UniMoE-Audio: Unified Speech and Music Generation with Dynamic-Capacity Mixture-of-Experts

Anonymous ACL submission

Abstract

Recent advances in unified multimodal models indicate a clear trend towards comprehensive content generation. However, the auditory domain remains a significant challenge, with music and speech often developed in isolation, hindering progress towards universal audio synthesis. This separation stems from inherent task conflicts between semantic speech and structural music modeling, and severe data imbalances, which impede the development of a truly unified model. To address these challenges, we propose **UniMoE-Audio**, a unified speech and music generation model built upon a novel **D**ynamic-**C**apacity **M**ix-**o**f-**E**xperts (DCMoE) framework. Architecturally, UniMoE-Audio extends the conventional MoE paradigm by introducing a Top-P routing strategy for adaptive capacity allocation. To tackle data imbalance, we introduce a three-stage training curriculum: 1) Independent Specialist Training leverages original datasets to instill domain-specific knowledge into each specialists without interference; 2) MoE Integration and Warmup incorporates these specialists into the UniMoE-Audio architecture, warming up the gate module and shared expert using a subset of balanced dataset; and 3) Synergistic Joint Training trains the entire model end-to-end on the fully balanced dataset, fostering enhanced cross-domain synergy. Extensive experiments show that UniMoE-Audio not only achieves state-of-the-art performance on major speech and music generation benchmarks, but also demonstrates superior synergistic learning, mitigating the performance degradation typically seen in naive joint training. Our findings highlight the substantial potential of specialized MoE architecture and curated training strategies in advancing universal audio generation.

1 Introduction

A hallmark of human intelligence is the seamless ability to perceive, reason, and create across multiple modalities, effortlessly blending language, vision, and audio. Emulating this holistic capability represents a grand challenge and a core objective in the pursuit of more general artificial intelligence. The recent ascendancy of Large Language Models (LLMs) has served as a powerful catalyst, paving the way for unified models that can understand and generate content across these diverse data streams (Alayrac et al., 2022). Significant progress has been made in systems that jointly process text, images, video, and even speech within a single architecture (Zhan et al., 2024; Wu et al., 2025b; Xu et al., 2025; AI et al., 2025; KimiTeam et al., 2025; Huang et al., 2025). Nevertheless, a critical imbalance persists in the treatment of the auditory domain. While speech has been a primary focus of integration (KimiTeam et al., 2025; Huang et al., 2025), music—a domain of comparable complexity and cultural richness—remains largely siloed and excluded from these unified frameworks. This limitation not only hinders the pursuit of universal audio synthesis but also stands as a significant impediment to developing AI with truly comprehensive multimodal intelligence.

The primary obstacle to unifying speech and music generation stems from two fundamental challenges. The first is **task conflict**, arising from the divergent objectives of speech and music generation. Fundamentally, speech is primarily semantic, whereas music is primarily structural (Borsos et al., 2023). Consequently, the former prioritizes semantic intelligibility and speaker identity, while the latter demands the capture of long-term dependencies and complex hierarchies like harmony and rhythm. This divergence creates conflicting optimization pressures within a shared model, where progress on one task can impede the other. Recently, the MoE paradigm has emerged as a promising architecture for mitigating conflicts of multimodal understanding (Lin et al., 2024; Li et al., 2025; AI et al., 2025). Despite these advances, its application and further optimization for unified audio generation remain

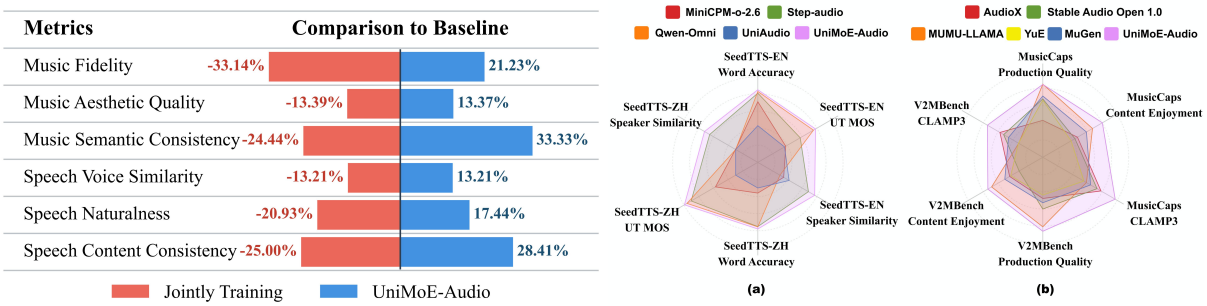

Figure 1: Performance of UniMoE-Audio. **Left:** Comparison against specialized baselines reveals the failure of naive joint training, which causes a clear performance degradation on speech generation and more significant decline on music generation. In contrast, our UniMoE-Audio yields synergistic gains across both tasks. **Right:** Radar charts show UniMoE-Audio achieving the best comprehensive performance against leading models on a wide array of speech (a) and music (b) metrics.

largely unexplored.

