strategy yields a dataset that combines the acoustic fidelity of natural audio with the large-scale diversity needed for robust model training.

4.1 Preliminary

Rectified Flow. Our generative backbone is built upon the framework of Rectified Flow [Liu et al. 2022], a powerful class of generative models. Unlike traditional diffusion models that often follow stochastic paths, Rectified Flow simplifies the generation process by modeling a straight-line trajectory between noise and data. This trajectory connects a random noise sample $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a data sample $\mathbf{x}_0$ via a simple ordinary differential equation (ODE) with a constant velocity field $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$. The ODE is defined for a time variable $t \in [0, 1]$ as:

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}. \tag{1}$$

The solution at any time t along this linear path is given by the interpolation:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1. \tag{2}$$

A neural network, denoted $v_\theta(\mathbf{x}_t, t, \mathbf{c})$, is trained to predict this velocity field $\mathbf{v}$ conditioned on the noisy state $\mathbf{x}_t$, time t, and a set of conditioning signals $\mathbf{c}$. During inference, generation starts from a random noise sample $\mathbf{x}_1$ at $t = 1$, and the ODE in Equation 1 is solved backwards to $t = 0$ using a numerical solver, guided by the predictions of v_θ. The training objective for v_θ is detailed in Section 4.3.

4.2 Model Architecture

As illustrated in Figure 3, our Audio-Omni framework is built upon a decoupled architecture comprising a frozen MLLM for understanding and a trainable DiT-based backbone for versatile audio generation and editing.

Understanding Module. Our understanding module is a pre-trained and frozen MLLM, which serves as the primary reasoning and understanding core of our framework. It processes a diverse set of inputs, including a textual instruction ($\mathbf{T}_{\text{in}}$), an audio waveform ($\mathbf{A}_{\text{in}}$), and a video ($\mathbf{V}_{\text{in}}$), after they are tokenized by their respective encoders. The MLLM fulfills a dual role within our framework. For *understanding tasks*, it directly generates textual responses based on the multimodal inputs. For *generative tasks*, its crucial function is to produce a powerful conditioning signal for the generative module. To this end, we extract the hidden states from its penultimate layer (i.e., the second-to-last hidden state), as we empirically find that this provides a more generalizable and richer semantic representation for downstream generative tasks compared to the final layer (see Sec. 5.4). We denote this multimodal feature as $\mathbf{F}_{\text{mm}} \in \mathbb{R}^{L_{\text{mm}} \times D_{\text{mm}}}$.

Generative Module. The generative module consists of a DiT backbone, trained with a Rectified Flow objective, which synthesizes the final audio waveform. This module is conditioned on two distinct streams of features, allowing it to handle a wide array of tasks.

The first stream, which we term *High-Level Semantic Features*, provides the primary instructional signal for the synthesis process. This stream is formed by concatenating the multimodal features $\mathbf{F}_{\text{mm}}$ from the MLLM with a transcript-derived feature $\mathbf{F}_{\text{trans}}$ for speech-related tasks:

$$\mathbf{c}_{\text{high}} = \text{Concat}(\mathbf{F}_{\text{mm}}, \mathbf{F}_{\text{trans}}). \tag{3}$$

The feature $\mathbf{F}_{\text{trans}}$ is produced by a Transcript Encoder, which performs a character-level encoding of the input transcript $\mathbf{T}_{\text{trans}}$ (for TTS/VC tasks). Specifically, it first converts the text into a sequence of character embeddings, which are then processed by a ConvNeXtV2-based architecture.

The second stream, termed *Low-Level Signal Features*, provides concrete, temporally-aligned references crucial for editing and synchronization tasks. This stream is formed by concatenating features from a reference audio and a video sync signal:

$$\mathbf{c}_{\text{low}} = \text{Concat}(\mathbf{F}_{\text{sync}}, \mathbf{F}_{\text{mel}}). \tag{4}$$

The mel-spectrogram feature, $\mathbf{F}_{\text{mel}}$, is extracted by a Mel Encoder from either a reference audio $\mathbf{A}_{\text{ref}}$ (for editing) or a speech prompt $\mathbf{S}_{\text{prompt}}$ (for voice conversion). The synchronization feature, $\mathbf{F}_{\text{sync}}$, is extracted from the input video $\mathbf{V}_{\text{in}}$ by a pre-trained Synchformer [Iashin et al. 2024] model.

