UniMoE-Audio: Unified Speech and Music Generation with Dynamic-Capacity MoE

Zhenyu Liu, Yunxin Li, Xuanyu Zhang, Qixun Teng, Shenyuan Jiang, Xinyu Chen, Haoyuan Shi, Haolan Chen, Fanbo Meng, Mingjun Zhao, Yu Xu, Yancheng He, Yaowei Wang, Baotian Hu, Min Zhang

***Abstract*—Recent advancements in unified multimodal models indicate a clear trend towards comprehensive content generation, yet 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 fundamental task conflicts and severe data imbalances, which impede the development of a truly unified audio generation model. To address this challenge, we propose UniMoE-Audio, a novel approach that unifies speech and music generation within a single Mixture-of-Experts (MoE) framework. Architecturally, UniMoE-Audio introduces a dynamic-capacity routing mechanism using Top-P sampling for dynamic expert allocation, and a hybrid expert design that decouples domain-specific computation (dynamic experts) from universal features (shared experts). To enable effective joint learning from imbalanced data, we introduce a three-stage training curriculum: 1) Independent specialists training by training dense "proto-experts" on domain-specific data; 2) integrating them into the MoE architecture and warming up the new routing gates on a curated seed dataset; and 3) conducting synergistic joint training on the full balanced dataset. Extensive experiments show that UniMoE-Audio not only achieves state-of-the-art performance 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 architectures and curated training strategies in advancing the field of universal audio generation.**

I. 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. Significant progress has been made in systems that jointly process text, images, video, and even speech within a single architecture [1], [2], [3], [4], [5], [6]. Nevertheless, a critical imbalance persists in the treatment of the auditory domain. While speech has been a primary focus of integration [5], [6], music—a domain of comparable complexity and cultural richness—remains largely siloed and excluded from these unified frameworks. This fundamental omission not only curtails the ambition 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 the speech and music generation. The former is primarily concerned with semantic intelligibility and speaker identity, whereas the latter focuses on capturing complex structures 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 Mixture-of-Experts (MoE) paradigm has emerged as a promising architecture for mitigating conflicts of multimodal understanding [7], [8], [4], its application and further optimization for unified audio generation remain an open question. Moreover, this issue is compounded by the second challenge: **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 [9]. Consequently, a naive jointly training approach often allows the data-rich speech task to dominate the learning process, leading to a substantial degradation in musical quality. Our preliminary experiments empirically confirm this degradation (Figure 1a), 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 central scientific question we address is: *how to overcome both task conflict and data imbalance, enabling a shared model to master speech and music generation synergistically?*

To address these challenges, we introduce UniMoE-Audio, which leverages a novel dynamic-capacity MoE for mitigating task conflict. Instead of directly applying the conventional MoE, we provide two key architectural optimization to improve both routing flexibility and functional decoupling. First, we introduce a Top-P routing strategy that replaces the conventional fixed-capacity routing. Based on the Top-P sampling, this strategy dynamically adjusts the number of experts allocated to each token based on their complexity., thus enabling more flexible expert combinations. Second, we present a hybrid expert design to establish clear functional specialization, comprising: (1) conditional routed experts for domain-specific knowledge; (2) constantly active shared experts to handle domain-agnostic

Zhenyu Liu, Yunxin Li, Xuanyu Zhang, Qixun Teng, Shenyuan Jiang, Xinyu Chen, Haoyuan Shi, Yaowei Wang, Baotian Hu and Min Zhang are with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. (e-mail: liuzhenyuhit@gmail.com, hubaotian@hit.edu.cn, and zhangmin2021@hit.edu.cn)

Yaowei Wang and Baotian Hu are also with the Pengcheng Laboratory, Shenzhen, China.

Haolan Chen, Fanbo Meng, Mingjun Zhao, Yu Xu and Yancheng He are researchers with the Tencent PCG group, Shenzhen, China. (e-mail: haolanchen@tencent.com, fanbomeng@tencent.com, henrysxu@tencent.com, collinhe@tencent.com)

Baotian Hu is the corresponding author. (e-mail: hubaotian@hit.edu.cn)

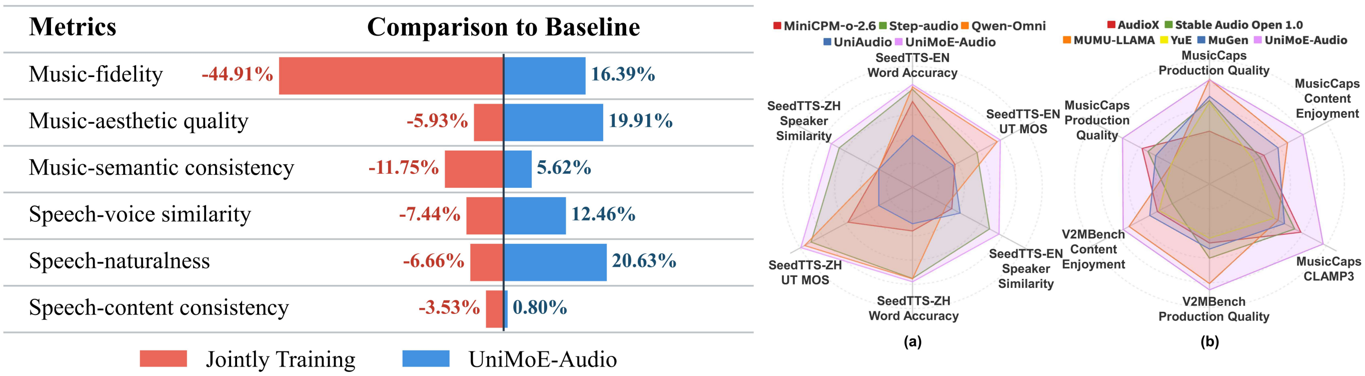

Fig. 1: Performance of UniMoE-Audio. **(Left)** Comparison against specialized baselines reveals the failure of naive joint training, which causes a clear performance degradation on speech generation and more significant decline on music generation. In contrast, our UniMoE-Audio yields synergistic gains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

features; and (3) null experts to enable adaptive computation skipping.

While our architecture provides the structural means to mitigate task conflict, we introduce a tightly coupled, three-stage training curriculum to address data imbalance. The curriculum unfolds as follows: (1) **Independent Specialist Training** leverages the original, uncurated datasets to instill domain-specific knowledge into each “proto-expert” without interference. (2) **MoE Integration and Warmup** then integrates these specialists into the UniMoE-Audio architecture. This stage begins by creating a curated, balanced dataset via a rigorous data filtering pipeline. To ensure training stability, the newly added components (i.e. the gate module and the shared expert) are then exclusively warmed up on a small subset of this curated data. (3) **Synergistic Joint Training** finally trains the entire, model end-to-end on the full, curated balanced dataset, fostering effective knowledge transfer across domains.

