UniAudio 2.0: A Unified Audio Language Model with Text-Aligned Factorized Audio Tokenization

Dongchao Yang[1] **Yuanyuan Wang**[1] **Songxiang Liu**[2]
Dading Chong[3] **Xixin Wu**[1] **Helen Meng**[1] *
[1] **The Chinese University of Hong Kong** [2] **Meituan** [3] **Peking University**

Abstract

We present UniAudio 2.0, a unified audio foundation model built on FactorCodec, a purely discrete tokenizer that factorizes audio into two role-separated representations. Analysis tokens are optimized to retain language-aligned, text-expressible information that can be decoded by a text LLM head into grounded natural-language analyses, while reconstruction tokens preserve the information required for waveform synthesis and are used exclusively for audio decoding. This factorization yields understanding performance competitive with continuous Whisper features on audio understanding tasks, while reducing the perplexity of reconstruction-token modeling for generation. UniAudio 2.0 further adopts functional layer specialization, partitioning the backbone into audio-understanding, cross-modal, and audio-generation experts. We train the model with a four-stage recipe on 100B text tokens and 60B unique audio tokens, together with a composed audio sequence construction strategy for unified multi-task pre-training. At 3B parameters, UniAudio 2.0 is competitive with strong 7B unified baselines on in-domain tasks and exhibits non-trivial *compositional generalization* to unseen tasks under few-shot and zero-shot evaluation. Unlike prior unified systems that rely on hybrid continuous-discrete inputs or do not support generation beyond speech, UniAudio 2.0 provides a discrete interface for both understanding and generation across speech, sound, and music. Code and checkpoints are available at the project repository

1 Introduction

Large language models [OpenAI, 2023, Grattafiori et al., 2024] have shown that a single autoregressive backbone can unify a remarkably broad range of text tasks. Extending this recipe to audio is appealing: in principle, one model could support both understanding and generation across speech, sound, and music. Yet this goal remains challenging, because audio foundation models must simultaneously reason over dense perceptual signals and generate high-fidelity waveforms. A central question is therefore: *what discrete audio interface should an LLM operate on?*

Existing approaches [Xu et al., 2025a, KimiTeam et al., 2025] reflect a central representation trade-off: Continuous audio features, such as Pre-trained Whisper representations, preserve rich perceptual information and are effective for understanding, but they are inconvenient for scalable autoregressive generation and often require hybrid continuous-discrete interfaces KimiTeam et al. [2025], Xu et al. [2025a] or specialized decoders Yan et al. [2025]. Discrete neural codecs, in contrast, are naturally compatible with language model style generation, but often entangle language-relevant abstractions with waveform-synthesis details, making them a suboptimal interface for LLM-based understanding. As a result, existing unified systems either rely on hybrid continuous-discrete inputs KimiTeam et al. [2025], Xu et al. [2025a] or do not support generation beyond speech Tang et al. [2024], Chu et al. [2024], leaving open the question of whether a *purely discrete* interface can unify understanding and generation across speech, sound, and music.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

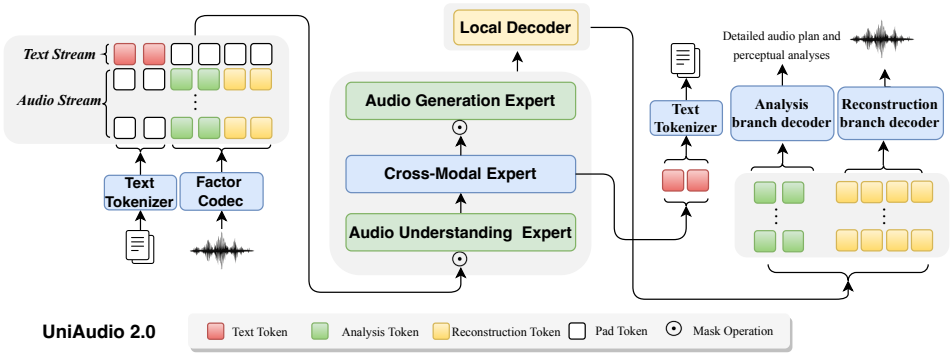

Figure 1: Overview of UniAudio 2.0. The figure illustrates a generation example. UniAudio 2.0 first predicts analysis tokens and then reconstruction tokens conditioned on them, while text positions activate only text tokens. The backbone is partitioned into audio-understanding, cross-modal, and audio-generation experts. Figure 2 shows the FactorCodec in detail.

We address this question with **FactorCodec**, a purely discrete tokenizer built on *role-separated factorization*. Instead of using a single token stream as a compromise representation, FactorCodec decomposes each audio signal into an *analysis stream* and a *reconstruction stream*. Analysis tokens are optimized to retain language-aligned, text-expressible information that can be decoded by a text LLM head into grounded natural-language analyses, while reconstruction tokens preserve the synthesis-oriented information used for waveform decoding. This separation yields a discrete interface that is better suited to both understanding and generation: it approaches the performance of continuous Whisper features on audio understanding tasks, while the factorized design reduces the perplexity of reconstruction-token modeling for autoregressive generation.

Building on this interface, we develop **UniAudio 2.0**, a unified autoregressive audio foundation model with *functional layer specialization*. Rather than processing text and audio uniformly in all transformer layers, UniAudio 2.0 partitions the backbone into audio-understanding, cross-modal, and audio-generation experts, preserving specialized inductive biases within a single autoregressive model. We train UniAudio 2.0 with a four-stage recipe on 100B text tokens and 60B unique audio tokens, together with *composed audio sequences* for scalable multi-task pre-training across speech, sound, and music. At 3B parameters, UniAudio 2.0 is competitive with strong 7B unified baselines on in-domain tasks and exhibits non-trivial *compositional generalization* [Lake and Baroni, 2018, Bai et al., 2023] to unseen tasks: novel task instantiations that recombine compositional skills induced during pre-training. To our knowledge, it is the first unified audio model to provide a *purely discrete* interface for both understanding and generation across speech, sound, and music. Our contributions include:

- **FactorCodec.** We introduce a purely discrete audio tokenizer that factorizes audio into an analysis stream and a reconstruction stream, yielding a role-separated interface for unified understanding and generation.
- **UniAudio 2.0 architecture and training.** We develop a unified autoregressive model with functional layer specialization, together with a four-stage training recipe and composed-audio-sequence construction for scalable multi-task pre-training.
- **Unified evaluation across domains.** We show that a single 3B model supports both understanding and generation across speech, sound, and music, achieving competitive in-domain performance and non-trivial compositional generalization to unseen tasks under few-shot and zero-shot evaluation.

2 Related Work

Audio language models. Recent audio language models have explored both audio understanding and audio generation within language model style frameworks Borsos et al. [2023a], Wang et al. [2023], Tang et al. [2024], Zhang et al. [2025], KimiTeam et al. [2025]. Understanding-oriented systems typically pair continuous audio representations with pre-trained LLMs, yielding strong perception

and reasoning performance Tang et al. [2024], Goel et al. [2025]. In parallel, autoregressive audio models based on discrete tokenization Défossez et al. [2022], Borsos et al. [2023b], Défossez et al. [2024] have achieved success in speech Wang et al. [2023] and music generation Agostinelli et al. [2023], Copet et al. [2023]. These directions highlight the promise of unified audio modeling, but also suggest that the choice of audio interface remains a central bottleneck.