Beyond task conflict, another major hurdle is **data imbalance**. High-quality, large-scale speech corpora are far more abundant than their musical counterparts. The detrimental effects of this disparity are evident in prior work (Yang et al., 2024). Consequently, a naive joint training approach often allows the data-rich speech task to dominate the learning process, resulting in a substantial degradation in musical quality. Our preliminary experiments empirically confirm this degradation (Figure 1), showing that a jointly trained model performs significantly worse than specialized models, with the performance drop being particularly severe for the data-scarce music task. Therefore, the core research question we address is: ***how to overcome both task conflict and data imbalance, enabling a shared model to master speech and music generation synergistically?***

We address these challenges at both the architectural and training levels. Architecturally, we propose the Dynamic-Capacity Mixture-of-Experts (DCMoE) framework to mitigate task conflict. Unlike standard MoE models that rely on fixed-capacity routing (e.g., top-k), we introduce a Top-P routing strategy. This mechanism dynamically adjusts the number of experts allocated to each token according to its complexity, thereby enabling more flexible expert combinations. Complementing our architectural design, we introduce a three-stage training curriculum specifically designed to address data imbalance: (1) **Independent Specialist Training** utilizes raw, uncurated datasets to impart domain-specific knowledge to each proto-expert in isolation. (2) **MoE Integration and Warmup** incorporates these specialists into the UniMoE-Audio framework. We construct a balanced dataset via a rigorous filtering pipeline. Then we warm up the newly initialized router on a subset of this data to ensure stability. (3) **Synergistic Joint Training** optimizes the entire model on the full balanced dataset, facilitating effective cross-domain knowledge transfer. Our main contributions can be summarized as follows:

- We propose **UniMoE-Audio**, a unified speech and music generation model built on a novel Dynamic-Capacity Mix-of-Experts architecture. By combining a Top-P routing strategy for adaptive capacity allocation, it can effectively mitigate the intrinsic task conflict between speech and music generation.

- To fully exploit this architecture and tackle data imbalance, we introduce a three-stage training curriculum. It first trains independent dense specialists on each domain, then integrates their FFN modules as proto-expert into the UniMoE-Audio architecture, and finally performs synergistic joint training on a curated balanced dataset. This enables robust learning from highly imbalanced sources without relying on ad-hoc resampling strategies.

- Extensive experiments demonstrate that UniMoE-Audio achieves SOTA performance on major speech and music generation benchmarks. In-depth analyses of the dynamic activation patterns of experts further illuminate how the unified model allocates capacity and coordinates knowledge across diverse audio generation tasks.

2 Related Work

Domain-Specific Audio Models. Recent advancements have seen a convergence in speech and music generation towards autoregressive modeling over discrete audio tokens (Zhang et al., 2023; Huang et al., 2023, 2025; Liu et al., 2024). In text-to-speech, VALL-E (Chen et al., 2024) pioneered the use of neural codec (Défossez et al., 2022) for zero-shot synthesis, inspiring subsequent robust systems like CosyVoice (Du et al., 2024a) and SpearTTS (Kharitonov et al., 2023) which scale up training for high-fidelity generation. Parallel evolution has occurred in music generation; while diffusion models remain active (Agostinelli et al., 2023; Evans et al., 2025; Tian et al., 2025a), autoregressive frameworks like MusicGen (Copet et al., 2023) and YuE (Yuan et al., 2025) have demonstrated superior controllability and long-form generation capabilities. While the aforementioned studies demonstrate substantial advancements in speech and music generation, they primarily focus on advancing the state-of-the-art within their respective domains. Our work, in contrast, shifts the focus from domain-specific excellence to the challenge of cross-domain unification, aiming to broaden the scope of what autoregressive audio models can achieve.

Unified Audio Generation. Efforts to unify diverse audio tasks into a single framework remain nascent. A notable early attempt, UniAudio (Yang et al., 2024), proposed a general-purpose model via naive joint training but reportedly suffered from severe data imbalance, leading to suboptimal performance on data-scarce tasks like music. More recently, AudioX (Tian et al., 2025a) explored multimodal music generation but excluded speech synthesis, failing to bridge the core gap between sematic and structural audio. In contrast, our work focus on addressing the task conflict and data imbalance of unified audio generation. Instead of naive joint training, we introduce a DCMoE architecture and a three-stage curriculum, aiming to provide a more principled and effective pathway toward truly unified and high-fidelity audio generation

3 UniMoE-Audio

We present UniMoE-Audio, a unified generative framework designed to synthesize high-fidelity speech and music from multimodal inputs, including text, audio, and video. As illustrated in Figure 2, the cornerstone of our architecture is the **Dynamic-Capacity Mix-of-Experts** framework. This framework not only leverage MoE paradigms for handle task conflict between speech and music, and also introduce a Top-P routing strategy that adaptively allocates the number of experts based on token processing difficulty.

3.1 Input Representation and Tokenization

Audio Tokenization. Following established practices in audio generation, we employ a neural audio codec to transform continuous waveforms into a sequence of discrete acoustic tokens. Specifically, we utilize the DAC codec (Kumar et al., 2023), which represents each audio frame using a multi-channel codebook. Unlike some works (Défossez et al., 2024; Yang et al., 2024) that employ the Depth Transformer to predict tokens for each channel sequentially, we adopt a more parameter-efficient approach. Our model predicts all channels with a multi-head output layer. This design avoids the introduction of additional sequential modules, thereby reducing the overall parameter count and computational latency.

Visual Embedding. To process visual inputs (e.g., from video), we follow the Qwen-VL (Wang et al., 2024), using a Visual Transformer (ViT) to encode the input image into patches. These visual features are then mapped into the language model's embedding space via a projector module, yielding a sequence of soft visual tokens that can be seamlessly integrated with text and audio representations.

3.2 Top-P Routing

A primary limitation of conventional MoE models is their static Top-K routing strategy, which allocates a fixed number of experts to each token. This approach is computationally sub-optimal, as it may over-allocate computational resources to simple tokens while under-powering complex ones that require more extensive processing. To address this, we introduce a Top-P routing mechanism that dynamically allocates the number of activated experts for each token based on the routing probability of the router module.