Our main contributions can be summarized as follows:

- We propose UniMoE-Audio, a unified speech and music generation model based on a novel dynamic-capacity Mixture-of-Experts framework. By integrating a Top-P routing strategy for adaptive resource allocation and a hybrid expert design for functional decoupling, our architecture effectively mitigates the inherent task conflict between speech and music generation.
- To leverage this architecture and tackle data imbalance, we introduce a data-aware, three-stage training curriculum. This curriculum systematically overcomes the data imbalance challenge by orchestrating independent specialist training, router warmup, and synergistic joint training, enabling robust and effective learning from highly imbalanced data sources without resorting to conventional data sampling.
- We provide extensive experiments to show the UniMoE-Audio’s effectiveness, achieving state-of-the-art or competitive performance on major speech and music generation benchmarks. Furthermore, our in-depth analysis reveals the dynamic activation patterns of the MoE model, offering valuable insights into how unified MoE model navigates diverse audio generation tasks.

II. Related Work

A. Domain-Specific Audio Generation Models

a) Large Spoken Models: The paradigm of generative AI, powered by Large Language Models (LLMs), has recently catalyzed a revolution in text-to-speech (TTS), giving rise to the field of Large Spoken Models. This approach fundamentally reframes speech synthesis as a conditional language modeling problem. Typically, a Speech LLM consists of a large, decoder-only Transformer and a neural audio codec. Given a textual prompt and optional voice conditions, the Transformer autoregressively generates a sequence of discrete audio tokens, which are then converted back into a continuous waveform by the codec. This framework has enabled unprecedented capabilities in zero-shot voice cloning and expressive, controllable speech generation. The seminal work in this area, VALL-E [10], pioneered this approach by discretizing speech into acoustic tokens via the EnCodec [11] and modeling them conditioned on text. This breakthrough laid the groundwork for a proliferation of subsequent models, including VALL-E X [12], SpearTTS [13], and Make-a-Voice [14], which further refined tokenization schemes and text-to-acoustic alignment. Building on this foundation, the field has seen rapid advancements towards greater robustness and versatility. For instance, CosyVoice [15] leverages a multi-task, multi-stage training curriculum to achieve state-of-the-art performance across a wide array of speech synthesis tasks. Concurrently, StepAudio [6] demonstrate the power of training on massive-scale synthetic data to produce exceptionally high-fidelity speech with rich emotional and stylistic diversity.

b) Large Music Models: Mirroring the evolution in speech synthesis, the field of music generation has also increasingly adopted the Large Language Model paradigm, re-framing music composition as a sequence generation task guided by textual or visual prompts. While diffusion-based models like

MusicLM [16] and Stable Audio Open [17] have achieved remarkable results, autoregressive models have demonstrated a compelling alternative. MusicGen [18] was a pivotal work that validated the feasibility of modeling music with a single Transformer decoder, generating high-fidelity music from discrete tokens. Pushing the boundaries further, subsequent works have explored more complex architectures and functionalities. Built upon the architecture of Llama2 [19], YuE [20] introduced a track-decoupled prediction strategy to handle long-form music generation. MuMuLlama [21] introduce multimodal music generation by jointly training on text-to-music and vision-to-music tasks. These advancements collectively establish that the powerful and viable of autoregressive framework for controllable music synthesis.

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. This line of inquiry is prompted by the observation that both fields, despite their distinct objectives, have independently converged on a similar technical paradigm: autoregressive modeling of discrete audio tokens. This parallel evolution suggests the potential for a single unified architecture that master both speech and music generation, yet the feasibility and inherent complexities of such unification remain largely unexplored. Therefore, our work represents a foundational investigation into this underexplored area, aiming to broaden the scope of what generative audio models can achieve.

B. Unified Audio Generation Models

The ambition of a universal audio model has prompted several initial investigations into unifying diverse audio generation tasks within a single framework. A notable early attempt, UniAudio [9], proposed a general-purpose text-to-audio model capable of generating various audio. However, as a naive co-training approach, it reportedly suffered from the effects of severe data imbalance, leading to limited performance on data-scarce tasks such as music generation. More recently, AudioX [22] demonstrated impressive capabilities in generating sound effects and music from multimodal inputs like text, images, and video, utilizing a Diffusion Transformer architecture. While powerful, its scope notably omits speech generation, a prevalent and critical audio modality, thus not addressing the full challenge of speech-music unification. In contrast to these approaches, our work directly confronts the core challenges that have hindered previous unification efforts. Rather than relying on simple joint training, we propose a framework that explicitly accounts for the inherent differences between audio modalities. Specifically, we leverage a Mixture-of-All architecture to mitigate task conflict and a data-aware, three-stage training curriculum to address data imbalance, aiming to provide a more principled and effective pathway toward truly unified and high-fidelity audio generation.

III. UniMoE-Audio

Our proposed model, UniMoE-Audio, is a unified generative framework designed to synthesize both speech and music from multimodal inputs, including text, audio, and video. As illustrated in Figure 2, the core innovation of the architecture lies in the Dynamic-Capacity Mixture-of-Experts implementation, which deviates from conventional MoE in two aspects: (1) a novel Top-P routing strategy for dynamic token-level expert number allocation, and (2) a hybrid expert design comprising routed, shared, and null experts.

A. Input Representation and Tokenization

a) 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 [23], which represents each audio frame using a multi-channel codebook. Unlike some works [24], [9] that employ a additional 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.

b) Visual Embedding: To process visual inputs (e.g., from video), we follow the Qwen-VL [25], using Visual Transformer (ViT) first encodes 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.