Audio tokenization for understanding and generation. Existing audio representations can be broadly grouped into continuous features, acoustic discrete codecs, and semantic-rich tokenizers. Continuous features Chen et al. [2022], Radford et al. [2023], Baevski et al. [2020] preserve rich perceptual information and are effective for understanding, but are inconvenient for scalable autoregressive generation. Acoustic codecs enable LM-style generation, but typically emphasize waveform fidelity over language-aligned abstraction. More recent tokenizers inject semantic or linguistic supervision into discrete audio tokens Défossez et al. [2024], Zhang et al. [2023b], Ye et al. [2025a], Du et al. [2024], improving their usefulness for downstream tasks. Yet the underlying tokenization paradigm remains unchanged: semantic supervision is added on top of a VQ layer, all VQ levels are trained end-to-end for reconstruction and collectively serve waveform synthesis. In contrast, our FactorCodec introduces a role-separated discrete interface, decomposing audio into an analysis stream for language-grounded understanding and a reconstruction stream for synthesis.

Unified audio foundation models. Several recent systems move toward unified audio foundation models KimiTeam et al. [2025], Zhang et al. [2025], Tian et al. [2025b], Cai et al. [2026], combining understanding and generation within a single architecture. However, many rely on hybrid continuous-discrete inputs, or focus generation primarily on speech or sound. Our work differs in targeting a single purely discrete interface that supports both understanding and generation across speech, sound, and music, together with nontrivial transfer to unseen tasks.

3 FactorCodec: A Role-Separated Discrete Audio Interface

We introduce **FactorCodec**, a purely discrete tokenizer built on *role-separated factorization*. Instead of using a single token stream as a compromise representation, FactorCodec decomposes each audio signal into two discrete streams: an *analysis stream* for language-grounded understanding and a *reconstruction stream* for waveform synthesis.

Given an audio waveform x, FactorCodec maps it to two discrete token sequences,

$$(z^A, z^R) = \mathrm{FactorCodec}(x), \tag{1}$$

where z^A denotes the *analysis tokens* and z^R denotes the *reconstruction tokens*. The key property of FactorCodec is not merely the existence of two streams, but their *role-separated recoverability*: analysis tokens are optimized to preserve language-aligned, text-expressible information for understanding, while reconstruction tokens preserve the synthesis-oriented information used for waveform decoding. In our design, waveform reconstruction is performed only from the reconstruction stream $\hat{x} = D(z^R)$ rather than from the joint pair (z^A, z^R). In the following, we introduce the details of FactorCodec, including the analysis branch and reconstruction branch.

Analysis branch. The analysis branch is implemented as a low-rate discrete representation derived from frozen audio encoder features (Whisper and Music SSL encoders) via query-based compression and quantization [Yang et al., 2025]. Let $h \in \mathbb{R}^{T\times d}$ denote fused continuous audio features. We introduce learnable queries $Q \in \mathbb{R}^{M\times d}$ and summarize h into

$$z = \mathrm{Enc}(Q, h) \in \mathbb{R}^{M\times d}, \qquad M = \lceil T/W \rceil. \tag{2}$$

In our implementation, $W = 5$, yielding a low-rate sequence at approximately 5 Hz, which is then discretized using 8-codebook RVQ to obtain z^A. The analysis branch is trained with textual supervision so that a text LLM head can decode it into grounded natural-language analyses. The branch is trained in two stages: supervised fine-tuning (SFT) on a mixture of audio understanding tasks, followed by GRPO [Shao et al., 2024] with an LLM-based judge to encourage diverse grounded perceptual analyses (Appendix A.2). Previous approaches (e.g., CosyVoice 3 [Du et al., 2025]) also use multi-task supervision but retain waveform-reconstruction capability in the same token stream; our analysis tokens offload reconstruction to a separate branch, enabling a $5\times$ lower frame rate (5 vs. 25 Hz) and strict confinement to text content (refer to Appendix Box A.4). Refer to Appendix A.3.8 for more details about the difference between semantic tokens and the analysis tokens.

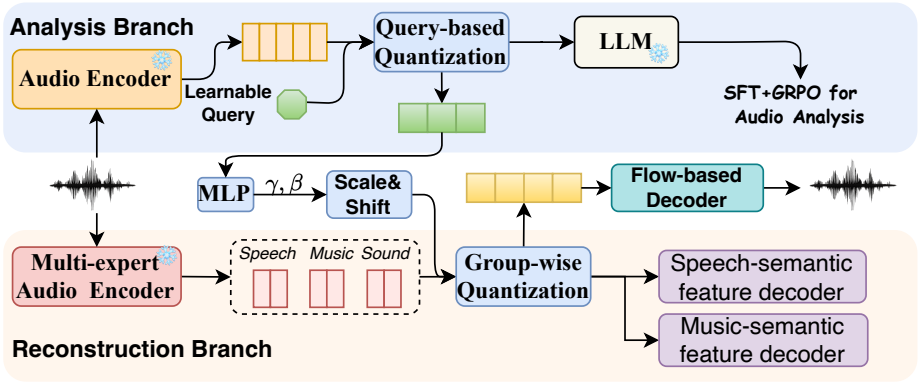

Figure 2: The overview of FactorCodec. The analysis branch (top) produces analysis tokens. The reconstruction branch (bottom) produces reconstruction tokens for waveform reconstruction. The two branches are coupled through FiLM-based feature modulation, providing high-level conditioning while preserving role separation.

Reconstruction branch. The reconstruction branch is built from multi-expert features and group-wise quantization for speech, music, and general audio (we explain why we choose group-wise quantization in Appendix A.3.2). We extract three groups of features from frozen experts: h^{ph} from WavLM [Chen et al., 2022] (phone-level speech semantics), h^{mu} from a music SSL encoder [Zhu et al., 2025] (music structure), and h^{env} from Whisper [Radford et al., 2023] (environmental sound and residual acoustics). We quantize them as

$$z^R = \big(\mathrm{VQ}_1(h^{ph}),\ \mathrm{VQ}_2(h^{mu}),\ \mathrm{RVQ}_{3:8}(h^{env})\big), \tag{3}$$

where the first two groups mainly capture structured semantic content and the remaining groups retain synthesis-relevant acoustic detail. Let $S_e = (h^{ph}, h^{mu}, h^{env})$ denote the pre-quantization features. To reduce redundancy between the analysis and reconstruction tokens, S_e is modulated by the analysis stream through FiLM [Perez et al., 2018] *before* group-wise quantization:

$$\mathrm{FiLM}(S_e; \hat{A}) = \gamma(\hat{A}) \odot S_e + \beta(\hat{A}), \tag{4}$$

where $\hat{A}$ denotes the quantized continuous representation of analysis tokens z^A. Importantly, the analysis stream is not used as a decoding representation; waveform synthesis remains determined by reconstruction tokens only. We decode the waveforms from z^R with a flow-based scalar latent diffusion decoder [Yang et al., 2024b]. The branch is trained with a flow loss $\mathcal{L}_{flow}$ and a semantic reconstruction loss Ye et al. [2025a] $\mathcal{L}_{sem}$ for h^{ph} and h^{mu}. More details of architecture and training are deferred to the Appendix A.2. In Appendix A.3.10, we analyze why the flow decoder is used.

Bitrate accounting. With codebook size 1024 (10 bits per codebook index), the 5-Hz analysis stream with 8 codebooks has a nominal rate of $5 \times 8 \times 10 = 400$ bps. The 12.5-Hz reconstruction stream with 8 codebooks has a nominal rate of $12.5 \times 8 \times 10 = 1000$ bps. Thus, the complete discrete interface contains 1400 bps, but waveform decoding uses only the 1000-bps reconstruction stream; the analysis stream is used for understanding and conditioning, not waveform decoding. These are rates for the evaluated configuration, not a multi-bitrate evaluation.

Why factorization helps. On the understanding side, the analysis token compresses audio into a low-rate, language-aligned representation, discarding acoustic details irrelevant to text-level understanding. On the generation side, separating out high-level structure reduces the uncertainty that would otherwise have to be modeled within the reconstruction stream alone, leading to lower reconstruction-token perplexity (refer to Figure 3). Together, these properties make FactorCodec a more suitable discrete interface for unified modeling. Refer to Appendix A.3.5 for more details.