Given an input tensor $X \in \mathbb{R}^{N \times d}$ for an FFN layer, where N is the sequence length and d is the hidden dimension, a linear module first computes the gating probabilities for all E experts:

$$P = \text{Softmax}(XW_g), \tag{1}$$

where $W_g \in \mathbb{R}^{d \times E}$ is the trainable gating matrix

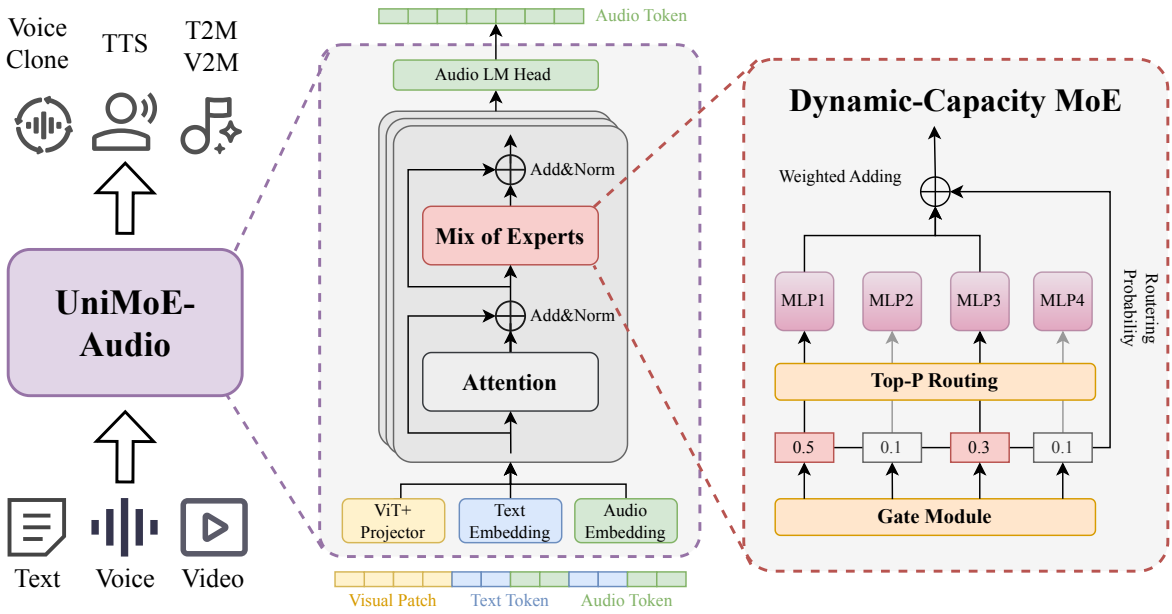

Figure 2: An overview of the UniMoE-Audio framework. **Left:** UniMoE-Audio is a unified model capable of performing speech and music generation by leveraging multimodal conditional inputs, including Voice Cloning, Text-to-Speech (TTS), Text-to-Music (T2M), and Video-to-Music (V2M). **Center:** The core architecture of our model is a Transformer with Dynamic-Capacity MoE layers. **Right:** We propose a novel Top-P routing , which dynamically selects the number of experts allocated to each token based on their processing difficulty.

and $P \in \mathbb{R}^{N \times E}$ represents the probability distribution over experts for each token.

We interpret this distribution P as the router's confidence. The objective is to select the smallest set of experts whose cumulative probability exceeds a predefined threshold p, thereby balancing computational cost and predictive accuracy. This can be formulated as finding an index set I for each token such that:

$$I = \arg\min_{I'} |I'| \quad \text{s.t.} \quad \sum_{i \in I'} P_i \geq p. \tag{2}$$

To solve this, we employ the Top-P sampling algorithm, sorting expert probabilities in descending order and selecting the smallest set whose cumulative sum exceeds the threshold P. The experts included in this sum are selected for computation. This approach naturally links the number of selected experts to the complexity of token, which is reflected in the router's probability distribution: low-entropy distributions correspond to simpler tokens, while high-entropy ones indicate more complex tokens requiring more experts.

The final output of the MoE layer is a weighted sum of the outputs from the selected experts, where the weights are the normalized gating probabilities:

$$O = \sum_{i \in I} \frac{P_i}{\sum_{j \in I} P_j} E_i(X), \tag{3}$$

where I is the set of selected expert, and $E_i(X)$ is the output of the i-th expert.

4 Training

The successful unification of speech and music generation hinges not only on the model architecture but also on a training strategy that can effectively navigate the challenges of data imbalance and task conflict. To this end, we devise a comprehensive approach encompassing both rigorous data governance and a three-stage training curriculum.

4.1 Training Data

Raw Multitask Corpus. We first construct a large-scale raw corpus covering four generation tasks: Mandarin Text-to-Speech, English Text-to-Speech, Text-to-Music, and Video-to-Music. Overall statistics are given in Table 2. For speech, we start from in-house studio-quality human recordings and then expand the corpus by speaker-cloning synthesis. Concretely, about 20% of the utterances are real recordings, and the remaining 80% are generated with CosyVoice2 (Du et al., 2024b). Most speech clips have durations between 3 and 10 seconds. For music, we collect audio from open-source music datasets and segment long tracks into fixed 20-second clips. Each clip is paired with a textual description produced by Gemini-2.5-Flash, providing

Name	Task	Architecture	Activated Param	Total Param
Expert-ZhTTS	Mandarin TTS	Dense	3.1B	3.1B
Expert-EnTTS	English TTS	Dense	3.1B	3.1B
Expert-T2M	Text to Music	Dense	3.1B	3.1B
Expert-V2M	Video to Music	Dense	3.1B	3.1B
Unify-Dense	Unify Audio Generation	Dense	7.1B	7.1B
UniMoE-Audio	Unify Audio Generation	DCMoE	Avg: 4.8B (Min: 2.8B, Max: 5.9B)	7.1B

Table 1: Model configurations and parameters of all model variants used in our main experiments.