B. Dynamic-Capacity MoE

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 allocate the number of activated experts for each token based on the routing probability of gate 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 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:

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

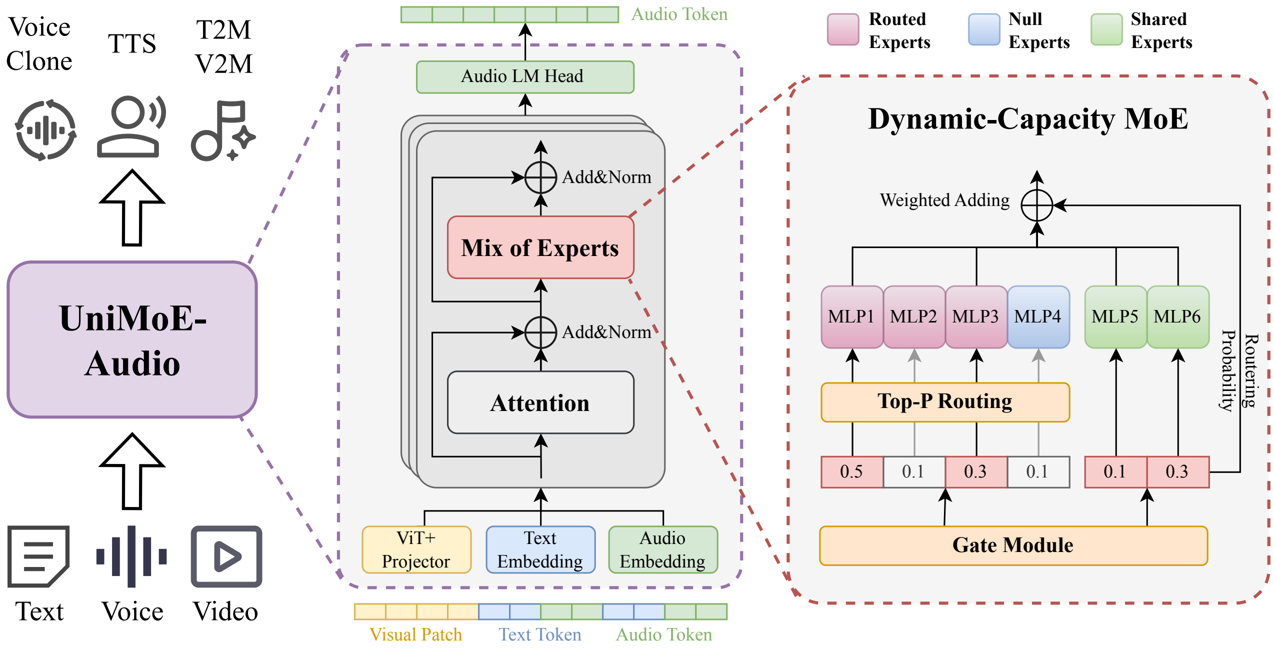

Fig. 2: An overview of the UniMoE-Audio framework. **Left:** UniMoE-Audio is a unified model capable of performing speech and music generation by leveraging multimodal conditional inputs, including Voice Cloning, Text-to-Speech (TTS), Text-to-Music (T2M), and Video-to-Music (V2M). **Center:** The core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

To efficiently solve this, we employ the classic Top-P sampling algorithm. For each token, we sort the expert probabilities in descending order and accumulate them until the sum reaches the threshold p. The experts included in this sum are selected for computation. This approach naturally links the number of selected experts to the router's confidence. A high-entropy probability distribution (signifying low confidence) requires accumulating more experts to reach the threshold. Conversely, a low-entropy distribution (high confidence) meets the threshold with fewer experts, thus reducing computation.

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 indices for a given token, and $E_i(X)$ is the output of the i-th expert.

While routed experts excel at learning domain-specific knowledge through conditional activation, they are inefficient for acquiring common knowledge, as inactive experts are excluded from the learning process. To address this, we functionally decouple the expert pool. Specifically, we incorporates a set of shared experts that operate in parallel with the routed ones. which is constantly activated for all tokens, aiming at capturing common knowledge and offloading computational burden in routed experts, allowing the routed experts to dedicate their full capacity to mastering domain-specific patterns.

Furthermore, while our proposed routing strategy enables expert number allocation, the range of selected expert number is inherently constrained. For a set of $N_{r]}$ routed experts and the probability threshold p, the number of activated experts is confined to the range $[1, \lceil pN_r \rceil]$. This prevents true computation skipping for simple tokens or activating all the router experts for the most demanding ones, limiting the model's adaptive potential. To overcome this, we introduce the null expert: a parameter-free module whose output is a constant zero tensor. By incorporating N_n null experts into the routing pool, the effective number of activated routed experts activated now spans the expanded range of $[0, \lceil p(N_r + N_n) \rceil]$. This not only enhances the combinatorial flexibility of expert selection but also enables true adaptive computation skipping.

IV. 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 principled, three-stage training curriculum.

A. Training Data

To support the unified generation of speech and music, we constructed a comprehensive, multi-task dataset encompassing four distinct categories: Chinese voice cloning, English voice

TABLE I: Overview of Datasets Used in Different Tasks

Task	Datasets	Number	Duration (hours)
Voice Cloning	Chinese in-house data	180K	20K
	English in-house data	100K	10K
Text-to-Music	Free-music-archive [26]	106K	8.2K
	MusicNet [27]	320	37
	MU2Gen [21]	22K	1.2K
Video-to-Music	V2M [28]	20K	6̃00

cloning, text-to-music generation, and video-to-music generation. Our data curation process began with the collection of extensive raw data, which was then automatically annotated in detail using a large-scale model. Subsequently, we employed a rigorous pipeline for multi-metric filtering, deduplication, and cleaning to ensure data quality. For voice cloning, we compiled approximately 20,000 hours of single-speaker Mandarin Chinese data and 10,000 hours of single-speaker English data. Each sample consists of a voice prompt, the target audio clip, and its corresponding transcription. For music generation, the dataset comprises approximately 9.4K hours of text-music pairs and 600 hours of video-music pairs, with each audio clip having a duration of around 20 seconds.

B. 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 data-aware, three-stage training curriculum, designed to decouple task-specific learning from synergistic optimization.

a) 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 (i.e., Chinese TTS, English TTS, T2M, V2M). This complete isolation allows each model—which will serve as a "proto-expert"—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.

b) MoE Integration and Warmup: In this stage, the pre-trained experts are fused into the single MoE architecture, and their weights are initially frozen. The key challenge here is to stably integrate the newly introduced, randomly initialized components: the gate module and the shared experts. A naive joint training would expose the well-trained experts to arbitrary and potentially erroneous routing decisions, risking catastrophic forgetting of their specialized knowledge. To prevent this, we perform a crucial calibration step. Using only the small, balanced seed dataset, we exclusively train the gate modules and the shared expert. This allows the routers to learn meaningful dispatch patterns based on the experts' pre-trained specializations and stabilizes the shared expert before full-model training. The rapid initial decrease in training loss observed during this phase (Figure 3) underscores the initial instability and validates the necessity of this stabilization step.

c) Synergistic Joint Training: With a stable and calibrated routing mechanism in place, the final stage aims to foster synergistic learning across all tasks. We unfreeze the entire model and conduct end-to-end fine-tuning on the larger, balanced fine-tuning dataset. To maintain routing efficiency and prevent expert specialization collapse during joint training, we employ an auxiliary load-balancing loss. The weight of this loss is linearly annealed over the course of training. Initially, a high weight encourages the model to prioritize balanced expert utilization, promoting exploration. As training progresses, the weight decreases, shifting the optimization focus toward maximizing the primary sequence generation objective and exploiting the learned, efficient routing patterns for superior performance.