4 UniAudio 2.0: Unified Autoregressive Modeling on FactorCodec

Connection to the UniAudio series. UniAudio Yang et al. [2023c] focused on unified audio generation; UniAudio 1.5 Yang et al. [2024a] mapped audio into a frozen LLM's vocabulary for

few-shot understanding and generation. UniAudio 2.0 continues the UniAudio series toward a universal audio foundation model. UniAudio 2.0 jointly trains on text and FactorCodec tokens, with functional layer specialization and composed audio sequences supporting multi-task learning and compositional transfer.

4.1 Unified Token Interface

Our joint vocabulary comprises text, FactorCodec analysis and reconstruction tokens, and control symbols. Following Défossez et al. [2024], Tian et al. [2025a], we pack them into a multi-stream tensor $X \in \mathbb{Z}^{B\times T\times S}$, where B is the batch size, T is the packed sequence length, and S is the number of streams. In our implementation, the audio side uses $K{=}8$ codebooks and the last stream is reserved for text, giving $S = K+1$. At text positions, only the text stream is active and audio streams are padded; at audio positions, the text stream is padded and audio streams are active according to the FactorCodec token layout. For generated audio, the analysis and reconstruction segments are produced at 5 Hz and 12.5 Hz, respectively, yielding 17.5 global audio-frame positions per second when serialized. Each audio-frame position contains up to eight codebook indices, which the local decoder predicts autoregressively within that frame. These within-frame predictions add local computation but do not each create a separate position in the global backbone sequence.

We assign an embedding table to each stream and fuse them by masked summation. Let $x_{t,i}$ denote the token at time step t and stream i, and let $m_{t,i} \in \{0,1\}$ indicate whether $x_{t,i}$ is valid (non-`PAD`). The fused representation is $h_t = \sum_{i=1}^{S} m_{t,i} E_i(x_{t,i})$, where $E_i(\cdot)$ is the embedding lookup for stream i. This design allows a single transformer to process text-only, audio-only, and mixed text-audio sequences under one autoregressive interface.

4.2 Functional Layer Specialization

Rather than letting all transformer layers process all token types uniformly, UniAudio 2.0 adopts *functional layer specialization*. We partition the backbone into three consecutive blocks:

$$H^{(u)} = F_u(h), \qquad H^{(c)} = F_{cm}(H^{(u)}), \qquad H^{(g)} = F_g(H^{(c)}), \tag{5}$$

where F_u denotes the audio-understanding experts, F_{cm} the cross-modal experts, and F_g the audio-generation experts. The cross-modal block is initialized from a pre-trained text LLM, while the audio-specific experts are newly introduced to support understanding and generation oriented specialization. A key design choice is that the understanding and generation experts operate only on audio positions, while text positions remain unchanged. Let $M_{\text{aud}} \in \{0,1\}^{B\times T}$ be a binary mask indicating whether a position corresponds to audio. For a transformer block output $f(\cdot)$, we apply audio-only updates as

$$H' = H + M_{\text{aud}} \odot \big(f(H) - H\big), \tag{6}$$

which updates hidden states only at audio positions and preserves the pre-trained text pathway at text positions. This makes the role separation introduced by FactorCodec compatible with the internal specialization of the backbone: lower and middle layers emphasize analysis and cross-modal alignment, while upper layers focus on reconstruction-token modeling.

The function-specific blocks execute serially; "experts" here does not denote a mixture-of-experts layer. There is no router or top-k expert selection. The architecture uses 3 audio-understanding layers, a 28-layer cross-modal LLaMA-3.2-3B backbone, and 2 audio-generation layers, followed by a separate 4-layer local audio decoder (Table 22).

4.3 Autoregressive Objective and Training

UniAudio 2.0 is trained under a unified autoregressive framework over the packed multi-stream sequence. Text tokens are predicted with standard language modeling loss $\mathcal{L}_{\text{text}} = -\sum_{t\in\mathcal{T}} \log p_\theta(x_{t,\text{text}} \mid X_{<t})$ where $\mathcal{T}$ denotes the set of text positions. Audio tokens are modeled at the frame level using a lightweight local decoder conditioned on the hidden states from the audio-generation experts. For each audio frame t, the decoder autoregressively predicts the K audio tokens $\mathcal{L}_{\text{audio}}$. The overall training objective is

$$\mathcal{L}_{\text{AR}} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{audio}}\mathcal{L}_{\text{audio}}. \tag{7}$$

Table 1: Unified codec reconstruction results on speech, sound, and music. PESQ is reported as WB/NB. AB denotes AudioBox score (CE/CU/PC/PQ). For RVQ baselines with adjustable depth, we use configurations near 100 tokens/s; token-rate matching does not imply bitrate matching. The evaluated FactorCodec reconstruction stream is 1000 bps; nominal baseline rates are summarized in Appendix A.4. We do not report the MimiCodec's performance on the sound and music data due to it only trained on speech data.

Model	**Speech**					**Sound**		**Music**	
	PESQ	*STOI*	*UTMOS*	*VISQOL*	*SIM*	*VISQOL*	*AB*	*VISQOL*	*AB*
DAC-Codec	2.10/2.29	0.81	3.13	3.67	0.91	3.02	3.34/4.25/3.78/5.44	**4.06**	6.98/6.97/6.25/7.09
EnCodec	2.00/2.24	0.81	2.58	3.64	0.92	2.99	3.61/4.62/3.82/5.49	4.04	6.60/6.57/**6.27**/6.71
MimiCodec	2.09/2.82	0.82	3.65	3.82	0.96	–	–	–	–
Higg-Audio	2.20/2.90	0.78	3.90	3.84	0.96	**3.20**	4.01/4.97/3.65/5.87	4.01	6.95/7.48/5.01/7.65
X-Codec	2.08/2.72	0.83	3.75	3.90	0.94	3.01	3.97/4.82/**3.98**/5.87	3.82	7.43/7.19/6.21/7.24
ALMTokenizer	2.00/2.30	0.81	3.76	3.78	0.92	2.99	4.02/4.65/3.24/5.66	3.96	6.44/6.68/6.12/6.94
Ours	**2.36/2.93**	**0.85**	**3.91**	**3.94**	**0.97**	3.10	**4.12/5.06**/3.58/**5.96**	4.03	**7.51/7.68**/6.12/**7.87**

We optimize this objective with a four-stage training recipe. The first two stages separately warm up the understanding and generation pathways, reducing early interference between analysis-oriented and synthesis-oriented learning. In particular, the understanding warm-up stage includes an auxiliary semantic distillation objective, which provides additional continuous supervision to complement the discrete quantization and encourage richer semantic representations. The later stages perform joint multi-task training over text and audio, and further introduce *composed audio sequences*: long-context sequences of related audio/text segments, to support scalable pre-training across speech, sound, and music. Stage-wise training details are deferred to Appendix B.4.

4.4 Composed Audio Sequences

To improve compositional generalization Lake and Baroni [2018], Bai et al. [2023], we introduce *composed audio sequences*, a compositional curriculum constructed from multiple related segments rather than independent task-specific pairs. By arranging related segments in a single sequence, the next-token objective implicitly supervises compositional skills (decomposition, alignment, attribute disentanglement) that recombine at inference, supporting *compositional generalization* to unseen task instantiations. A composed sequence is $C = [u_1, \ldots, u_N]$, $u_i \in \mathcal{A} \cup \mathcal{T} \cup (\mathcal{A} \times \mathcal{T})$, where each segment u_i can be an audio segment, a text segment, or a paired audio-text segment. We instantiate five types, each targeting a distinct skill: long-form segmentation (long-context dependency); same-speaker speech-text interleaving (speaker tracking); audio-caption interleaving (caption-audio alignment); mixture-clean triples (source decomposition); and content-consistent style-varied speech (content-style disentanglement). Empirically, this stage is the main driver of compositional generalization. Construction details and per-type analyses are deferred to Appendix B.3.