Task	Datasets	Number	Duration
Speech Synthesis	Mandarin TTS	180K	20K
	English TTS	100K	10K
Text-to-Music	FMA (Defferrard et al., 2017)	106K	8.2K
	MusicNet (Thickstun et al., 2017)	320	37
	MU2Gen (Liu et al., 2024)	22K	1.2K
Video-to-Music	V2M (Tian et al., 2025b)	20K	600

Table 2: Overview of datasets used in different tasks with instance number and duration (hours).

captions for both T2M and V2M scenarios. This raw corpus serves as the foundation from which we later derive a smaller, curated subset.

Quality Filtering and Curated Subset. On top of the raw corpus, we build a high-quality, integrated dataset through an automatic filtering pipeline. For Mandarin TTS and English TTS, we compute UTMOS (Saeki et al., 2022) scores and discard samples whose perceptual quality falls below a predefined threshold. For T2M and V2M, we estimate Aesthetic Quality (Tjandra et al., 2025) for the audio and semantic similarity scores between audio and text using CLAP (Wu et al., 2023), and remove clips with low aesthetic quality or poor audio–text alignment. From the remaining data, we select a moderate-scale subset that is approximately balanced across the four tasks, resulting in about 60K high-quality samples in total. By fine-tuning on this curated subset, we can prevent the model from becoming biased toward data-rich speech tasks while fostering genuine cross-domain synergy.

4.2 Three-stage Training Curriculum

A naive joint training approach on the imbalanced dataset would inevitably lead to the data-rich speech task dominating the learning process. Conversely, simple up-sampling or down-sampling from the outset either sacrifices data diversity or discards valuable resources. To systematically circumvent this dilemma, we propose a three-stage training curriculum, designed to decouple large scale task-specific learning from synergistic optimization

Independent Specialist Training. The primary objective of this stage is to mitigate task conflict at its source and maximize data utilization. We leverage the full, imbalanced raw datasets to train separate, dense models for each task, as listed in Table 1. This complete isolation allows each model to master its domain-specific knowledge without interference from other tasks. This process effectively injects specialized knowledge into the parameters of each future expert, pre-assigning their intended function before they are integrated.

MoE Integration and Warmup. We then integrate these specialists into the UniMoE-Audio framework. Specifically, the FFN blocks of the specialists serve as the proto-expert. Shared components (e.g., attention modules) are initialized by averaging parameters across all dense specialist, while the vision encoder is inherited directly from the V2M specialist. To prevent catastrophic forgetting, we initially freeze the pre-trained experts. We then exclusively train the router on the balanced dataset. This warmup step enables the router to learn effective dispatching policies based on the experts' existing specializations.

Synergistic Joint Training. Finally, we unfreeze the entire model for end-to-end fine-tuning on the balanced dataset. To ensure routing efficiency, we apply an auxiliary load-balancing loss with a linearly decaying weight. This annealing strategy initially encourages exploration through balanced expert usage, then gradually shifts focus to exploiting learned routing patterns for maximizing generation performance.

5 Experiments

5.1 UniMoE-Audio Setting

Table 1 summarizes the specifications of all model variants evaluated in our experiments, including:

Specialist Baselines: We employ four task-specific dense models (Expert-ZhTTS, Expert-EnTTS, Expert-T2M, and Expert-V2M), each with

Method	SeedTTS-EN			SeedTTS-ZH			LibriSpeech		AISHELL-3	
	WER↓	UTMOS↑	SIM↑	CER↓	UTMOS↑	SIM↑	WER↓	UTMOS↑	CER↓	UTMOS↑
UniAudio (Tian et al., 2025a)	7.2	3.46	0.40	-	-	-	20.2	3.26	-	-
Mini-CPM-O-2.6 (Yao et al., 2024)	3.4	3.49	0.36	13.0	2.94	0.47	11.1	3.76	13.1	3.30
Qwen2.5-Omni (Xu et al., 2025)	2.1	4.16	-	1.6	3.28	-	7.6	4.19	2.5	3.38
Step-audio (Huang et al., 2025)	2.2	3.84	0.52	1.0	3.23	0.62	5.0	4.37	2.7	3.69
Step-audio 2 mini (Wu et al., 2025a)	1.6	4.22	0.47	1.6	3.40	0.63	3.5	4.35	3.2	4.00
Higgs audio V2 (Boson AI, 2025)	**1.0**	4.00	**0.67**	**0.8**	3.27	**0.73**	3.6	4.26	5.9	3.89
MiMo (Xiaomi, 2025)	4.6	3.06	-	1.0	2.35	-	7.3	2.83	6.9	2.32
Expert-EnTTS	2.5	3.57	0.48	-	-	-	6.3	3.46	-	-
Expert-ZhTTS	-	-	-	3.5	3.25	0.58	-	-	4.8	3.47
Unify-Baseline	3.1	2.62	0.43	4.8	2.96	0.49	7.9	2.55	5.3	2.73
UniMoE-Audio	1.3	**4.36**	0.64	**0.8**	**3.78**	0.67	**3.4**	**4.46**	**1.0**	**4.13**

Table 3: Performance on English and Mandarin speech synthesis benchmarks. The best performance for each metric is highlighted in **bold**, and the second best is underlined. ↑ indicated higher is better. WER and CER measure content intelligibility, UTMOS measure perceptual quality, and SIM measure speaker similarity.