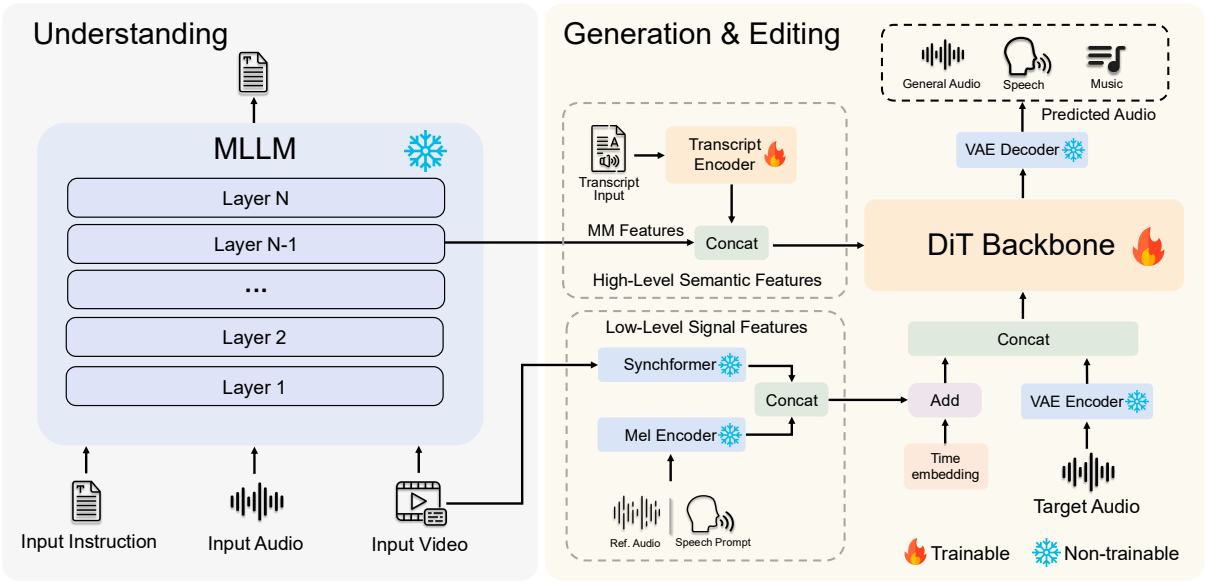

Fig. 3. **The Audio-Omni Framework.** Our framework utilizes a decoupled design with two distinct conditioning streams to guide a trainable DiT backbone. The High-Level Semantic Features stream provides global, instructional guidance. It is formed by concatenating features from a frozen MLLM (MM Features) with character-level embeddings from a trainable Transcript Encoder. The Low-Level Signal Features stream offers precise, temporal guidance for editing and synchronization. It combines features from Synchformer and Mel Encoder. These two streams are injected into the DiT via different mechanisms: the high-level stream as context for cross-attention, and the low-level stream concatenated with the input noise and time embedding.

These two conditioning streams are injected into the DiT backbone through distinct mechanisms tailored to their properties. The *High-Level Semantic Features* ($\mathbf{c}_{\text{high}}$) are injected as context via cross-attention, allowing the model to flexibly query abstract instructions at each step of the generation process. In contrast, the *Low-Level Signal Features* ($\mathbf{c}_{\text{low}}$) are first fused with a time embedding via element-wise addition, and this combined signal is then concatenated with the VAE-encoded noisy audio latent ($\mathbf{x}_t$) to form the main input to the DiT. This concatenation provides strong, frame-by-frame guidance, ensuring precise alignment for editing and synchronization tasks. The DiT backbone then processes this composite input to predict the velocity field, and its final denoised latent output is passed to a VAE Decoder to reconstruct the audio waveform.

4.3 Training Objective

We train our Audio-Omni end-to-end using a single, unified loss based on the Rectified Flow objective. For each training sample, we aim to train the network v_θ to predict the constant velocity $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$, where $\mathbf{x}_0$ is the VAE-encoded latent of the target audio and $\mathbf{x}_1$ is a random noise sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Following the standard RF training procedure, we first sample a random timestep t from a uniform distribution $\mathcal{U}(0, 1)$. This timestep is used to create an interpolated latent state $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. The training objective is then to minimize the mean squared error between the network's predicted velocity and the ground-truth velocity $\mathbf{v}$. The loss function, $\mathcal{L}$, is defined as:

$$\mathcal{L} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\mathbf{x}_0,\mathbf{x}_1,\mathbf{c}}\left[||v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)||^2\right] \tag{5}$$

Table 2. Quantitative results on multimodal understanding benchmarks.

Methods	MMSU↑	MMAU↑
Specialized Models		
Audio Flamingo3 [Goel et al. 2025]	**61.40**	**72.42**
Qwen2-Audio-Instruct [Chu et al. 2024]	53.27	57.40
Qwen2.5-Omni-3B [Xu et al. 2025]	56.83	63.30
Unified Models		
Ming-Omni [AI et al. 2025]	47.53	70.80
Unified-IO2 [Lu et al. 2024]	30.74	3.20
MuMuLLaMA [Liu et al. 2024a]	6.32	12.87
Audio-Omni	56.83	63.30

where $\mathbf{c}$ represents the full set of conditioning signals available for the given training sample.

5 Experiments

5.1 Implementation Details

Model Architecture. For the MLLM, we use a pre-trained Qwen2.5-Omni-3B model [Xu et al. 2025], which processes input videos at 5 fps and audio at 16 kHz. We extract features from its penultimate layer as the primary semantic condition. The DiT backbone is a transformer architecture with a depth of 36 blocks, a hidden dimension of 2048, and 32 attention heads. The entire model has approximately 7.9B total parameters, of which 3.05B (the DiT and conditioners) are trainable. Our conditioning modules include a Synchformer [Iashin et al. 2024] that extracts synchronization features from videos at 25 fps; a Mel Encoder that computes 100-dimensional

Table 3. Quantitative results on multimodal generation benchmarks.