5 Experiments

5.1 Does FactorCodec provide a better discrete interface?

A suitable interface should satisfy three properties: it should (i) preserve sufficient information for high-quality reconstruction, (ii) be easy to model autoregressively for generation, and (iii) retain language-relevant information for understanding.

Table 2: PESQ from decoding each FactorCodec stream with the reported decoder.

Decoding tokens	PESQ-WB	PESQ-NB
Analysis only	1.13	1.47
Reconstruction only	2.39	2.92
FactorCodec (full)	2.36	2.93

Reconstruction quality. We first compare FactorCodec against prior universal audio tokenizers on reconstruction over speech, sound, and music. We include strong discrete baselines such as DAC Kumar et al. [2023], EnCodec Défossez et al. [2022], Higgs-Audio Boson AI [2025], X-Codec Ye et al. [2025a], and ALMTokenizer Yang et al. [2025]. As reported in

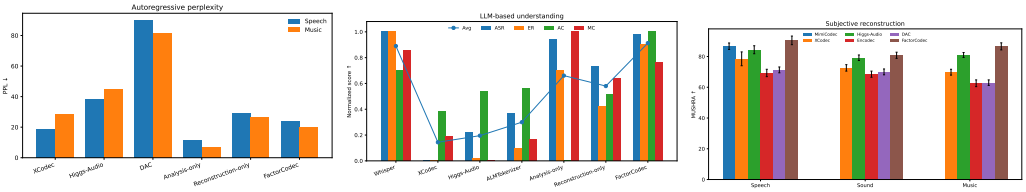

Figure 3: Tokenizer-level evaluation of FactorCodec. Left: autoregressive perplexity on speech and music. Middle: normalized LLM-based understanding performance comparing continuous and discrete token interfaces. For visualization, each task is normalized to $[0, 1]$, and ASR-WER is reversed so that higher is better for all metrics. Right: subjective reconstruction quality (MUSHRA) on speech, sound, and music. Full numerical tables are deferred to Appendix Table 15 and 16

Table 3: Performance comparison on seen tasks between previous SOTA models and UniAudio 2.0. Bold indicates the best result. For TTS and ASR tasks, Seed-ZH, Seed-EN, LS-clean denote the SeedTTS-eval and LibriSpeech benchmarks. We use WER for English and CER for Chinese. CP denotes the CLAP-Score. More comprehensive comparisons are provided in Appendix B.6.

Task	Model	Performance
TTS (Seed-ZH / Seed-EN / LS-clean)	MiMo-Audio-7B-Instruct	1.93 / 5.37 / 4.74
	Qwen2.5-Omni-7B	**1.21** / **3.10** / 4.28
	Ours	2.30 / 3.63 / **3.46**
InstructTTS (WER / Style-Acc / UTMOSv2)	MiMo-Audio-7B-Instruct	7.8 / 40.5 / 3.17
	CapSpeech-AR	9.1 / **52.2** / 3.18
	Ours	**7.3** / 42.3 / **3.38**
ASR (LS-clean / Seed-EN / Seed-ZH)	MiMo-Audio-7B-Instruct	3.5 / 29.8 / 7.0
	Qwen2.5-Omni-7B	3.9 / **1.3** / 2.9
	Ours	**2.7** / 2.6 / **2.1**
Audio Cap & Gen (CIDEr / KL / FD/CP)	Qwen2.5-Omni	0.39 / - / - / -
	Stable Audio Open	- / **2.14** / 78.2 / **0.29**
	Ours	**0.69** / 3.26 / **50.7** / 0.17
Music Cap & Gen (gpt-score / KL / FAD/CP)	Qwen2.5-Omni-7B	**5.33** / - / - / -
	MusicGen	- / **1.31** / 5.0 / 0.28
	Ours	5.14 / 1.8 / **3.44** / **0.29**
Song Generation (WER / CE / CU / PQ)	SongGen	40.58 / 6.77 / 6.86 / 7.19
	UniAudio 2.0	**36.5** / **6.87** / **7.41** / **7.62**
Lyric Recognition (WER)	Qwen2.5-Omni-7B	56.99
	Ours	**28.57**

Table 1, FactorCodec achieves strong reconstruction quality across modalities in terms of objective metrics. In Figure 3 right part, we can see that FactorCodec has better subjective performance than other baselines at similar token rates. Token-rate similarity does not imply equal bitrate; the nominal BPS and configurations for the evaluated tokenizer settings are listed in Appendix Table 13. To examine role separation directly, Table 2 shows that reconstruction tokens alone achieve PESQ comparable to the full codec, while analysis tokens alone score lower under the same decoder. These results indicate that role separation does not sacrifice the waveform side of the interface.

Autoregressive modeling suitability. We next ask whether the proposed factorization makes the discrete representation easier for language models to generate. To isolate the effect of tokenization, we keep the LLM backbone, training data, and evaluation setup fixed, and vary only the tokenizer. We report perplexity and token prediction accuracy, following [Wang et al., 2025b]. As Figure 3 left part shows, the results show two consistent patterns: semantically enhanced tokenizers (e.g. XCodec, Higg-Audio) outperform purely acoustic codecs (e.g. DAC-Codec), and FactorCodec further improves autoregressive modeling quality. In particular, analysis tokens are easier to model than reconstruction tokens, and conditioning reconstruction-token modeling on the analysis stream significantly improves reconstruction-token prediction accuracy. This supports our claim that role separation reduces the uncertainty that would otherwise be absorbed by a single-stream representation.

Understanding quality. Finally, we evaluate whether the analysis stream preserves understanding-relevant information. We compare FactorCodec against discrete audio tokenizers as well as a continuous Whisper encoder on multiple audio understanding benchmarks. As reported in Figure 3 middle part, FactorCodec achieves the strongest performance among discrete tokenizers and approaches the performance of continuous Whisper features on several tasks. Figure 4 illustrates the training loss for different tokenizers. This result is central to our goal: FactorCodec is not only LM-friendly for generation, but also competitive as a language-facing interface for understanding. In Appendix A.3.5, we provide a theoretical analysis of why introducing FactorCodec is effective.

5.2 Seen-task evaluation of UniAudio 2.0

We evaluate whether UniAudio 2.0 achieves competitive *in-domain multi-task performance* across speech, sound, and music. We compare UniAudio 2.0 with strong unified audio baselines such as MiMo-Audio Zhang et al. [2025] and Qwen2.5-Omni Xu et al. [2025a], as well as task-specific systems where appropriate. Table 3 summarizes representative results across understanding and generation tasks in the three domains, with full comparisons provided in Appendix B.6.

UniAudio 2.0 is competitive with strong 7B unified baselines despite using only 3B parameters, indicating a favorable trade-off between model scale and multi-task capability. It maintains balanced performance across both understanding and generation under a single purely discrete interface, and remains competitive beyond speech on general-audio and music tasks. These results demonstrate that UniAudio 2.0 is a strong in-domain unified model with a compact architecture. However, we also note that UniAudio 2.0 does not perform well on all benchmarks; we summarize the limitations in Appendix E.

5.3 Compositional generalization to unseen tasks

We further evaluate whether UniAudio 2.0 exhibits *compositional generalization* [Lake and Baroni, 2018] to unseen tasks: novel task instantiations that recombine compositional skills induced during pre-training (Section 4.4; boundary discussed in Appendix B.3). Few-shot evaluation focuses on *in-context adaptation*: whether the model can infer a new task pattern from a small number of demonstrations. Zero-shot instead focuses on *instruction-level and compositional transfer*: whether the model can perform unseen tasks without any in-context examples. Since these two settings probe different capabilities, we evaluate them on different task sets that are most informative for each setting.