3.1B parameters and initialized from Qwen2.5-VL. These models are trained on their respective tasks. They serve a dual purpose: acting as the proto-experts for DCMoE and providing a performance benchmark for dedicated, single-task systems.

Unify-Dense: To isolate the benefits of our DCMoE architecture from mere parameter scaling, we propose Unify-Dense. Unify-Dense is developed base on Qwen-2.5-VL, designed to match the total parameter count of our primary UniMoE-Audio model but is trained via standard joint training on the combined dataset.

UniMoE-Audio: Our primary model, built upon the Qwen2.5-VL architecture with a total of 7.1B parameters. It features our DCMoE with Top-P routing ($p = 0.7$). As detailed in Table 1, while the total capacity is large, the number of activated parameters varies dynamically based on token complexity, averaging approximately 4.8B (ranging from 2.8B to 5.9B).

5.2 Implementation Details

We employ the AdamW (Loshchilov and Hutter, 2019) optimizer in conjunction with a cosine learning rate scheduler across all training stages. Subsequently, in the independent specialist training stage, we utilize 48 Ascend 910B2 GPUs, with a global batch size of 48 and a base learning rate of 1e-4. In the warmup stage, we utilize 196 Ascend 910B GPUs for training, with a global batch size of 784 and a base learning rate of 3e-5. Finally, in the synergistic joint training stage, we utilize 196 Ascend 910B GPUs, with a global batch size of 3136 and a base learning rate of 1e-5. We adopt expert parallelism with four-way partitioning, meaning only one routed expert are loaded on each GPU. For inference, we employ greedy sampling for both text and speech token generation. We evaluate our model with three random seed and report their average performance.

5.3 Evaluation Setting

Speech Synthesis. For speech synthesis, we follow the setting of Seed-TTS (Anastassiou et al., 2024) and evaluate models on both English and Mandarin benchmarks, focusing on three primary aspects: content consistency, speaker similarity, and perceptual quality. Our evaluation benchmark includes the Seed-TTS test set, the LibriSpeech test-clean set (Panayotov et al., 2015), and AISHELL-3 (Shi et al., 2021). For content intelligibility and perceptual quality, we utilize a predefined voice prompt to isolate the model's generative quality from prompt variations.

- **Content Consistency** is measured by Word Error Rate (WER) for English and Character Error Rate (CER) for Mandarin, computed with the Whisper-large-v3 (Radford et al., 2023) and Paraformer-zh (Gao et al., 2022) as ASR engines, respectively.
- **Perceptual Quality** is assessed using UTMOS (Saeki et al., 2022) as an objective proxy for subjective human ratings.
- **Speaker Similarity** is quantified by the cosine similarity of speaker embeddings extracted from a fine-tuned WavLM model, following the methodology of Seed-TTS.

Music Generation. For music generation, we evaluate both T2M and V2M tasks, assessing semantic alignment, audio quality, and aesthetic quality. The T2M task is evaluated on MusicCaps (Agostinelli et al., 2023) and V2M-bench(Tian et al., 2025b),

Dataset	Method	Task	PC↑	PQ↑	CE↑	CLAP↑	KL↓	CLaMP3↑	IS↑	FAD↓
MusicCap	YuE (Yuan et al., 2025)	T2M	3.45	7.25	5.84	0.18	2.12	0.09	2.09	9.02
	Stable Audio Open 1.0 (Evans et al., 2025)	T2M	3.70	7.29	6.02	0.30	1.44	0.11	2.74	3.72
	AudioX (Tian et al., 2025a)	T2M	5.00	6.67	6.14	0.25	**1.20**	0.12	**3.02**	**1.64**
	MusicGen (Copet et al., 2023)	T2M	4.78	7.37	6.57	0.26	1.21	0.10	1.68	7.02
	MUMU-LLAMA (Liu et al., 2024)	T2M	5.15	7.71	6.87	0.20	1.27	0.10	1.44	8.57
	Expert-T2M	T2M	5.52	6.67	6.23	0.20	1.41	0.10	1.38	7.52
	Unify-Baseline	T2M	5.40	6.52	5.79	0.11	1.43	0.07	1.16	8.72
	UniMoE-Audio	T2M	**6.17**	**7.89**	**7.34**	**0.34**	1.25	**0.15**	2.87	3.43
V2M-bench	YuE (Yuan et al., 2025)	T2M	3.78	7.25	6.01	0.15	1.27	0.13	1.79	4.29
	Stable Audio Open 1.0 (Evans et al., 2025)	T2M	3.41	7.46	5.69	0.34	1.91	0.16	3.13	2.94
	AudioX (Tian et al., 2025a)	T2M	4.60	7.30	6.06	0.30	2.12	0.11	**3.64**	4.26
	MusicGen (Copet et al., 2023)	T2M	4.64	7.37	6.24	0.28	1.27	0.15	1.70	3.39
	MUMU-LLAMA (Liu et al., 2024)	T2M	5.19	**7.73**	6.75	0.17	**0.92**	0.13	1.42	**2.54**
	Expert-T2M	T2M	5.40	6.77	6.51	0.25	1.70	0.15	1.85	3.81
	Unify-Baseline	T2M	5.25	6.59	5.30	0.23	1.99	0.15	1.23	5.81
	UniMoE-Audio	T2M	**5.91**	7.58	**6.85**	**0.38**	1.04	**0.19**	2.31	3.14
V2M-bench	AudioX (Tian et al., 2025a)	V2M	4.44	7.44	6.06	-	1.8	-	3.14	2.94
	Expert-V2M	V2M	5.14	7.34	6.71	-	1.89	-	2.34	4.45
	Unify-Baseline	V2M	4.98	5.47	4.13	-	1.95	-	1.62	6.48
	UniMoE-Audio	V2M	**5.88**	**7.52**	**6.85**	-	**1.69**	-	**3.34**	**2.91**