V. Experiments

A. UniMoE-Audio Setting

The setting of UniMoE-Audio are detailed in Table II. Since the activation parameter count for UniMoE-Audio is not constrain: it is at a minimum of 2.8B when only the shared expert is active (i.e., only null expert is activated), and reaches a maximum of 5.9B when 6 routed experts are activated. By randomly simulating , we calculate an average activation of 4.8B parameters, which is on par with a standard MoE model using a Top-4 routing strategy.

Our initialization process is designed to effectively transfer knowledge from specialized models into the unified model. We begin by training four independent dense models, each based on the Qwen2.5VL-3B architecture, on our four respective tasks. To construct the UniMoE-Audio model, the FFN block from each of dense models is split into two, yielding a total of eight routed experts. The ViT module inherits its parameters directly from the model trained on Video-to-Music, while other shared components like attention and layer normalization modules are initialized by averaging the parameters from all four dense models. The shared expert and the routing gates are randomly initialized. For our dense baseline (i.e UniMoE-Dense), we utilize the larger Qwen2.5VL-7B architecture to ensure a fair comparison.

B. Implementation Details

We employ the AdamW [31] optimizer in conjunction with a cosine learning rate scheduler to train our models in all stages. Subsequently, in the independent specialist training stage, we utilize 48 Ascend 910B NPUs for each dense model training, with a global batch size of 48 and a base learning rate of 1e-4. In the MoE integration and warmup stage, we utilize 196 Ascend 910B NPUs for MoE training, with a global batch size of 784 and a base learning rate of 3e-5. Notably, during this stage, only the routing module and shard experts are trained. Finally, in the synergistic joint training stage, we utilize 196 Ascend 910B NPUs, with a global batch size of 3136 and a base learning rate of 1e-5. We employ four expert parallelism, which means only two routed experts are loaded on each NPUs (8 routed expert in total).

TABLE II: Detailed Architecture of Uni-MoE and Comparison with two common topK setting

Name	Experts	Routing Strategy	Activated Param	Total Param
UniMoE-Audio-Dense	1	-	7.1B	7.1B
UniMoE-Audio-MoE-8A4	8	TopK (K=4)	4.7B	7.1B
UniMoE-Audio-MoE	8 routed; 1 null; 2 shard	TopP (P=0.7)	Avg: 4.8B (Min: 2.8B, Max: 5.9B)	7.1B

TABLE III: Performance on Music Generation

Dataset	Method	Task	PC↑	PQ↑	CE↑	CLAP↑	KL↓	CLaMP3↑	IS↑	FAD↓
MusicCap	YuE [20]	T2M	3.45	7.25	5.84	0.18	2.12	0.09	2.09	9.02
	Stable Audio Open 1.0 [17]	T2M	3.70	7.29	6.02	**0.30**	1.44	0.11	2.74	3.72
	AudioX [22]	T2M	5.00	6.67	6.14	0.25	**1.20**	**0.12**	**3.02**	**1.64**
	MusicGen [18]	T2M	4.78	7.37	6.57	0.26	1.21	0.10	1.68	7.02
	MUMU-LLAMA [21]	T2M	5.15	7.71	6.87	0.20	1.27	0.10	1.44	8.57
	UniMoE-Audio-Dense	T2M	5.66	6.48	5.30	0.14	1.57	0.07	1.57	9.64
	UniMoE-Audio-MoE	T2M	**6.00**	**7.77**	**7.34**	0.29	1.39	**0.12**	1.93	6.43
V2M-bench	YuE [20]	T2M	3.78	7.25	6.01	0.15	1.27	0.13	1.79	4.29
	Stable Audio Open 1.0 [17]	T2M	3.41	7.46	5.69	**0.34**	1.91	0.16	3.13	2.94
	AudioX [22]	T2M	4.60	7.30	6.06	0.30	2.12	0.11	**3.64**	4.26
	MusicGen [18]	T2M	4.64	7.37	6.24	0.28	1.27	0.15	1.70	3.39
	MUMU-LLAMA [21]	T2M	5.19	7.73	6.75	0.17	**0.92**	0.13	1.42	**2.54**
	UniMoE-Audio-Dense	T2M	5.71	6.68	5.83	0.23	1.89	0.15	1.83	3.27
	UniMoE-Audio-MoE	T2M	**5.75**	**7.58**	**6.85**	0.31	1.06	**0.19**	2.17	3.11
V2M-bench V2M-bench	AudioX [22]	VT2M	4.44	7.44	6.06	-	1.84	-	3.14	2.94
	UniMoE-Audio-MoE	VT2M	5.88	7.62	6.96	-	1.69	-	3.31	2.89

TABLE IV: Performance on Speech Synthesis

Method	SeedTTS-EN			SeedTTS-ZH			librispeech		AISHELL-3	
	WER↓	UTMOS ↑	SIM↑	CER↓	UTMOS↑	SIM↑	WER↓	UTMOS ↑	CER↓	UTMOS ↑
UniAudio [22]	7.2	3.46	0.40	-	-	-	20.2	3.26	-	-
Mini-CPM-O-2.6 [29]	3.4	3.49	0.36	13.0	2.94	0.47	11.1	3.76	13.1	3.30
Qwen2.5-Omni [3]	2.1	4.16	-	1.6	3.28	-	7.6	4.19	2.5	3.38
Step-audio [6]	2.2	3.84	0.52	1.0	3.23	0.62	5.0	4.37	2.7	3.69
Higgs audio V2 [30]	**1.0**	4.00	**0.67**	10.8	3.27	**0.73**	**3.6**	**4.26**	5.9	**3.89**
UniMoE-Audio-Dense	2.5	3.67	0.47	2.0	3.29	0.57	10.8	3.97	4.2	3.45
UniMoE-Audio-MoE	1.9	**4.36**	0.56	**0.8**	**3.73**	0.65	4.4	4.23	**1.6**	3.86

C. Evaluation Datasets

We employ the benchmark of Seed-TTS [32] for English and Chinese speech synthesis and voice cloning, as well as LibriSpeech [33] and AISHELL-3 [34] for speech synthesis. Seed-TTS contains 1088 English speech samples and 2020 Chinese speech samples, with each samples contain one reference utterance. For LibriSpeech [33], we use the libriSpeech-test-clean set which contains 2619 English speech samples. AISHELL-3 [34] contains 24773 Chinese speech samples.