Table 4: Few-shot results. For speech enhancement (SE) task, we use PESQ/STOI/WER/DNS-MOS as the metrics. For voice conversion (VC) task, we use WER/Speaker Similarity (SIM)/DNS-MOS as the metrics. For emotion recognition and sound classification tasks, we use accuracy as the metric. For emotion recognition, we report English and Chinese results in ESD dataset Zhou et al. [2022].

Task	Model	1-shot	2-shot
SE	MiMo-Audio	1.20/0.21/65.29/3.46	1.16/0.21/29.29/3.48
	Ours	**1.55/0.64/14.13/3.76**	**1.52/0.66/14.82/3.77**
VC	MiMo-Audio	20.95/**0.90/3.80**	**14.05/0.93/3.78**
	Ours	**18.61**/0.89/3.74	19.01/0.90/3.71
Emotion	MiMo-Audio	42.5/45.0	53.6/52.4
	Ours	**67.0/59.8**	**70.2/62.8**
Sound	MiMo-Audio	45.3	**76.0**
	Ours	**59.8**	62.8

Few-shot setting.

For few-shot generation, we follow Zhang et al. [2025] and evaluate speech denoising and voice conversion; for understanding, we consider sound and emotion classification. These tasks require recombining multiple compositional skills induced by composed sequences, e.g., denoising relies on the decomposition skill from mixture-clean triples, while voice conversion combines speaker tracking and content-style disentanglement. Crucially, the evaluation data (WHAM Wichern et al. [2019] noise types, VCTK Veaux et al. [2017] speakers) are disjoint from pre-training. Table 4 shows that UniAudio 2.0 matches or surpasses MiMo-Audio-7B on most benchmarks.

Table 6: Controlled tokenizer comparison under the same LM evaluation protocol. Entries are speech/music PPL, averaged over the first four RVQ layers. Both tokenizers use eight RVQ layers and codebook size 1024, but their native token rates differ (25 Hz for the CosyVoice3-style baseline and 5 Hz for FactorCodec); this is not a bitrate-matched comparison. Lower is better.

Tokenizer	Standalone PPL ↓	With reconstruction PPL ↓
CosyVoice3-style (25 Hz)	29.34 / 30.23	37.45 / 39.16
FactorCodec analysis (5 Hz)	**11.57 / 6.56**	**23.77 / 20.10**

Zero-shot setting. We evaluate zero-shot performance on unseen tasks without demonstrations, probing instruction following and compositional transfer. We consider standard benchmarks (e.g., MMLU Hendrycks et al. [2020]) and newly designed compositional tasks: speech-sound generation (Speech-S, composing TTS and sound generation) and audio-prompt- and caption-guided speech generation (A-I-TTS, composing speaker tracking, and instructed TTS). Results in Table 5 show that UniAudio 2.0 retains strong text understanding without major degradation from multimodal training, and can compose operations at inference, e.g., leveraging the timbre of an audio prompt and the style from a caption to synthesize the desired speech. While non-trivial, performance is not yet saturated; we discuss the boundary between compositional generalization and fundamentally novel operations in Appendix E.

Table 5: Zero-shot generalization results. A-I-TTS denotes audio and caption guided speech generation, we use SIM / StyleAcc/WER/UTMOSv2 as the metrics; Speech-S denotes generate speech and sound, we use WER/CLAP-score/UTMOSv2 as the metrics.

Task	Model	Score
Text	LLaMA 3.2 3B	**47.63**
	Ours	44.10
A-I-TTS	Mimo-Audio	**0.95** / 31.02 / 15.0 / **2.96**
	Ours	0.89 / **32.62** / **11.57** / 2.87
Speech-S	Mimo-Audio	7.49 / -0.09 / **3.2**
	Ours	**6.15** / **0.11** / 2.96

5.4 Ablation studies

We conduct ablations to analyze the contributions of the key components in UniAudio 2.0.

FactorCodec: role separation. Comparing analysis-only, reconstruction-only, and the full factorized representation (Figure 3, Left and Middle) confirms a clear division of labor: analysis tokens specialize in understanding, reconstruction tokens specialize in synthesis, and the joint representation combines both strengths. Detailed analysis is provided in Appendices A.3.5, A.3.6, and A.3.7.

Comparison with semantic tokenizers. To test whether generic semantic supervision alone produces equally LM-friendly tokens, we compare our analysis tokens with a CosyVoice3-style tokenizer under the same language-model evaluation protocol. Both use eight RVQ layers and a codebook size of 1024. As Table 6 shows, FactorCodec's analysis tokens achieve lower standalone PPL and lower reconstruction-token PPL when paired with the reconstruction stream. This comparison controls RVQ depth and codebook size, but not token rate: the baseline operates at 25 Hz and our analysis stream at 5 Hz. The results therefore show that codebook size alone does not explain the PPL difference, without claiming a bitrate-matched comparison.

Effectiveness of GRPO in the analysis branch. In Section 3, we introduce two training stages: SFT and GRPO. GRPO is designed to improve the model's reasoning and analysis capability. In Appendix A.3.1, we present ablation experiments to evaluate the effectiveness of these stages. As shown in Table 14, GRPO further improves performance on audio understanding tasks.

Effectiveness of FiLM conditioning. As introduced in Section 3, the analysis stream provides high-level guidance to the reconstruction pathway through FiLM Perez et al. [2018], without directly participating in waveform reconstruction. We conduct an ablation study, as shown in Table 11, which demonstrates that FiLM effectively connects the analysis and reconstruction tokens. Furthermore, Table 12 shows that introducing FiLM conditioning does not degrade reconstruction performance.

Effectiveness of multi-expert encoders in the reconstruction branch. To preserve richer semantic information in the reconstruction tokens, we employ three separate pre-trained encoders to extract features, followed by three group-wise VQ modules. To evaluate this design, we construct two baselines: (1) concatenating all semantic features and applying query-based quantization; and (2)

Table 7: Ablation studies on multi-stage training, layer specialization, and model size. gpt denotes GPT-score, S denotes the speaker similarity, D-MOS and U-MOS denote DNS-MOS and UT-MOSv2. FS-VC denotes the few-shot voice conversion task. N/A denotes the model cannot get meaningful results. Due to space limitation, we report more results in Appendix Table 34.

Setting	MMLU	ASR (EN/ZH)	TTS (EN/ZH)	Audio Cap&Gen (CIDEr/KL/FD)	Music Cap&Gen (gpt/KL/FAD)	FS-VC (WER/S/D-MOS)	Speech-S (WER/clap/U-MOS)
Ours	44.1	2.6 / 2.1	3.63 / 2.30	0.69 / 3.26 / 50.7	5.14 / 1.8 / 3.44	18.61/0.89/3.74	6.15 / 0.11 / 2.96
w/o Stage 4	38.7	3.87 / 4.3	6.21 / 1.97	0.34 / 3.08 / 47.6	4.56 / 1.87 / 3.66	N/A	12.45 / 0.03 / 2.88
& w/o Experts	37.2	4.22 / 5.1	5.31 / 2.76	0.22 / 3.98 / 60.8	4.33 / 4.12 / 6.98	N/A	10.33 / 0.03 / 2.92
1B	30.2	3.3 / 3.5	4.2 / 5.3	0.30 / 3.78 / 82.4	4.52 / 4.3 / 6.9	21.3/0.82/3.66	6.79 / 0.04 / 2.82

concatenating all semantic features and applying standard RVQ. As shown in Table 21, query-based quantization achieves better reconstruction performance, while the proposed group-wise VQ yields improved token modeling performance, particularly for music.