Methods	T2A_FAD↓	T2M_FAD↓	V2A_FAD↓	V2M_FAD↓	TTS_WER↓
Specialized Models					
Tango2 [Majumder et al. 2024]	3.20	-	-	-	-
MMAudio [Cheng et al. 2025]	4.71	-	2.04	-	-
Stable-Audio-Open [Evans et al. 2024]	3.15	3.23	-	-	-
MusicGen [Copet et al. 2024]	-	3.94	-	-	-
VATT [Liu et al. 2024b]	-	-	2.55	-	-
AudioX [Tian et al. 2025a]	**1.86**	**1.53**	**1.13**	2.12	-
VidMuse [Tian et al. 2025c]	-	-	-	2.46	-
F5-TTS [Chen et al. 2025a]	-	-	-	-	1.83
MaskGCT [Wang et al. 2024]	-	-	-	-	2.62
CosyVoice3 [Du et al. 2025]	-	-	-	-	2.46
Unified Models					
Ming-Omni [AI et al. 2025]	-	-	-	-	4.31
Unified-IO2 [Lu et al. 2024]	7.81	3.17	-	-	21.63
MuMuLLaMA [Liu et al. 2024a]	-	5.89	-	52.25	-
Audio-Omni	**1.86**	1.94	1.71	**1.58**	**1.77**

Table 4. Quantitative results on audio editing benchmarks.

Methods	AE_FAD↓	AE_LSD↓	CLAP↑
Specialized Models			
ZETA [Manor and Michaeli 2024]	3.81	3.80	0.30
SDEdit [Meng et al. 2021]	3.51	4.40	0.22
MMEDIT [Tao et al. 2025]	3.95	4.05	0.15
Unified Models			
Audio-Omni	**3.27**	**2.27**	**0.32**

mel-spectrograms (44.1 kHz, FFT size 1024, hop 256); and a Transcript Encoder composed of four ConvNeXtV2 blocks for character-level text encoding. The final audio is encoded and decoded by a pre-trained VAE adopted from [Evans et al. 2024].

Datasets. The training data sources and their scale are summarized in Appendix Table A4.

Training Details. We train the DiT backbone and all learnable conditioning modules for approximately 80k steps on a total batch size of 5120. We use the AdamW optimizer with a learning rate of 5e-5, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1e-3. The entire model is trained end-to-end using the Rectified Flow objective (Equation 5). To enhance voice conversion and speech editing capabilities, we employ a masking strategy on the *speech prompt* during training following [Chen et al. 2025a]. Specifically, we randomly mask 20% to 75% of the prompt's mel-spectrogram, forcing the model to infer the global speaker timbre from a partial acoustic signal while reconstructing the full utterance using the complete transcript. This scheme is crucial for developing the model's robust zero-shot voice cloning and content editing abilities.

Inference Details. At inference time, we use a numerical ODE solver with 100 steps to generate the audio latent from a random noise sample. For conditional generation, we employ classifier-free guidance with a guidance scale of 6.0.

5.2 Evaluation Metrics

To evaluate Audio-Omni, we employ a comprehensive suite of metrics for its understanding, generation, and editing capabilities. For understanding, we report scores on the MMSU [Wang et al. 2025b] and MMAU [Sakshi et al. 2024] benchmarks to assess multi-task reasoning. For generation, we measure distributional similarity and quality using KL divergence, Inception Score (IS), Fréchet Audio Distance (FAD), and Fréchet Distance (FD) on PANNs [Kong et al. 2020] embeddings. Speech synthesis performance is specifically evaluated using Word Error Rate (WER). For editing, we use Log-Spectral Distance (LSD) to measure fidelity and content preservation, and FAD to assess perceptual quality.

5.3 Main Results

In this section, we present quantitative and qualitative results for Audio-Omni, evaluating its performance on understanding, generation, and editing tasks against a wide range of specialized and unified models.

5.3.1 Overall Performance. As shown in Table 2, Table 3, and Table 4, Audio-Omni demonstrates strong and comprehensive capabilities across understanding, generation, and editing tasks.

Understanding. Benefiting from our decoupled architecture, Audio-Omni inherits a strong understanding ability from the frozen MLLM core, whose audio encoder has been pre-trained on a large-scale audio-related data, enabling strong performance on multi-domain audio tasks [Xu et al. 2025]. We evaluate on MMSU (covering 47 spoken-language tasks) and MMAU (evaluating 27 reasoning skills across sound, music, and speech). As shown in Table 2, Audio-Omni achieves strong understanding performance, outperforming most unified models and approaching the level of dedicated understanding specialists.