For music generation, we evaluate our model on both text-to-music (T2M) and video-and-text-to-music (VT2M) tasks. The **MusicCaps** benchmark [16] is utilized for its rich expert-authored text descriptions reflecting human perceptual interpretations of music. To complement this, we employ the **V2M-bench** [28], for which we generate objective, content-grounded textual annotations using Gemini. These annotations serve a dual purpose: first, they are used for a direct T2M task, allowing us to assess performance on factual descriptions in contrast to the subjective style of MusicCaps. Second, they are paired with their corresponding video data to evaluate the model's advanced multi-modal capabilities in the VT2M task.

D. Evaluation Metrics

a) Speech Synthesis: For objective evaluation, we adopt the word error rate (WER) and Character Error Rate (CER) for English and Chinese speech synthesis separately, and speaker similarity (SIM) metrics for voice cloning. We utilize Whisper-large-v3 [35] and Paraformer-zh [36] as the automatic speech recognition (ASR) engines for English and Chinese. For SIM, we compute the cosine similarity of speaker embedding from a fine-tuned WavLM from Seed-TTS [32]. In addition, we quantify the perceptual quality and naturalness of generated speech via UTMOS [37], a neural MOS predictor trained on human rating data for subjective evaluation. We use a predefined voice prompt for WER, CER and UTMOS to isolates the influence of voice quality.

b) Music Generation: We evaluate the generated music in terms of fidelity, semantic relevance, and aesthetic quality. Audio fidelity and diversity are measured using Fréchet Audio Distance (FAD) [38], where embeddings are extracted with openl3 [39], Kullback–Leibler Divergence (KL) based on class probability distributions predicted by PaSST [40], and Inception Score (IS). For IS computation, instead of the original images-based Inception-V3 network, we extract features from the audio transformer PaSST model to capture audio-oriented representation. To capture semantic alignment between text prompts and generated music, we report CLAP score [41], computed as the cosine similarity between audio and text embeddings extracted by the pretrained CLAP model. Beyond this, for a more robust evaluation of semantic relevance, we also employ CLaMP3 [42], a framework trained on large-scale music–text data with diverse and semantically rich annotations, and equipped with a multilingual text encoder, which together provide a superior assessment of audio–text consistency. In addition, we adopt three specialized metrics—Production Complexity (PC), Production Quality (PQ), and Content Enjoyment (CE)—as proposed in [43], to assess the aesthetic aspects of musical generation.

E. Main Result

We conducted a systematic evaluation of our UniMoE-Audio model on a variety of music and speech generation tasks. The main results, presented in Table III and Table IV,

We evaluate UniMoE-Audio against state-of-the-art specialized models for music and speech generation. The results, summarized in Table III and Table IV, demonstrate that our single, unified model not only avoids the performance degradation typical of naive joint training but achieves competitive or even superior results in both domains.

a) Music Generation: In the realm of music generation (Table III), UniMoE-Audio demonstrates a remarkable strength in producing aesthetically pleasing content. On the MusicCap benchmark, our MoE model is the only method to achieve the top scores across all three human-rated aesthetic metrics: PC, PQ, and CE. This indicates that despite being trained on speech data, its ability to generate rich, high-quality music is not compromised but enhanced. This pattern of superior aesthetic quality is replicated on the V2M-bench dataset, where our model again leads in PC, PQ, and CE, while also achieving the best semantic alignment (CLaMP3). This proves our model's capability to generate high-fidelity music that aligns well with diverse textual descriptions, a key challenge for unified models.

b) Speech Synthesis: The model's prowess in speech synthesis is equally compelling (Table IV). For English synthesis on SeedTTS-EN, UniMoE-Audio-MoE establishes a new state-of-the-art in intelligibility, reducing the WER to just 1.9, while simultaneously achieving the highest perceptual quality (UTMOS 4.36). This performance surpasses strong, dedicated TTS models like Higgs audio v2. Similarly, for Mandarin synthesis on SeedTTS-ZH, our model dominates by achieving the best scores across all metrics, including the lowest CER of 0.8.

This consistent, top-tier performance across two distinct languages and multiple benchmarks is the strongest evidence of our model's success. It demonstrates that UniMoE-Audio, as a single unified system, can master both the semantic precision of speech and the complex artistry of music without sacrificing performance in either domain, effectively overcoming the long-standing challenge of task conflict.

F. Ablation Study

To isolate the contribution of our proposed architecture, we conducted an ablation study comparing the full UniMoE-Audio-MoE model with its dense counterpart, UniMoE-Audio-Dense, which has a similar parameter count but lacks the expert-based structure. The results, presented in Tables III and IV, reveal a stark and consistent performance gap across all tasks. For instance, in speech synthesis on LibriSpeech, the MoE model achieves a WER of 4.4, more than halving the 10.8 WER of the dense model. A similar dramatic improvement is seen in Mandarin synthesis (SeedTTS-ZH), where the CER drops from 2.0 to a mere 0.8. In music generation, the MoE model consistently achieves higher aesthetic scores, indicating superior quality.

This significant and universal improvement strongly suggests that the dense model suffers from severe negative interference, where the conflicting optimization objectives of music and speech degrade performance in both domains. In contrast, the MoE architecture effectively mitigates this task conflict by allowing different experts to specialize. This result provides compelling evidence that a specialized, expert-based structure is not just beneficial but essential for building a high-performance, unified audio generation model.

VI. VISUALIZATION

A. Loss Visualization

To analyze the training dynamics and validate our two-stage training curriculum, we visualize the training loss for both TTS and Music tasks in Figure 3. The training process is divided into the MoE Warmup stage (Stage 1, blue curve) and the Synergistic Joint Training stage (Stage 2, orange curve).

First, the figures clearly demonstrate the effectiveness of the warmup stage. In both tasks, we observe a smooth transition from Stage 1 to Stage 2. Crucially, at the beginning of Stage 2 (step 5000), where all model parameters are unfrozen for joint training, the loss experiences a sharp, immediate drop without any initial spikes or instability. This smooth handover validates our warmup strategy, proving that it successfully calibrates the randomly initialized routing gates and shared expert. This pre-training prevents the training collapse that could occur from abrupt joint optimization and allows the full model to start learning synergistically from the outset.