Functional layer specialization. We first conduct ablation studies on 1B model size, Table 26 shows that two experts and semantic reconstruction loss have an impact on final performance. In Table 7, we also conduct 3B version experiments to explore the influence of layer specialization (without the understanding and generation experts). These results indicate that functional specialization helps different parts of the model focus on distinct roles, improving overall multi-task capability.

Training stage. We analyze the role of the final training stage with composed audio sequences. As shown in Table 7, removing Stage 4 has task-dependent effects rather than uniformly limited impact on seen tasks. It reduces audio-caption CIDEr from 0.69 to 0.34 and worsens English ASR and TTS WER from 2.60 to 3.87 and from 3.63 to 6.21, respectively, while Chinese TTS WER changes from 2.30 to 1.97. In the Speech-S zero-shot evaluation, WER increases from 6.15 to 12.45 (Appendix Table 34). These results suggest that Stage 4 benefits selected in-domain tasks and compositional transfer, with effects that vary by task.

Influence of model size. We evaluate the effect of model scale by using LLaMA 3.2 1B and 3B as the cross-modal expert under the same training setup. As shown in Table 7, the 1B model shows consistent performance degradation, particularly in few-shot and zero-shot settings, indicating that model capacity is crucial for multi-task generalization.

Benefits of unified modeling. In a controlled speech setup (Table 27), unified training improves ASR over task-specific models, and text-only data provides an additional gain. Unified training slightly degrades TTS at 1B, but the gap nearly disappears at 3B, suggesting that task interference diminishes with model scale.

6 Conclusion

In this work, we present UniAudio 2.0, a unified audio foundation model that supports both understanding and generation under a purely discrete interface. We introduce FactorCodec, which factorizes audio into role-separated analysis and reconstruction tokens, enabling a unified representation that is both text-aligned and suitable for autoregressive generation. Our results suggest that role-separated discrete tokenization, together with structured sequence construction, provides a promising direction for building scalable and generalizable audio foundation models. Future work will explore further scaling of model capacity and training data, as well as more advanced mechanisms for improving long-context reasoning and generalization. The limitations of our approach are discussed in Appendix E.

Impact Statement

UniAudio 2.0 may support audio accessibility, speech interfaces, and creative production. Its generation capabilities could also enable voice imitation, misleading audio, or copyright and consent violations. We do not evaluate technical safeguards. Deployments should obtain consent for voice imitation, respect applicable rights and dataset terms, and disclose synthetic audio.

Acknowledgments and Disclosure of Funding

This study was supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675*, 2016. URL https://arxiv.org/abs/1609.08675.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating Music From Text. *arXiv preprint arXiv:2301.11325*, 2023. URL https://arxiv.org/abs/2301.11325.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR, 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential Modeling Enables Scalable Learning for Large Vision Models. *arXiv preprint arXiv:2312.00785*, 2023. URL https://arxiv.org/abs/2312.00785.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011. URL https://millionsongdataset.com/.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, 2019. URL https://mtg.github.io/mtg-jamendo-dataset/.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023a. doi: 10.1109/taslp.2023.3288409. URL https://arxiv.org/abs/2209.03143.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. SoundStorm: Efficient Parallel Audio Generation. *arXiv preprint arXiv:2305.09636*, 2023b. URL https://arxiv.org/abs/2305.09636.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, 2020. URL https://arxiv.org/abs/2005.14165.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6. URL https://sail.usc.edu/publications/html/b2hd-Busso2008IEMOCAPInteractiveemotionaldyadic.html.

Runyuan Cai, Yu Lin, Yiming Wang, Chunlin Fu, and Xiaodong Zeng. Unifying Speech Recognition, Synthesis and Conversion with Autoregressive Transformers. *arXiv preprint arXiv:2601.10770*, 2026. URL https://arxiv.org/abs/2601.10770.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/jstsp.2022.3188113. URL https://arxiv.org/abs/2110.13900.

Xueyuan Chen, Dongchao Yang, Dingdong Wang, Xixin Wu, Zhiyong Wu, and Helen Meng. CoLM-DSR: Leveraging Neural Codec Language Modeling for Multi-Modal Dysarthric Speech Reconstruction. *arXiv preprint arXiv:2406.08336*, 2024. URL https://arxiv.org/abs/2406.08336.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284*, 2023. URL https://arxiv.org/abs/2306.05284.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A Dataset For Music Analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. URL https://arxiv.org/abs/1612.01840.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High Fidelity Neural Audio Compression. *arXiv preprint arXiv:2210.13438*, 2022. URL https://arxiv.org/abs/2210.13438.

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

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. LP-MusicCaps: LLM-Based Pseudo Music Captioning. *arXiv preprint arXiv:2307.16372*, 2023. URL https://arxiv.org/abs/2307.16372.

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. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. *arXiv preprint arXiv:2407.05407*, 2024. URL https://arxiv.org/abs/2407.05407.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training. *arXiv preprint arXiv:2505.17589*, 2025. URL https://arxiv.org/abs/2505.17589.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio Open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. doi: 10.1109/icassp49660.2025.10888461. URL https://arxiv.org/abs/2407.14358.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. *arXiv preprint arXiv:2409.06666*, 2024. URL https://arxiv.org/abs/2409.06666.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. URL https://research.google/pubs/audio-set-an-ontology-and-human-labeled-dataset-for-audio-events/.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Music Flamingo: Scaling Music Understanding in Audio Language Models. *arXiv preprint arXiv:2511.10289*, 2025a. URL https://arxiv.org/abs/2511.10289.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. *arXiv preprint arXiv:2503.03983*, 2025b. URL https://arxiv.org/abs/2503.03983.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *arXiv preprint arXiv:2507.08128*, 2025. URL https://arxiv.org/abs/2507.08128.

Yitian Gong, Luozhijie Jin, Ruifan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. XY-Tokenizer: Mitigating the Semantic-Acoustic Conflict in Low-Bitrate Speech Codecs. *arXiv preprint arXiv:2506.23325*, 2025. URL https://arxiv.org/abs/2506.23325.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: A Large-Scale, Extensive, Multilingual, and Diverse Dataset for Speech Generation. *arXiv preprint arXiv:2501.15907*, 2025. URL https://arxiv.org/abs/2501.15907.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020. URL https://arxiv.org/abs/2009.03300.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. URL https://arxiv.org/abs/2106.07447.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction. *arXiv preprint arXiv:2502.11946*, 2025. URL https://arxiv.org/abs/2502.11946.

Rongjie Huang, Ruofan Hu, Yongqi Wang, Zehan Wang, Xize Cheng, Ziyue Jiang, Zhenhui Ye, Dongchao Yang, Luping Liu, Peng Gao, and Zhou Zhao. InstructSpeech: Following Speech Editing Instructions via Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19886–19903. PMLR, 2024. URL https://proceedings.mlr.press/v235/huang24k.html.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. *arXiv preprint arXiv:2408.16532*, 2024. URL https://arxiv.org/abs/2408.16532.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9052942. URL https://arxiv.org/abs/1912.07875.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Interspeech*, pages 1741–1744, 2008. doi: 10.21437/Interspeech.2008-480. URL https://www.isca-archive.org/interspeech_2008/kim08c_interspeech.html.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, et al. Kimi-Audio Technical Report. *arXiv preprint arXiv:2504.18425*, 2025. URL https://arxiv.org/abs/2504.18425.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. *arXiv preprint arXiv:2402.01831*, 2024. URL https://arxiv.org/abs/2402.01831.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation. *arXiv preprint arXiv:2209.15352*, 2022. URL https://arxiv.org/abs/2209.15352.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-Fidelity Audio Compression with Improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36, pages 27980–27993, 2023. doi: 10.52202/075280-1214. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Brenden Lake and Marco Baroni. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Max W. Y. Lam, Yijin Xing, Weiya You, Jingcheng Wu, Zongyu Yin, Fuqiang Jiang, Hangyu Liu, Feng Liu, Xingda Li, Wei-Tsung Lu, Hanyu Chen, Tong Feng, Tianwei Zhao, Chien-Hung Liu, Xuchen Song, Yang Li, and Yahui Zhou. Analyzable Chain-of-Musical-Thought Prompting for High-Fidelity Music Generation. *arXiv preprint arXiv:2503.19611*, 2025. URL https://arxiv.org/abs/2503.19611.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless Speech-to-Speech Translation on Real Data. *arXiv preprint arXiv:2112.08352*, 2021. URL https://arxiv.org/abs/2112.08352.