Generation and Editing. For generation tasks, Audio-Omni exhibits state-of-the-art or highly competitive performance. We report the FAD on standard test sets: AudioCaps for T2A, Musicaps for T2M, VGGSound for V2A, and V2M-bench for V2M. For speech synthesis, we evaluate WER on the Seed-TTS [Anastassiou et al. 2024] en benchmark. In generation tasks, our model demonstrates strong overall performance, consistently and significantly outperforming all other unified models. Notably, it further surpasses specialized expert models in T2M and TTS. For editing, we report the average FAD and LSD across four tasks (`add`, `remove`, `extract`, `style transfer`) on our proposed AudioEdit test set. We additionally report the CLAP score (averaged over `add`, `extract`, and `style transfer`) to evaluate instruction adherence. As shown in Table 4, Audio-Omni achieves the best performance on all metrics, with detailed per-task results in Appendix Table A3.

In summary, Table 2, Table 3, and Table 4 validate the effectiveness of our unified framework: Audio-Omni inherits strong understanding capabilities from the frozen MLLM while achieving state-of-the-art or highly competitive results across generation and editing tasks, demonstrating that a single unified model can serve as a strong generalist across the full spectrum of audio tasks.

5.3.2 Zero-shot Cross-lingual Text-to-Audio Generation. A remarkable ability of our framework is its zero-shot cross-lingual generation, inherited from the frozen MLLM's multilingual understanding. As shown in Appendix Table A2, Audio-Omni maintains strong performance across multiple languages (CN, ES, DE, FR, JP) despite being trained on English, with quality comparable to English-only specialist models.

5.3.3 Inherited Abilities and Zero-Shot Capabilities. We further highlight several representative capabilities of Audio-Omni. The first two are emergent abilities inherited from the frozen MLLM's world knowledge and in-context reasoning, while the latter two are enabled by our masking-based training strategy. We qualitatively showcase these in Figure 4.

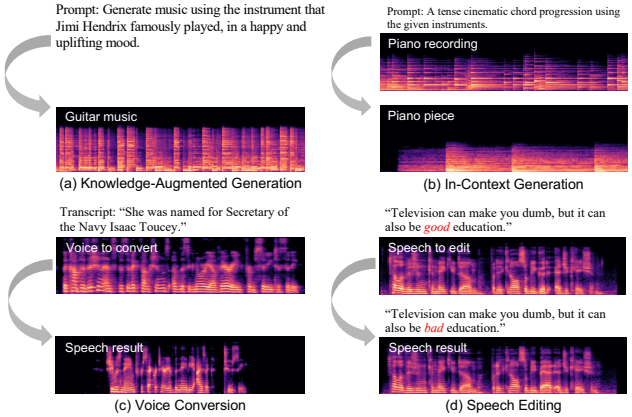

Fig. 4. **Qualitative showcase of Audio-Omni's capabilities**, including (a) knowledge-augmented generation, (b) in-context generation, (c) zero-shot voice conversion, and (d) zero-shot speech editing.

Knowledge-Augmented Generation. Audio-Omni successfully handles knowledge-intensive prompts that require external world knowledge. For instance, when prompted with "*Generate music using the instrument Jimi Hendrix played in a happy mood.*" it correctly infers the instrument is an electric guitar and synthesizes a corresponding melody, a task beyond standard text-to-audio models.
In-Context Generation. Our model demonstrates strong in-context learning. Given a piano recording and the instruction "*Generate a dramatic cinematic chord progression that builds tension*," it extracts the piano's timbre and applies it to the newly synthesized piece.
Zero-Shot Voice Conversion and Speech Editing. During speech training, we randomly mask either speaker identity or speech content, forcing the model to infer the missing component from context. This naturally enables zero-shot voice conversion and speech editing at inference time without task-specific supervision.

These capabilities demonstrate that Audio-Omni goes beyond a standard multi-task model, exhibiting inherited intelligence and flexible zero-shot control across diverse audio generation scenarios.

5.4 Ablation Studies

To validate our design choices, we conduct a series of ablation studies in this section, more ablation studies are included in the Appendix 1.3.

Ablation on Dataset Composition. To assess the impact of our data sources, we evaluate three training configurations: using only synthetic data, only real-world data, or a mix of both. As shown in Table 5, the mixed-data approach yields the best overall performance. This suggests a synergy where synthetic data provides broad coverage of diverse editing operations, while real-world data contributes crucial acoustic realism and fidelity. Notably, training on synthetic data alone is insufficient for achieving robust generalization to the acoustic complexities of real-world audio.

Table 5. Ablation study on dataset composition for audio editing training.

Setting	KL↓	IS↑	FD↓	FAD↓	LSD↓
Only Real	**1.27**	5.54	20.69	2.67	1.84
Only Synthesis	1.93	5.04	37.96	3.80	5.17
Syn. + Real (Ours)	1.30	**5.93**	**20.48**	**2.48**	**1.82**

Table 6. Ablation study on conditioning injection strategies.