Second, a comparison between the two figures reveals the divergent learning dynamics of speech and music generation. The music generation task (Figure 4) consistently exhibits a significantly higher overall loss than the TTS task (Figure 3) throughout both stages, eventually converging to a loss of approximately 60, compared to 40 for TTS. Furthermore, the convergence process for music is visibly less stable, characterized by greater loss variance (a wider shaded area) and a more jagged descent. These observations empirically support our initial hypothesis that music generation, with its complex structures of harmony and rhythm, is an inherently more challenging task for the model to learn compared to speech synthesis. This inherent difficulty underscores the necessity of our specialized architecture and curated training strategy to

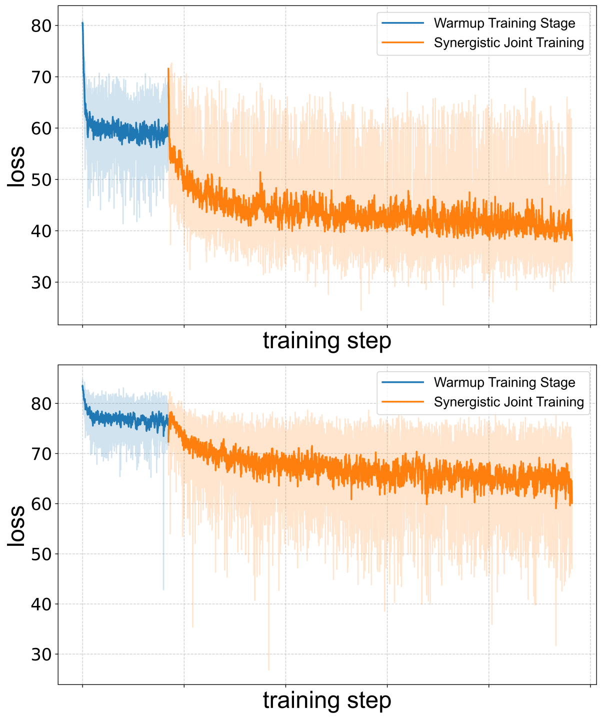

Fig. 3: Training dynamics of UniMoE-Audio's two-stage curriculum for TTS (top) and Music (bottom) tasks. The Warmup Training Stage (blue) exclusively trains the new routing gates and shared expert. The subsequent Synergistic Joint Training stage (orange) trains the entire model end-to-end. The sharp, immediate drop in loss at the transition point (step 5000) validates the warmup strategy's effectiveness in ensuring a stable start to joint training. Furthermore, the consistently higher and more volatile loss curve for music generation empirically demonstrates its greater learning difficulty compared to speech, underscoring the need for our proposed methods.

prevent the simpler TTS task from dominating the learning process.

B. Allocated Expert Number Visualization

To investigate the operational dynamics of our Top-P routing strategy, we analyze the distribution of the number of activated experts per token across different layers, as shown in Figure [Your Figure Number]. The visualization reveals a clear pattern of hierarchical computation. In the initial, shallower layers (e.g., layers 0-3), a significant proportion of tokens are routed to a smaller number of experts (1-3). As the model deepens, it allocates a larger computational budget, with the majority of tokens in the middle layers (approximately 4-13) activating 4-5 experts, peaking around layer 12. Subsequently, in the final layers (14-17), the trend reverses, and the allocated budget decreases. This ”rise-and-fall” pattern strongly suggests that the model has learned to concentrate its most intensive computations in the middle layers for complex feature abstraction, while using fewer resources for initial feature extraction (shallow layers) and final integration (deep layers).

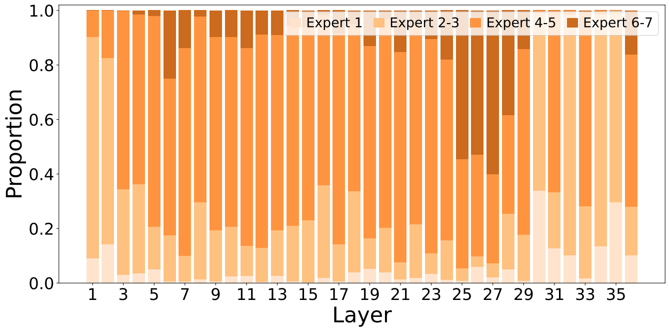

Fig. 4: Visualization of the dynamic computational budget allocated by our Top-P routing mechanism. The figure illustrates the proportion of tokens activating a varying number of experts at each layer, demonstrating the model's ability to adaptively assign more computational resources to the middle layers.

Crucially, this dynamic allocation pattern highlights a core advantage of our approach. Unlike a conventional Top-K router that would enforce a rigid, uniform number of experts for all tokens, our figure demonstrates true token-level adaptive computation at every layer. Even within the same layer, the model dynamically assigns a larger budget to ”hard” tokens while conserving resources on ”easy” ones. This inherent flexibility validates the efficacy of Top-P routing in creating a more efficient and intelligent model that tailors its computational effort to the complexity of the task at hand.

C. Expert Routing Visualization

To delve into the internal mechanisms of UniMoE-Audio and validate our design, we visualize the expert routing statistics across all transformer layers, as shown in Figure [Your Figure Number]. The figure is composed of a summary plot for all experts (top-left) and individual plots detailing the task-specific activation (Music vs. TTS) for each of the eight routed experts (E1-E8) and the null expert (E9). The analysis reveals several key operational dynamics of our model.

a) Balanced Expert Utilization: The “All Experts” subplot demonstrates a remarkably balanced utilization of all experts across the layers. No single routed expert dominates the computation, and the null expert (E9) is also consistently selected. This indicates that our training methodology, particularly the auxiliary load-balancing loss, is highly effective in encouraging a distributed workload and preventing expert collapse—a common failure mode in MoE training. Each expert is actively contributing, fulfilling its role as a functional component of the larger model.

b) Clear Task Specialization: The individual expert plots (E1-E8) provide striking evidence of task specialization. Each expert exhibits a distinct and consistent preference for either the TTS or the Music task. For instance, Expert 1 is overwhelmingly activated by TTS tokens, whereas Expert 5 shows a strong preference for Music tokens. This strong specialization is direct evidence that our three-stage training curriculum, especially the initialization from pre-trained ”proto-experts,” successfully instills and preserves domain-specific