Keon Lee, Kyumin Park, and Daeyoung Kim. DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10095751. URL https://arxiv.org/abs/2207.01063.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering. *arXiv preprint arXiv:2503.11197*, 2025a. URL https://arxiv.org/abs/2503.11197.

Longhao Li, Zhao Guo, Hongjie Chen, Yuhang Dai, Ziyu Zhang, Hongfei Xue, Tianlun Zuo, Chengyou Wang, Shuiyuan Wang, Jie Li, Jian Kang, Xin Xu, Hui Bu, Binbin Zhang, Ruibin Yuan, Ziya Zhou, Wei Xue, and Lei Xie. WenetSpeech-Yue: A Large-scale Cantonese Speech Corpus with Multi-dimensional Annotation. *arXiv preprint arXiv:2509.03959*, 2025b. URL https://arxiv.org/abs/2509.03959.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023. URL https://www.wavlab.org/activities/2023/foundations/.

Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Zhou Zhao, and Wei Xue. FlashAudio: Rectified Flow for Fast and High-Fidelity Text-to-Audio Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13694–13710, 2025a. doi: 10.18653/v1/2025.acl-long.673. URL https://aclanthology.org/2025.acl-long.673/.

Zihan Liu, Shuangrui Ding, Zhixiong Zhang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation. *arXiv preprint arXiv:2502.13128*, 2025b. URL https://arxiv.org/abs/2502.13128.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *arXiv preprint arXiv:2303.17395*, 2023. URL https://arxiv.org/abs/2303.17395.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. DCASE2017 Challenge Setup: Tasks, Datasets and Baseline System. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pages 85–92, 2017. URL https://dcase.community/workshop2017/proceedings.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. EXPRESSO: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. *arXiv preprint arXiv:2308.05725*, 2023. URL https://arxiv.org/abs/2308.05725.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussà, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved Spoken and Written Language Model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025. doi: 10.1162/tacl_a_00728. URL https://aclanthology.org/2025.tacl-1.2/.

OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015. doi: 10.1109/icassp.2015.7178964. URL https://www.openslr.org/12.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, 2015. doi: 10.1145/2733373.2806390. URL https://raw.githubusercontent.com/karolpiczak/ESC-50/master/README.md.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A Large-Scale Multilingual Dataset for Speech Research. *arXiv preprint arXiv:2012.03411*, 2020. URL https://arxiv.org/abs/2012.03411.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid Transformers for Music Source Separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/icassp49357.2023.10096956. URL https://arxiv.org/abs/2211.08553.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio LLM? *arXiv preprint arXiv:2505.09439*, 2025. URL https://arxiv.org/abs/2505.09439.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. AudioPaLM: A Large Language Model That Can Speak and Listen. *arXiv preprint arXiv:2306.12925*, 2023. URL https://arxiv.org/abs/2306.12925.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. SNAC: Multi-Scale Neural Audio Codec. *arXiv preprint arXiv:2410.14411*, 2024. URL https://arxiv.org/abs/2410.14411.

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

Team OLMo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. URL https://arxiv.org/abs/2512.13961.

Jinchuan Tian, William Chen, Yifan Peng, Jiatong Shi, Siddhant Arora, Shikhar Bharadwaj, Takashi Maekaku, Yusuke Shinohara, Keita Goto, Xiang Yue, Huck Yang, and Shinji Watanabe. OpusLM: A Family of Open Unified Speech Language Models. *arXiv preprint arXiv:2506.17611*, 2025a. URL https://arxiv.org/abs/2506.17611.

Jinchuan Tian, Sang-gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, Mohammad Shoeybi, Bryan Catanzaro, Rafael Valle, and Wei Ping. UALM: Unified Audio Language Model for Understanding, Generation and Reasoning. *arXiv preprint arXiv:2510.12000*, 2025b. URL https://arxiv.org/abs/2510.12000.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

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. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound. *arXiv preprint arXiv:2502.05139*, 2025. URL https://arxiv.org/abs/2502.05139.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. University of Edinburgh, Centre for Speech Technology Research, 2017. URL https://hdl.handle.net/10283/2651.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. URL https://arxiv.org/abs/2301.02111.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Thomas Thebaud, Laureano Moro Velazquez, Jesus Villalba, Zengyi Qin, Shrikanth Narayanan, Mounya Elhiali, and Najim Dehak. CapSpeech: Enabling Downstream Applications in Style-Captioned Text-to-Speech. *arXiv preprint arXiv:2506.02863*, 2025a. URL https://arxiv.org/abs/2506.02863.

Lu Wang, Hao Chen, Siyu Wu, Zhiyue Wu, Hao Zhou, Chengfeng Zhang, Ting Wang, and Haodi Zhang. AudioCodecBench: A Comprehensive Benchmark for Audio Codec Evaluation. *arXiv preprint arXiv:2509.02349*, 2025b. URL https://arxiv.org/abs/2509.02349.

Yuanyuan Wang, Dongchao Yang, Yiwen Shao, Hangting Chen, Jiankun Zhao, Zhiyong Wu, Helen Meng, and Xixin Wu. DualSpeechLM: Towards Unified Speech Understanding and Generation via Dual Speech Token Modeling with Large Language Models. *arXiv preprint arXiv:2508.08961*, 2025c. URL https://arxiv.org/abs/2508.08961.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending Speech Separation to Noisy Environments. *arXiv preprint arXiv:1907.01160*, 2019. URL https://arxiv.org/abs/1907.01160.

Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5840, 2019. doi: 10.1109/ICASSP.2019.8683091. URL https://doi.org/10.1109/ICASSP.2019.8683091.

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. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025b. URL https://arxiv.org/abs/2509.17765.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, et al. Ming-UniAudio: Speech LLM for Joint Understanding, Generation and Editing with Unified Representation. *arXiv preprint arXiv:2511.05516*, 2025. URL https://arxiv.org/abs/2511.05516.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. HiFi-Codec: Group-residual Vector quantization for High Fidelity Audio Codec. *arXiv preprint arXiv:2305.02765*, 2023a. URL https://arxiv.org/abs/2305.02765.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. InstructTTS: Modelling Expressive TTS in Discrete Latent Space with Natural Language Style Prompt. *arXiv preprint arXiv:2301.13662*, 2023b. URL https://arxiv.org/abs/2301.13662.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen Meng. UniAudio: An Audio Foundation Model Toward Universal Audio Generation. *arXiv preprint arXiv:2310.00704*, 2023c. URL https://arxiv.org/abs/2310.00704.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. *arXiv preprint arXiv:2406.10056*, 2024a.

Dongchao Yang, Dingdong Wang, Haohan Guo, Xueyuan Chen, Xixin Wu, and Helen Meng. SimpleSpeech: Towards Simple and Efficient Text-to-Speech with Scalar Latent Transformer Diffusion Models. *arXiv preprint arXiv:2406.02328*, 2024b. URL https://arxiv.org/abs/2406.02328.