Context	Cat.	T2A↓	V2A↓	TTS↓	AE↓
`mm, trans, sync, mel`	*none*	40.91	25.61	32.46	5.47
`mm, trans, sync`	`mel`	38.26	26.46	25.88	4.03
`mm, trans`	**`sync, mel`**	**28.90**	**18.55**	**19.33**	**3.62**
`mm`	`trans, sync, mel`	60.58	56.88	59.20	4.88

Ablation on Conditioning Strategies. We perform an ablation study to identify the optimal method for integrating our multimodal conditions. We compare four distinct injection strategies by varying how features including multimodal (mm), transcript (trans), synchronization (sync), and mel-spectrogram (mel) are delivered to the DiT backbone, either via cross-attention (Context) or channel-wise concatenation (Cat). As shown in Table 6, the results consistently demonstrate the superiority of one particular configuration across all tasks. This optimal strategy involves providing high-level conditions (mm, trans) as flexible Context, while simultaneously concatenating low-level conditions (sync, mel) with the input noise. This result provides the insight for future research on the optimal conditioning strategy for unified audio generation.

Table 7. Ablation study on the source of conditional features from the MLLM.

Feature Source	T2A		T2M	
	IS↑	FD↓	IS↑	FAD↓
Last Layer (-1)	9.36	4.21	2.90	3.26
Penultimate Layer (-2)	**11.26**	**2.75**	**3.46**	**3.05**
MetaQuery	7.44	8.34	2.38	8.15
Query	8.55	5.31	2.69	4.22

Ablation on Feature Source Selection. We ablate four strategies for extracting conditional features from the frozen MLLM, with results on T2A and T2M tasks shown in Table 7. The methods include using the Last Layer (-1) embeddings, the Penultimate Layer (-2) embeddings, MetaQuery (which appends learnable tokens to the input sequence, following [Pan et al. 2025]), and Query mechanism (which uses learnable tokens to attend to the penultimate layer features via cross-attention). Our results indicate that for audio generative tasks, using the unfiltered feature sequence from the penultimate layer is the most effective conditioning strategy. The

penultimate layer's superiority confirms that the final layer is overly specialized for text prediction, whereas the penultimate layer retains richer, uncompressed semantic and acoustic details. Furthermore, complex query mechanisms proved detrimental, suggesting that high-fidelity audio synthesis is sensitive to information bottlenecks and benefits most from direct access to dense features.

6 Conclusion

In this work, we introduced Audio-Omni, the first end-to-end framework to unify audio understanding, generation, and editing across the full spectrum of sound, music, and speech. Our decoupled architecture successfully synergizes a frozen MLLM for high-level reasoning with a trainable DiT, guided by a hybrid conditioning mechanism that separates high-level semantic and low-level signal features. To address a critical data bottleneck, we also constructed AudioEdit, a large-scale dataset of over one million instruction-guided editing pairs. Extensive experiments demonstrate that our single unified model not only matches or surpasses the performance of specialized expert models but also exhibits remarkable inherited abilities, including knowledge-augmented reasoning and zero-shot cross-lingual control, inherited from the MLLM. We believe Audio-Omni provides a powerful and scalable baseline that highlights a promising path toward universal generative audio intelligence.

References

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. 2025. Ming-Omni: A Unified Multimodal Model for Perception and Generation. *arXiv preprint arXiv:2506.09344* (2025).

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051* (2024).

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430* (2024).

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2025. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing* (2025).

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1206–1210.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811* (2025).

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025a. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6255–6271.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 28901–28911.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759* (2024).

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* (2023).

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems* 36 (2024).

Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, et al. 2024. Composerx: Multi-agent symbolic music composition with llms. *arXiv preprint arXiv:2404.18081* (2024).

Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. 2025. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512* (2025).

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407* (2024).

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589* (2025).

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024. Stable audio open. *arXiv preprint arXiv:2407.14358* (2024).

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731* (2023).

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S. Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net. https://openreview.net/forum?id=xWu5qpDK6U

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128* (2025).

Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, et al. 2024. Llms meet multimodal generation and editing: A survey. *arXiv preprint arXiv:2405.19334* (2024).

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 366–370.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474* (2023).

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024a. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23802–23804.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024b. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 23802–23804. doi:10.1609/AAAI.V38I21.30570

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5325–5329.

Siyu Jiao, Yiheng Lin, Yujie Zhong, Qi She, Wei Zhou, Xiaohan Lan, Zilong Huang, Fei Yu, Yingchen Yu, Yunqing Zhao, et al. 2025. ThinkGen: Generalized Thinking for Visual Generation. *arXiv preprint arXiv:2512.23568* (2025).

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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)*. 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=WYi3WKZjYe

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352* (2022).

Zitong Lan, Yiduo Hao, and Mingmin Zhao. 2025. Guiding audio editing with audio language model. *arXiv preprint arXiv:2509.21625* (2025).

Yan-Bo Lin, Yu Tian, Linjie Yang, Gedas Bertasius, and Heng Wang. 2024. VMAS: Video-to-Music Generation via Semantic Alignment in Web Music Videos. *arXiv preprint arXiv:2409.07450* (2024).

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. 2025c. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *arXiv preprint arXiv:2506.21448* (2025).

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, et al. 2025a. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *arXiv preprint arXiv:2512.22905* (2025).

Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, and Ying Shan. 2024a. MuMu-LLaMA: Multi-modal Music Understanding and Generation via Large Language Models. *arXiv preprint arXiv:2412.06660* (2024).

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).

Xiulong Liu, Kun Su, and Eli Shlizerman. 2024b. Tell What You Hear From What You See–Video to Audio Generation Through Text. *arXiv preprint arXiv:2411.05679* (2024).

Zhenyu Liu, Yunxin Li, Xuanyu Zhang, Qixun Teng, Shenyuan Jiang, Xinyu Chen, Haoyuan Shi, Jinchao Li, Qi Wang, Haolan Chen, et al. 2025b. UniMoE-Audio: Unified Speech and Music Generation with Dynamic-Capacity MoE. *arXiv preprint arXiv:2510.13344* (2025).

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*. 26439–26455.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7739–7751.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 564–572.

Hila Manor and Tomer Michaeli. 2024. Zero-shot unsupervised and text-based audio editing using DDPM inversion. *arXiv preprint arXiv:2402.10009* (2024).

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 8293–8316.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256* (2025).

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973* (2024).

Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1015–1018.

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 *International conference on machine learning*. PmLR, 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

Yan Rong, Jinting Wang, Guangzhi Lei, Shan Yang, and Li Liu. 2025. Audiogenie: A training-free multi-agent framework for diverse multimodality-to-multiaudio generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 8872–8881.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168* (2024).

Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. 2017. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 344–348.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. 2025. SAM Audio: Segment Anything in Audio. *arXiv preprint arXiv:2512.18099* (2025).

Kun Su, Judith Yue Li, Qingqing Huang, Dima Kuzmin, Joonseok Lee, Chris Donahue, Fei Sha, Aren Jansen, Yu Wang, Mauro Verzetti, et al. 2023. V2Meow: Meowing to the Visual Beat via Music Generation. *arXiv preprint arXiv:2305.06594* (2023).

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=14rn7HpKVk

Ye Tao, Xuenan Xu, Wen Wu, Shuai Wang, Mengyue Wu, and Chao Zhang. 2025. MMEDIT: A Unified Framework for Multi-Type Audio Editing via Audio Language Model. *arXiv preprint arXiv:2512.20339* (2025).

Jinchuan Tian, Sang-gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, et al. 2025b. Ualm: Unified audio language model for understanding, generation and reasoning. *arXiv preprint arXiv:2510.12000* (2025).

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025a. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522* (2025).

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025c. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 18782–18793.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. *arXiv preprint arXiv:2506.04779* (2025).

Le Wang, Jun Wang, Chunyu Qiang, Feng Deng, Chen Zhang, Di Zhang, and Kun Gai. 2025a. Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation. *arXiv preprint arXiv:2508.00733* (2025).

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. 2023b. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 71340–71357.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750* (2024).

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. 2025. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377* (2025).

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Zhifeng Xie, Qile He, Youjia Zhu, Qiwei He, and Mengtian Li. 2025a. FilmComposer: LLM-Driven Music Production for Silent Film Clips. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13519–13528.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. 2025b. Audiotime: A temporally-aligned audio-text benchmark dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7151–7161.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215* (2025).

Liumeng Xue, Ziya Zhou, Jiahao Pan, Zixuan Li, Shuai Fan, Yinghao Ma, Sitong Cheng, Dongchao Yang, Haohan Guo, Yujia Xiao, et al. 2025. Audio-flan: A preliminary release. *arXiv preprint arXiv:2502.16584* (2025).

M Yinghao, Ø Anders, R Anton, S Bleiz MacSen Del, S Charalampos, and D Chris. 2024. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340* (2024).

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al. 2025. Yue: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638* (2025).

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. 2024. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153* (2024).

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 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)*. 9637–9662.

1 Appendix

Table A1. **Ablation study on different encoders.** We evaluate the impact of different encoders for text-to-audio (T2A) and video-to-audio (V2A) tasks.

Task	Encoder	KL↓	IS↑	FD↓	FAD↓
T2A	T5	1.51	11.27	15.78	2.75
	CLAP	1.81	9.64	16.58	**1.84**
	Qwen	**1.25**	**11.65**	**12.21**	2.14
V2A	CLIP	2.20	7.42	12.21	2.80
	VideoMAE	2.29	6.58	14.72	3.06
	Qwen	**2.00**	**8.08**	**8.50**	**2.23**

Table A2. Zero-shot cross-lingual text-to-audio generation results.

Method	Language	KL↓	IS↑	FD↓	FAD↓
Tango2 [Majumder et al. 2024]	EN	**1.11**	10.37	12.22	3.20
AudioX [Tian et al. 2025a]	EN	1.34	**12.09**	**11.83**	**1.86**
MMAudio [Cheng et al. 2025]	EN	1.35	12.03	12.63	4.71
Stable-Audio-Open [Evans et al. 2024]	EN	2.01	10.37	29.01	3.15
Audio-Omni	EN	1.15	11.64	11.97	**1.86**
	CN	1.65	11.10	15.05	2.26
	ES	2.36	9.16	25.26	4.32
	DE	2.39	9.13	23.51	2.92
	FR	2.47	8.80	28.63	4.21
	JP	2.27	8.67	19.82	3.13

This section includes dataset details, zero-shot cross-lingual text-to-audio generation, and ablation studies.