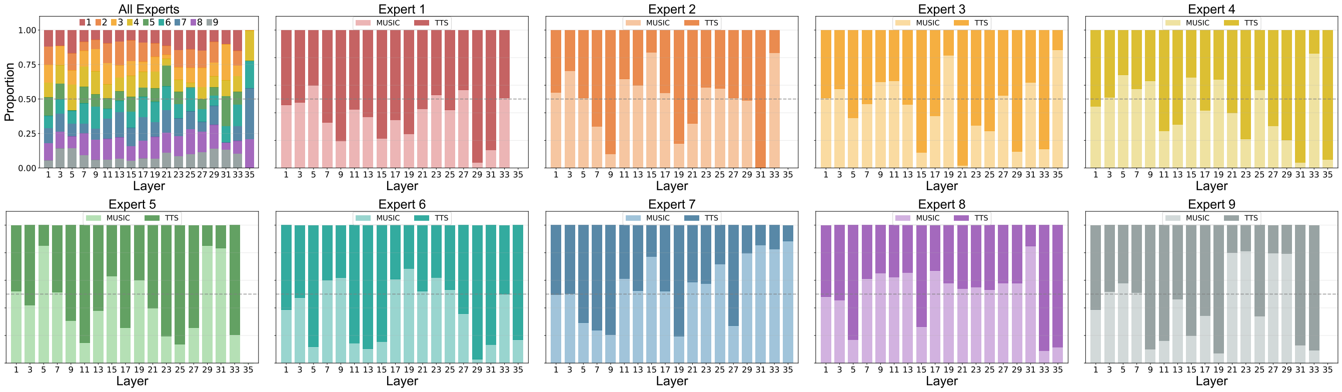

Fig. 5: Analysis of expert routing dynamics in UniMoE-Audio across transformer layers. The top-left ”All Experts” plot illustrates the routing frequency for each of the eight routed experts (E1-E8, colored) and the null expert (E9, gray), confirming balanced expert utilization. The subsequent nine plots provide a granular breakdown for each expert, showing the proportion of tokens routed from the Music (lighter shade) versus the TTS (darker shade) task. These plots reveal strong evidence of task specialization, where experts develop clear preferences (e.g., Expert 1 for TTS, Expert 5 for Music), validating our specialist pre-training strategy. Furthermore, the more balanced activation in the initial layers suggests a hierarchical processing scheme, where shallow layers handle universal features before deeper layers engage in domain-specific computation.

knowledge within designated experts. The model has learned to route tokens to the specialist best equipped to handle them.

c) Emergent Hierarchical Processing: A more nuanced, layer-wise pattern emerges upon closer inspection. While specialization is strong in the middle and deeper layers, the initial layers (e.g., layers 1-3) of most experts show a more balanced activation between TTS and Music. This suggests an emergent hierarchical processing scheme within the model. The shallower layers appear to handle more universal, low-level audio features common to both speech and music (e.g., basic frequency components or textures). As information propagates to deeper layers, the model routes tokens to experts that process more abstract, domain-specific representations (e.g., phonetic content for TTS, harmonic structures for Music).

d) The Role of the Null Expert in Adaptive Computation: The behavior of the null expert (Expert 9) provides critical insights into the model's adaptive computation capabilities. The plot for Expert 9 shows that it is activated by tokens from both Music and TTS tasks. This is a crucial finding: it indicates that the model learns to identify computationally simple or redundant tokens within each domain and dynamically skips their FFN computation by routing them to the parameter-free null expert. This is not just a mechanism for balancing load but a true form of learned computational efficiency, directly validating the utility of our null expert design for creating a genuinely dynamic-capacity model.

VII. Conclusion

In this paper, we addressed the long-standing challenge of unifying speech and music generation, a task hindered by task conflict and data imbalance. We introduced UniMoE-Audio, a framework that leverages a dynamic-capacity Mixture-of-Experts architecture to mitigate task conflict and a data-aware, three-stage training curriculum to overcome data imbalance. Extensive experiments demonstrate that our approach not only achieves state-of-the-art performance but, more importantly, fosters synergistic learning, effectively resolving the performance degradation seen in naive joint training. Our work provides a robust blueprint for building unified generative audio models, with future directions including the incorporation of a broader range of audio types and the exploration of model compression techniques.

References

[1] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li, H. Yan, J. Fu, T. Gui, T. Sun, Y. Jiang, and X. Qiu, “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*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 9637–9662. [Online]. Available: https://doi.org/10.18653/v1/2024.acl-long.521

[2] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen *et al.*, “Qwen-image technical report,” *arXiv preprint arXiv:2508.02324*, 2025.

[3] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, “Qwen2.5-omni technical report,” *CoRR*, vol. abs/2503.20215, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.20215

[4] I. AI, B. Gong, C. Zou, C. Zheng, C. Zhou, C. Yan, C. Jin, C. Shen, D. Zheng, F. Wang, F. Xu, G. Yao, J. Zhou, J. Chen, J. Sun, J. Liu, J. Zhu, J. Peng, K. Ji, K. Song, K. Ren, L. Wang, L. Ru, L. Xie, L. Tan, L. Xue, L. Wang, M. Bai, N. Gao, P. Chen, Q. Guo, Q. Zhang, Q. Xu, R. Liu, R. Xiong, S. Gao, T. Liu, T. Li, W. Chai, X. Xiao, X. Wang, X. Chen, X. Lu, X. Li, X. Dong, X. Yu, Y. Yuan, Y. Gao, Y. Sun, Y. Chen, Y. Wu, Y. Lyu, Z. Ma, Z. Feng, Z. Fang, Z. Qiu, Z. Huang, and Z. He, “Ming-omni: A unified multimodal model for perception and generation,” *CoRR*, vol. abs/2506.09344, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2506.09344

[5] KimiTeam, D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang, Z. Wang, C. Wei, Y. Xin, X. Xu, J. Yu, Y. Zhang, X. Zhou, Y. Charles, J. Chen, Y. Chen, Y. Du, W. He, Z. Hu, G. Lai, Q. Li, Y. Liu, W. Sun, J. Wang, Y. Wang, Y. Wu, Y. Wu, D. Yang, H. Yang, Y. Yang, Z. Yang, A. Yin, R. Yuan, Y. Zhang, and Z. Zhou, “Kimi-audio technical report,” *CoRR*, vol. abs/2504.18425, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2504.18425