Table 4: Performance on text-to-music and video-to-music generation benchmarks. The best performance is highlighted in **bold**, and the second best is underlined. ↑ indicated higher is better. PC, PQ, and CE measure the aesthetic quality. CLAP and CLaMP3 measure semantic alignment between the description and generated music. KL and FAD assess audio quality against reference tracks , while IS assess audio diversity.

and the V2M task is evaluated on V2M-bench. Notably, to align with the setting of MusicCap, all video and audio samples from V2M-Bench are segmented into 10-second clips.

- **Semantic Alignment** between text and audio is measured using CLAP score (Wu et al., 2023). To provide a more robust assessment, we also report the CLaMP3 score (Wu et al., 2025c), which leverages a more advanced multilingual framework.
- **Audio Quality and Diversity** are evaluated using a suite of metrics: Fréchet Audio Distance (FAD) with OpenL3 embeddings (Kilgour et al., 2019), Kullback-Leibler (KL) divergence based on PaSST (Koutini et al., 2022), and Inception Score (IS).
- **Aesthetic Quality** is evaluated using three specialized metrics from Tjandra et al. (2025): Production Complexity (PC), Production Quality (PQ), and Content Enjoyment (CE).

5.4 Overall Performance

We conducted a comprehensive evaluation of UniMoE-Audio against state-of-the-art specialized models and strong baselines. As detailed in Table 3 and Table 4, our results demonstrate that UniMoE-Audio achieve superior performance across both speech and music domains, overcoming the interference associated with multi-task learning,

Takeaway 1: UniMoE-Audio achieves SOTA speech synthesis with remarkable data efficiency. UniMoE-Audio demonstrates exceptional capabilities in speech synthesis, setting a new benchmark on SeedTTS-EN with a UTMOS of 4.36 and a WER of 1.3. Notably, this performance is achieved using only 280K hours of data, rivaling dedicated systems trained on 10M hours (e.g., Higgs Audio V2). This underscores the high data efficiency and strong representational power of our unified architecture.

Takeaway 2: The model excels in generating aesthetically superior and semantically aligned music. In the music domain (Table 4), UniMoE-Audio consistently prioritizes aesthetic quality. It obtains the highest scores across all aesthetic metrics (PC, PQ, CE) for both T2M and V2M tasks, indicating a superior ability to produce rich, enjoyable musical content. Furthermore, the model achieves precise semantic alignment, evidenced by leading CLAP and CLaMP3 scores. While reference-based metrics (FAD) are slightly lower, we attribute this to the model's tendency towards creative generation rather than mere imitation.

Takeaway 3: The DCMoE architecture is critical for resolving task conflict and data imbalance. A direct comparison with the Unify-Dense reveals

Method	TTS		Music	
	WER↓	UTMOS↑	PQ↑	CE↑
UniMoE-Audio	**2.2**	**3.97**	**7.54**	**7.39**
Unify-MoE	3.1	3.63	<u>7.47</u>	7.24
w/o Initialization	8.9	3.31	5.32	5.19
Confidence-Threshold:				
$p = 0.5$	3.7	3.57	7.21	6.56
$p = 0.9$	<u>2.5</u>	<u>3.85</u>	7.41	<u>7.26</u>
Expert Number:				
8 Experts	4.3	3.12	7.00	6.61
16 Experts	11.4	2.09	6.25	5.88

Table 5: Ablation studies of UniMoE-Audio across architecture, training algorithm, and hyper-parameter settings in TTS and music generation. The best performance is highlighted in **bold**, and the second best is <u>underlined</u>. ↑ indicated higher is better.

the necessity of our Dynamic-Capacity MoE design. Despite similar parameter counts, the dense baseline suffers from catastrophic forgetting, particularly on the data-scarce V2M task (PC: 4.98 vs. 5.88), where the dominant speech task overwhelms the music modality. In contrast, UniMoE-Audio maintains robust performance across all tasks, confirming that our strategy of pre-training proto-experts followed by dynamic routing effectively isolates task-specific knowledge and prevents cross-task interference.

5.5 Ablation Study

To delve into the contributions from our model architecture, training algorithm, and hyper-parameter settings, we conduct a series of ablation studies. For lightweight experiment, all models in this section are trained on a randomly sampled 20% subset of the full balanced dataset. We report the average WER and UTMOS for Seed-TTS-EN and Seed-TTS-ZH, along with PQ and CE for MusicCaps to evaluate performance across both speech and music domains.

Routing Algorithm Ablation. We first evaluate the efficacy of our Top-P routing by comparing it with a standard fixed-capacity strategy. The Unify-MoE variant employs the conventional Top-2 routing strategy. As shown in Table 5, Unify-MoE exhibits a clear performance drop compared to UniMoE-Audio, with WER increasing from 2.2 to 3.1 and PQ dropping from 7.54 to 7.33. This degradation suggests that a fixed computational budget is suboptimal for unified audio generation. It likely over-allocates resources to simple tokens while under-serving complex ones, whereas our Top-P routing adaptively aligns capacity with token difficulty.