1.1 Dataset Details

Our model is trained on a mixture of diverse datasets, as detailed in Table A4. The data composition for each task is as follows:

Text-to-Audio (T2A): Approximately 1.4k hours, sourced from AudioCaps [Kim et al. 2019], WavCaps [Mei et al. 2024], AudioSetCaps [Bai et al. 2025], and AudioTime [Xie et al. 2025b].

Video-to-Audio (V2A): Approximately 700 hours, sourced from VGGSound [Chen et al. 2020] and the AudioSet Strong [Hershey et al. 2021] benchmark.

Text-to-Music (T2M): Approximately 17k hours, combining data from V2M [Tian et al. 2025c] and MUCaps [Liu et al. 2024a].

Video-to-Music (V2M): Approximately 16k hours, sourced entirely from the V2M [Tian et al. 2025c] dataset.

Speech: Approximately 6k hours, using the English subset of Audio-FLAN [Xue et al. 2025].

Audio Editing: Approximately 3k hours from our internally constructed AudioEdit dataset, with the methodology detailed in Section 3.

Style-Transfer Keyword Generation. For style-transfer data construction, we prompt Gemini 2.5 Pro with each audio's original keyword to generate semantically related but stylistically different target keywords. The prompt template is: *"Given audio with keyword '[original_keyword]', generate a related but different keyword for style transfer."* Generated candidates are filtered by CLAP before guiding ZETA for style transformation.

1.2 More Comparison Results

Detailed Results on Audio Editing Tasks. Table A3 presents detailed performance on the four primary audio editing tasks using FAD and LSD on our AudioEdit benchmark. Audio-Omni consistently achieves SOTA performance across all individual tasks.

Detailed Results on Generation Tasks. We provide a comprehensive breakdown of our model's performance across all generation tasks with multiple evaluation metrics in Table A5. The table presents results on KL divergence, Inception Score (IS), Fréchet Distance (FD), and Fréchet Audio Distance (FAD) for Text-to-Audio (T2A), Text-to-Music (T2M), Video-to-Audio (V2A), Video-to-Music (V2M), and Text-to-Speech (TTS) tasks.

Zero-shot Cross-lingual Text-to-Audio Generation. A remarkable inherited capability of our framework is its zero-shot cross-lingual generation, inherited directly from the frozen MLLM's multilingual understanding. To evaluate this, we translated the AudioCaps test set into Chinese (CN), Spanish (ES), German (DE), French (FR), and Japanese (JP) using Gemini. As shown in Table A2, Audio-Omni maintains strong performance across all languages despite being trained almost exclusively on English. Notably, the quality of audio generated from non-English prompts (e.g., Chinese) is comparable to strong, English-only specialist models. This result validates that our decoupled architecture effectively transfers the MLLM's linguistic capabilities to the synthesis task, bridging the language gap in generative audio.

1.3 More Ablation Studies

Effect of the Unified MLLM Encoder. To validate our choice of the understanding module, we compare our frozen Qwen-Omni-3B (Qwen) MLLM against specialized single-modality encoders. For T2A, evaluated on AudioCaps [Kim et al. 2019], we replace it with text encoders (T5 [Raffel et al. 2020], CLAP [Elizalde et al. 2023]). For V2A, evaluated on VGGSound [Chen et al. 2020], we compare against vision encoders (CLIP [Radford et al. 2021], VideoMAE [Tong et al. 2022]). As shown in Table A1, the results consistently demonstrate the superiority of using a unified MLLM, which achieves significantly better performance across all metrics in both tasks. We attribute this to the MLLM's richer semantic understanding and, more importantly, its inherent ability to process multimodal contexts jointly, capturing cross-modal relationships that specialized encoders inherently miss.

1.4 Human Evaluation

To complement our objective metrics, we perform a comprehensive human evaluation study with 20 audio professionals. Following the evaluation protocols established in prior audio generation works [Kreuk et al. 2022; Majumder et al. 2024; Tian et al. 2025a], we assess the perceptual quality of our generated outputs against competitive baselines. For each task (T2A, T2M, V2A, V2M, and Audio Editing), we randomly select 20 test samples and present them to raters in a randomized, anonymized manner. Raters are asked to score each sample on two key dimensions: Overall Quality (OVL), which measures the general audio fidelity and naturalness, and Relevance (REL), which evaluates how well the output aligns with the given condition (text prompt, video content, or editing

Table A3. Detailed results on audio editing tasks. We report FAD/LSD for each task and their average.