[6] A. Huang, B. Wu, B. Wang, C. Yan, C. Hu, C. Feng, F. Tian, F. Shen, J. Li, M. Chen, P. Liu, R. Miao, W. You, X. Chen, X. Yang, Y. Huang, Y. Zhang, Z. Gong, Z. Zhang, H. Zhou, J. Sun, B. Li, C. Feng, C. Wan, H. Hu, J. Wu, J. Zhen, R. Ming, S. Yuan, X. Zhang, Y. Zhou, B. Li, B. Ma, H. Wang, K. An, W. Ji, W. Li, X. Wen, X. Kong, Y. Ma, Y. Liang, Y. Mou, B. Ahmidi, B. Wang, B. Li, C. Miao, C. Xu, C. Wang, D. Shi, D. Sun, D. Hu, D. Sai, E. Liu, G. Huang, G. Yan, H. Wang, H. Jia, H. Zhang, J. Gong, J. Guo, J. Liu, J. Liu, J. Feng, J. Wu, J. Wu, J. Yang, J. Wang, J. Zhang, J. Lin, K. Li, L. Xia, L. Zhou, L. Zhao, L. Gu, M. Chen, M. Wu, M. Li, M. Li, M. Li, M. Liang, N. Wang, N. Hao, Q. Wu, Q. Tan, R. Sun, S. Shuai, S. Pang, S. Yang, S. Gao, S. Yuan, S. Liu, S. Deng, S. Jiang, S. Liu, T. Cao, T. Wang, W. Deng, W. Xie, W. Ming, and W. He, "Step-audio: Unified understanding and generation in intelligent speech interaction," *CoRR*, vol. abs/2502.11946, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.11946

[7] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Zhang, M. Ning, and L. Yuan, "Moe-llava: Mixture of experts for large vision-language models," *CoRR*, vol. abs/2401.15947, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.15947

[8] Y. Li, S. Jiang, B. Hu, L. Wang, W. Zhong, W. Luo, L. Ma, and M. Zhang, "Uni-moe: Scaling unified multimodal llms with mixture of experts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3424–3439, 2025. [Online]. Available: https://doi.org/10.1109/TPAMI.2025.3532688

[9] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, H. Guo, X. Chang, J. Shi, S. Zhao, J. Bian, Z. Zhao, X. Wu, and H. M. Meng, "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, 2024. [Online]. Available: https://openreview.net/forum?id=SRmZw7nEGW

[10] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers," *CoRR*, vol. abs/2406.05370, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.05370

[11] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *CoRR*, vol. abs/2210.13438, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.13438

[12] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *CoRR*, vol. abs/2303.03926, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.03926

[13] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 1703–1718, 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00618

[14] R. Huang, C. Zhang, Y. Wang, D. Yang, L. Liu, Z. Ye, Z. Jiang, C. Weng, Z. Zhao, and D. Yu, "Make-a-voice: Unified voice synthesis with discrete representation," *CoRR*, vol. abs/2305.19269, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.19269

[15] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, Z. Gao, and Z. Yan, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *CoRR*, vol. abs/2407.05407, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2407.05407

[16] A. Agostinelli, T. I. Denk, Z. Borsos, J. H. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. H. Frank, "Musiclm: Generating music from text," *CoRR*, vol. abs/2301.11325, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.11325

[17] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 2025, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICASSP49660.2025.10888461

[18] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "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*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *CoRR*, vol. abs/2307.09288, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.09288

[20] R. Yuan, H. Lin, S. Guo, G. Zhang, J. Pan, Y. Zang, H. Liu, Y. Liang, W. Ma, X. Du, X. Du, Z. Ye, T. Zheng, Y. Ma, M. Liu, Z. Tian, Z. Zhou, L. Xue, X. Qu, Y. Li, S. Wu, T. Shen, Z. Ma, J. Zhan, C. Wang, Y. Wang, X. Chi, X. Zhang, Z. Yang, X. Wang, S. Liu, L. Mei, P. Li, J. Wang, J. Yu, G. Pang, X. Li, Z. Wang, X. Zhou, L. Yu, E. Benetos, Y. Chen, C. Lin, X. Chen, G. Xia, Z. Zhang, C. Zhang, W. Chen, X. Zhou, X. Qiu, R. B. Dannenberg, Z. Liu, J. Yang, W. Huang, W. Xue, X. Tan, and Y. Guo, "Yue: Scaling open foundation models for long-form music generation," *CoRR*, vol. abs/2503.08638, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.08638

[21] S. Liu, A. S. Hussain, Q. Wu, C. Sun, and Y. Shan, "Mumu-llama: Multi-modal music understanding and generation via large language models," *CoRR*, vol. abs/2412.06660, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.06660

[22] Z. Tian, Y. Jin, Z. Liu, R. Yuan, X. Tan, Q. Chen, W. Xue, and Y. Guo, "Audiox: Diffusion transformer for anything-to-audio generation," *CoRR*, vol. abs/2503.10522, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.10522

[23] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "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*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html

[24] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *CoRR*, vol. abs/2410.00037, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.00037

[25] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *CoRR*, vol. abs/2409.12191, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2409.12191

[26] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "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*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 316–323. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/75_Paper.pdf

[27] J. Thickstun, Z. Harchaoui, and S. M. Kakade, "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, 2017. [Online]. Available: https://openreview.net/forum?id=rkFBJv9gg

[28] Z. Tian, Z. Liu, R. Yuan, J. Pan, Q. Liu, X. Tan, Q. Chen, W. Xue, and Y. Guo, "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*. Computer Vision Foundation / IEEE, 2025, pp. 18 782–18 793. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Tian_VidMuse_A_Simple_Video-to-Music_Generation_Framework_with_Long-Short-Term_Modeling_CVPR_2025_paper.html

[29] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

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

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[32] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, M. Gong, P. Huang, Q. Huang, Z. Huang, Y. Huo, D. Jia, C. Li, F. Li, H. Li, J. Li, X. Li, X. Li, L. Liu, S. Liu, S. Liu, X. Liu, Y. Liu, Z. Liu, L. Lu, J. Pan, X. Wang, Y. Wang, Y. Wang, Z. Wei, J. Wu, C. Yao, Y. Yang, Y. Yi, J. Zhang, Q. Zhang, S. Zhang, W. Zhang, Y. Zhang, Z. Zhao, D. Zhong, and X. Zhuang, "Seed-tts: A family of high-quality versatile speech generation models," *CoRR*, vol. abs/2406.02430, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.02430

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "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*. IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[34] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "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*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2756–2760. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-755

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[36] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "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*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2063–2067. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-9996

[37] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "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*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 4521–4525. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-439

[38] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "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*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2350–2354. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2219

[39] J. Cramer, H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 3852–3856. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8682475

[40] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "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*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2753–2757. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-227

[41] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "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*. IEEE, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICASSP49357.2023.10095969

[42] S. Wu, Z. Guo, R. Yuan, J. Jiang, S. Doh, G. Xia, J. Nam, X. Li, F. Yu, and M. Sun, "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*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Association for Computational Linguistics, 2025, pp. 2605–2625. [Online]. Available: https://aclanthology.org/2025.findings-acl.133/

[43] A. Tjandra, Y. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W. Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *CoRR*, vol. abs/2502.05139, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.05139