Training Strategy Ablation. We investigate the necessity of our "Independent Specialist Training" phase. The *w/o Initialization* variant skips the pre-training of proto-experts and directly performs joint training from scratch. The results are catastrophic: WER surges to 8.9, and music metrics plummet (PQ 5.32). This confirms that without the domain-specific knowledge injected into the proto-experts, the router struggles to disentangle the modalities during the early stages of training, leading to severe task interference and optimization difficulties.

Hyper-parameters Search. We further explore the impact to the routing threshold p and expert granularity. Regarding the routing strategy, setting a lower value $p = 0.5$ results in a noticeable performance decline due to insufficient capacity allocation. Conversely, increasing the threshold to $p = 0.9$ forces the activation of nearly all experts, effectively reverting the model to a dense behavior. Since too many experts are simultaneously active, it may negates the benefits of sparse specialization. Finally, we examine the impact of expert quantity by dividing the proto-experts into 8 and 16 smaller experts while maintaining constant parameters. We observe a sharp deterioration in performance as the number of experts increases (e.g., 11.4 WER with 16 experts). We attribute this to routing instability; as the number of experts grows, the probability distribution output by the router tends to become flatter with diminished confidence gaps, rendering the routing selection sensitive to noise, thereby reduce robustness.

6 Conclusion

In this paper, we addressed the long-standing challenge of unifying speech and music generation, hindered by task conflict and data imbalance. We introduced UniMoE-Audio that leverages a dynamic-capacity Mixture-of-Experts architecture to mitigate task conflict, in conjunction with a three-stage training curriculum to overcome data imbalance. Experiments across diverse benchmarks show that UniMoE-Audio not only matches or surpasses strong domain-specific baselines, but also enables synergistic learning across audio domains—effectively avoiding the performance degradation observed in naive joint training. Our work provides a robust blueprint for building unified generative audio models, with future directions include the incorporation of a broader range of audio types and the optimization of MoE architecture.

Limitations

Despite our discoveries and improvements, we must acknowledge certain limitations in our work:

First, regarding long-form audio generation, the model occasionally exhibits challenges in maintaining rhythmic consistency over extended durations. While the underlying architecture excels at the short-term semantic coherence required for TTS task, adapting this mechanism to capture the long-range structural dependencies inherent in musical compositions remains unstable. This limitation may impact the overall listening experience in longer musical pieces.

Second, although the model performs well in general music generation, we observe minor instabilities in complex instruction following. In scenarios involving fine-grained user prompts, the model may not fully capture every specific element or constraint specified in the instruction. This can occasionally lead to deviations from the intended sub-genre or specific stylistic nuances, affecting the precision of generation in highly specific musical domains.

Third, while UniMoE-Audio achieves competitive results in zero-shot voice cloning, its generalization capability across diverse demographic attributes shows room for improvement. Specifically, the similarity and naturalness of cloned voices can vary when handling speakers with distinct accents or specific age groups. This variability may impact the reliability of voice cloning in specialized applications requiring high fidelity across a broad range of speaker characteristics.

These limitations highlight critical directions for future work, including designing mechanisms for better long-term structural modeling in music, refining instruction alignment for complex instruction, and enhancing the generalization of speaker representation to ensure consistent cloning performance.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. 2023. Musiclm: Generating music from text. *CoRR*, abs/2301.11325.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, Guangming Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, and 39 others. 2025. Ming-omni: A unified multimodal model for perception and generation. *CoRR*, abs/2506.09344.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *CoRR*, abs/2406.02430.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533.

Boson AI. 2025. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio. GitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *CoRR*, abs/2406.05370.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pages 316–323.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *CoRR*, abs/2210.13438.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2063–2067. ISCA.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, and 81 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *CoRR*, abs/2502.11946.

Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. 2023. Make-a-voice: Unified voice synthesis with discrete representation. *CoRR*, abs/2305.19269.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. Assoc. Comput. Linguistics*, 11:1703–1718.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2350–2354. ISCA.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *CoRR*, abs/2504.18425.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. 2022. Efficient training of audio transformers with patchout. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2753–2757. ISCA.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947.

Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, and Ying Shan. 2024. Mumu-llama: Multimodal music understanding and generation via large language models. *CoRR*, abs/2412.06660.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4521–4525. ISCA.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A multi-speaker mandarin TTS corpus. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2756–2760. ISCA.

John Thickstun, Zaïd Harchaoui, and Sham M. Kakade. 2017. Learning features of music from scratch. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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. *CoRR*, abs/2503.10522.

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025b. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 18782–18793. Computer Vision Foundation / IEEE.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *CoRR*, abs/2502.05139.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 81 others. 2025a. Step-audio 2 technical report. *CoRR*, abs/2507.16632.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025b. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. 2025c. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2605–2625. Association for Computational Linguistics.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen M. Meng. 2024. Uniaudio: Towards universal audio generation with large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, Xinrun Du, Zhen Ye, Tianyu Zheng, Yinghao Ma, Minghao Liu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, and 38 others. 2025. Yue: Scaling open foundation models for long-form music generation. *CoRR*, abs/2503.08638.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9637–9662. Association for Computational Linguistics.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *CoRR*, abs/2303.03926.

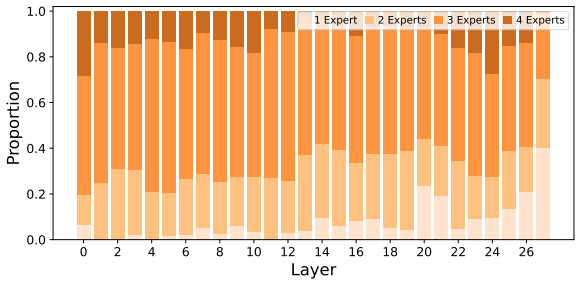

Figure 3: Layer-wise distribution of activated experts under Top-P routing. The visualization reveals a "dense-to-sparse" transition, where the model adaptively allocates high computational resources in shallow layers and shifts towards sparse specialization in deeper layers.