Method	Add		Remove		Extract		Style Transfer		Avg.	
	FAD↓	LSD↓	FAD↓	LSD↓	FAD↓	LSD↓	FAD↓	LSD↓	FAD↓	LSD↓
ZETA [Manor and Michaeli 2024]	5.36	4.15	3.43	4.19	2.86	3.63	3.59	3.27	3.81	3.81
SDEdit [Meng et al. 2021]	4.22	4.18	2.16	4.46	4.51	4.31	3.14	4.63	3.51	4.40
MMEDIT [Tao et al. 2025]	5.30	3.55	2.08	4.41	4.54	3.88	3.88	4.35	3.95	4.05
Audio-Omni	2.78	2.32	3.68	1.91	3.52	2.04	3.11	2.82	3.27	2.27

Table A4. Training data summary across tasks.

Task	**Hours**	**Datasets**
T2A	1.4k	AudioCaps [Kim et al. 2019] WavCaps [Mei et al. 2024] AudioSetCaps [Bai et al. 2025] AudioTime [Xie et al. 2025b] IF-Caps [Tian et al. 2025a]
V2A	0.7k	VGGSound [Chen et al. 2020] AudioSet Strong [Hershey et al. 2021]
T2M	17k	IF-Caps [Tian et al. 2025a] MUCaps [Liu et al. 2024a]
V2M	16k	V2M [Tian et al. 2025c] MMTrail [?]
Speech	6k	Audio-FLAN (English subset) [Xue et al. 2025]
Editing	3k	AudioEdit (Sec. 3)

instruction). Both metrics are rated on a scale from 1 to 100, with higher scores indicating better performance.

As shown in Table A6, Audio-Omni consistently achieves competitive or superior ratings across all evaluated tasks, demonstrating that our model's outputs align well with human perception of quality and instruction adherence.

Table A6. Human evaluation results on generation and editing tasks.

Task	Method	OVL↑	REL↑
T2A	Tango2 [Majumder et al. 2024]	72.3	74.1
	AudioX [Tian et al. 2025a]	**81.5**	83.2
	Audio-Omni	78.6	**83.5**
T2M	MusicGen [Copet et al. 2024]	70.8	72.5
	AudioX [Tian et al. 2025a]	79.2	81.3
	Audio-Omni	**82.7**	**81.6**
V2A	MMAudio [Cheng et al. 2025]	**80.2**	**81.8**
	AudioX [Tian et al. 2025a]	79.5	81.2
	Audio-Omni	75.3	77.1
V2M	VidMuse [Tian et al. 2025c]	73.5	75.2
	AudioX [Tian et al. 2025a]	78.9	80.7
	Audio-Omni	**80.3**	**81.0**
Editing	SDEdit [Meng et al. 2021]	68.4	70.2
	ZETA [Manor and Michaeli 2024]	74.6	76.3
	Audio-Omni	**79.8**	**81.5**

Table A5. Detailed quantitative results on multimodal generation benchmarks with multiple metrics.

Method	KL ↓	IS ↑	FD ↓	FAD ↓
Text-to-Audio (T2A)				
Tango2 [Majumder et al. 2024]	**1.11**	10.37	12.22	3.20
MMAudio [Cheng et al. 2025]	1.35	12.03	12.63	4.71
Stable-Audio-Open [Evans et al. 2024]	2.01	10.37	29.01	3.15
Unified-IO2 [Lu et al. 2024]	2.72	5.44	37.95	7.81
AudioX [Tian et al. 2025a]	1.34	**12.09**	**11.83**	**1.86**
Audio-Omni	1.15	11.64	11.97	**1.86**
Text-to-Music (T2M)				
Stable-Audio-Open [Evans et al. 2024]	1.51	2.94	36.33	3.23
MusicGen [Copet et al. 2024]	1.43	2.24	25.40	4.55
AudioX [Tian et al. 2025a]	1.02	**3.54**	10.63	1.53
MuMuLLaMA [Liu et al. 2024a]	1.00	1.25	52.25	5.10
Unified-IO2 [Lu et al. 2024]	**0.81**	2.47	18.94	3.17
Audio-Omni	0.84	3.49	**8.68**	**1.41**
Video-to-Audio (V2A)				
MMAudio [Cheng et al. 2025]	1.97	**14.95**	**6.18**	2.04
VATT [Liu et al. 2024b]	**1.40**	10.02	11.71	2.55
AudioX [Tian et al. 2025a]	2.57	12.16	8.83	**1.13**
Audio-Omni	1.98	10.35	8.33	1.71
Video-to-Music (V2M)				
VidMuse [Tian et al. 2025c]	0.73	1.32	22.95	2.46
AudioX [Tian et al. 2025a]	0.69	1.34	23.96	2.12
MuMuLLaMA [Liu et al. 2024a]	1.00	1.25	52.25	5.10
Audio-Omni	**0.64**	**1.36**	**17.46**	**1.51**

2 Ethics Statement

While Audio-Omni demonstrates powerful capabilities in audio understanding, generation, and editing, we acknowledge potential ethical risks associated with generative audio technologies. Tasks such as voice conversion and speech synthesis could be misused for creating deepfakes, impersonation, or spreading misinformation. To mitigate these risks, we will require users to accept responsible-use terms before accessing the model, explicitly prohibiting malicious applications. We encourage the community to develop robust audio watermarking and detection methods, and believe that with proper safeguards, Audio-Omni can serve as a valuable tool for creative expression and scientific research.