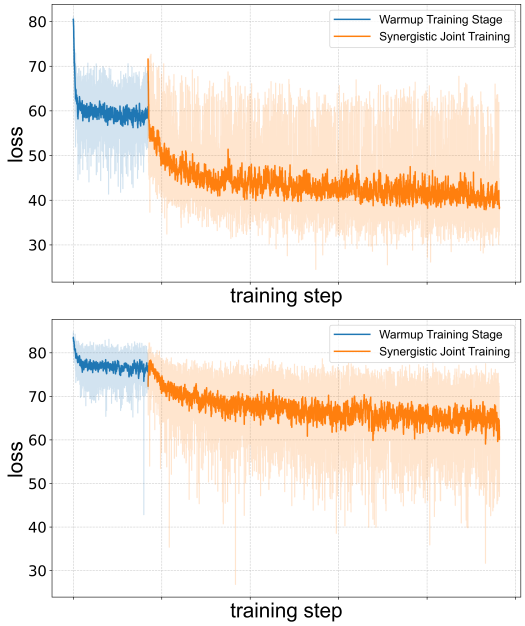

Figure 4: Training loss for the speech generation task (top) and music generation task (bottom). The plots show the transition from the Warmup Training Stage (blue) to the Synergistic Joint Training Stage (orange). The solid line represents the moving average of the loss.

A Expert Allocation Analysis

To understand how our model utilizes its dynamic capacity, we analyze the distribution of activated experts (ranging from 1 to 4) across different layers, as visualized in Figure 3. The results reveal a distinct "dense-to-sparse" transition in computational allocation.

In the shallow-to-middle layers (e.g., layers 0–12), the model maintains a high computational budget, with the majority of tokens activating 3 or 4 experts. The reason may be that the model is heavily engaged in extracting low-level features and fusing multimodal contexts at this stage, requiring the collective knowledge of most experts. However, as information propagates to the deeper layers (e.g., layers 13–27), a clear shift towards specialization emerges. The number of activated experts significantly decreases, with a growing proportion of tokens utilizing only 1 or 2 experts. This suggests that as the representations become more abstract and disentangled, the model becomes confident enough to route tokens to specific, task-dedicated experts, thereby pruning unnecessary computation.

This layer-wise adaptive behavior highlights the superiority of our confidence-based routing over static Top-K strategies. While Top-K enforces a constant cost regardless of complexity, our approach allows the model to self-regulate. This confirms that UniMoE-Audio effectively learns a hierarchical processing strategy, optimizing the trade-off between performance and efficiency.

B Expert Routing Visualization

To unravel the internal decision-making process of UniMoE-Audio, we visualize the expert routing patterns for the MLP (Figure 5). These figures present detail the task-specific activation ratios (Music vs. Speech) for each expert. Our analysis reveals three profound insights into the model's emergent behavior.

The visualization demonstrates a remarkable equilibrium in expert utilization. Across all layers, the workload is evenly distributed among the four routed experts, preventing the common *expert collapse* phenomenon where a few experts dominate the computation. However, a closer inspection of individual experts reveals a distinct *micro-level specialization*. Consistent with our initialization strategy, Experts 1 and 2 exhibit a strong preference for Speech tokens, while Experts 3 and 4 are predominantly activated by Music tokens. This clear division of labor validates the efficacy of our proto-expert specialization: the model successfully retains the domain-specific priors injected during pre-training, allowing it to route tokens to the most qualified specialists rather than learning from scratch.

C Training Dynamic Analysis

Figure 4 visualizes the loss trajectories across our three-stage curriculum, offering empirical evidence for the challenges inherent in multi-

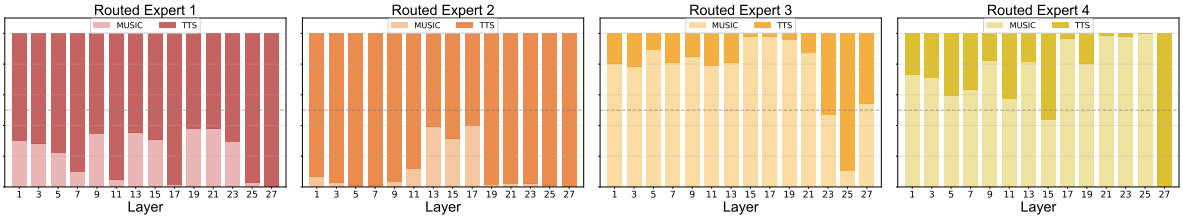

Figure 5: Visualization of layer-wise expert routing statistics. Experts 1 and 2 exhibit a predominant activation for Speech, whereas Experts 3 and 4 specialize in Music.

modal unification. In Stage 1, the distinct loss scales—significantly higher for music tasks compared to speech—highlight the intrinsic disparity in task complexity. This gap suggests that naive joint training would likely bias optimization toward easier tasks, validating our decision to pre-train isolated specialists. Crucially, the sharp loss reduction observed in Stage 3.1 confirms the necessity of the warmup phase; it indicates that calibrating the routing mechanism is non-trivial and essential for aligning the newly integrated Mix-of-Experts layers before full adaptation. Finally, the increased volatility during Stage 3.2 reflects the persistent tension in multi-task optimization. This variance underscores the value of our dynamic architecture, which absorbs these conflicts through conditional computation, preventing the catastrophic interference that typically plagues dense models in such heterogeneous landscapes.