Dongchao Yang, Songxiang Liu, Haohan Guo, Jiankun Zhao, Yuanyuan Wang, Helin Wang, Zeqian Ju, Xubo Liu, Xueyuan Chen, Xu Tan, Xixin Wu, and Helen Meng. ALMTokenizer: A Low-bitrate and Semantic-rich Audio Codec Tokenizer for Audio Language Modeling. *arXiv preprint arXiv:2504.10344*, 2025. URL https://arxiv.org/abs/2504.10344.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec Does Matter: Exploring the Semantic Shortcoming of Codec for Audio Language Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25697–25705, 2025a. doi: 10.1609/aaai.v39i24.34761. URL https://ojs.aaai.org/index.php/AAAI/article/view/34761.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis. *arXiv preprint arXiv:2502.04128*, 2025b. URL https://arxiv.org/abs/2502.04128.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/taslp.2021.3129994. URL https://research.google/pubs/soundstream-an-end-to-end-neural-audio-codec/.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot. *arXiv preprint arXiv:2412.02612*, 2024. URL https://arxiv.org/abs/2412.02612.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022. doi: 10.1109/icassp43922.2022.9746682. URL https://raw.githubusercontent.com/wenet-e2e/WenetSpeech/main/README.md.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. *arXiv preprint arXiv:2305.11000*, 2023a. URL https://arxiv.org/abs/2305.11000.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. MiMo-Audio: Audio Language Models are Few-Shot Learners. *arXiv preprint arXiv:2512.23808*, 2025. URL https://arxiv.org/abs/2512.23808.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified Speech Tokenizer for Speech Large Language Models. *arXiv preprint arXiv:2308.16692*, 2023b. URL https://arxiv.org/abs/2308.16692.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, 137:1–18, 2022. ISSN 0167-6393. doi: 10.1016/j.specom.2021.11.006. URL https://arxiv.org/abs/2105.14762.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-Supervised Music Representation Learning with Mel Residual Vector Quantization. *arXiv preprint arXiv:2501.01108*, 2025. URL https://arxiv.org/abs/2501.01108.

	FactorCodec
Input shape	(B, 1, N)
Analysis Branch	
Encoder modules	Whisper & Music Encoder
Token Interleaving and Retrieval	w=5
Dimension of transformer encoder	768
The number of transformer encoder	4
Codebook size	1024
VQ layers	8
Reconstruction Branch	
Encoder modules	Whisper & WavLM & Music Encoder
Codebook size	1024
VQ layers	8
Semantic decoder layers	4 convolutional layers
Flow-based Scalar Diffusion Decoder	
The number of transformer decoder	24
Dimension of transformer decoder	768
Latent space dimension	136

Table 8: FactorCodec model backbone configurations

A FactorCodec

In this section, we provide additional details about the proposed FactorCodec. Figure 2 provides an overview of the FactorCodec framework.

A.1 Model Structure of FactorCodec

Table 8 gives the details of FactorCodec configuration.

A.2 The training details of FactorCodec

The training of FactorCodec includes three stages: (1) supervised fine-tuning (SFT) on large-scale audio understanding tasks for the analysis branch; (2) GRPO-style reinforcement learning to encourage detailed and grounded analytical descriptions and improve the perceptual reasoning capability of the analysis branch; and (3) freezing the analysis branch and training the reconstruction branch with a flow-based decoder. Table 10 presents the training data and configurations for each stage. We build the mixture of audio reasoning data [2] based on Qwen3-Omni Xu et al. [2025b] and Gemini 2.5 Pro. Boxes A.4–A.4 show examples. Below, we describe each stage in detail.

A.2.1 Stage 1: SFT for Analysis Branch

In this stage, we follow SALMONN Tang et al. [2024] and use multiple audio understanding tasks as training objectives. During training, we only update the VQ modules and LoRA parameters to preserve pretrained linguistic knowledge. For different tasks, we prepare multiple prompts and randomly sample one prompt per instance.

A.2.2 Stage 2: GRPO for Analysis Branch

In this stage, we follow common GRPO practice Shao et al. [2024], Li et al. [2025a] and design an accuracy-based reward to train the model. Specifically, we classify audio understanding tasks into two categories: (1) rule-based verifiable tasks, such as ASR and audio classification, where we use WER and label accuracy as rewards; and (2) tasks that are difficult to verify automatically, such

[2] We use the term reasoning to denote fine-grained, perceptual analyses and inference cues grounded in audio (e.g., events, rhythm, timbre, and scene context), rather than symbolic multi-step logical deduction.

as fine-grained audio analysis and interpretation. For the latter, we introduce an LLM-based judge to score each rollout. We use LLaMA 3.1-Instruct 8B as the judge and ask it to evaluate the gap between the rollout and the ground-truth answer. Table 9 presents the hyper-parameters used for GRPO training. Given an input query q, the model first samples G distinct outputs $\{o_1, o_2, \ldots, o_G\}$. Each output is evaluated by our reward model R, which assigns a scalar reward r_i to each sample. Based on these rewards, the relative advantage for each output can be computed as shown in Equation 8. The full GRPO objective is summarized in the Equation 9.

$$\hat{A}_{i,t} = \frac{r(q, o_i) - \text{mean}\{r(q, o_1), \ldots, r(q, o_G)\}}{\text{std}\{r(q, o_1), \ldots, r(q, o_G)\}} \tag{8}$$

$$\mathcal{L}_{\text{GRPO},i,t} = \min \Bigg[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \hat{A}_{i,t}, \\ \text{clip}\Big(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, 1-\epsilon, 1+\epsilon \Big) \hat{A}_{i,t} \Bigg] \tag{9}$$

where ϵ denotes the PPO clipping range.

A.2.3 Stage 3: Waveform Reconstruction with a Flow-based Diffusion Decoder

In this stage, we freeze the analysis branch and train the reconstruction branch to recover waveforms from discrete tokens. Specifically, we first employ a multi-expert semantic encoder to extract semantically rich features from speech, music, and sound signals. These features are then quantized into discrete tokens using three parallel VQ groups, with a 1:1:6 allocation of VQ layers across the three groups. We assign six VQ layers to the last group to encode finer-grained acoustic details for high-fidelity reconstruction.

Overall, the three groups produce eight discrete tokens per frame, which we refer to as *reconstruction tokens*. Following SimpleSpeech Yang et al. [2024b], we further construct a flow-based scalar diffusion decoder to reconstruct waveforms from these tokens.

We do not directly condition waveform reconstruction on analysis tokens for two main reasons. First, reconstruction tokens already capture the essential information provided by analysis tokens. Second, explicit conditioning would significantly increase the sequence length of the diffusion transformer. In preliminary experiments, upsampling analysis tokens and concatenating them with reconstruction tokens resulted in only marginal improvements, as Table 2 shows. Instead, we adopt FiLM Perez et al. [2018] to inject analysis-token information into the multi-expert semantic encoder. Empirically, this design introduces no noticeable degradation in reconstruction quality.

For SQ-Codec, we follow the training protocol of SimpleSpeech Yang et al. [2024b] and set the latent dimension to 136. The model is trained on the same dataset used in Stage 3, as summarized in Table 10.

A.2.4 Implementation Details

We train FactorCodec in multiple stages. For each stage, we train the model on 8 NVIDIA A100 GPUs. The learning rate is set to $1e-4$ with a cosine annealing schedule.

A.3 Experiments

In this part, we include additional more experiments for FactorCodec, focusing on the effectiveness of GRPO, why choose multi-expert semantic features, the effectiveness of FiLM, the subjective evaluation for audio tokenizer, why the analysis tokens are important, the comparison between FactorCodec and other semantic tokenizers, and the influence of Classifier-free guidance for reconstruction performance.

A.3.1 The effectiveness of GRPO for the analysis branch training

To better understand the influence of GRPO training, we use the same audio understanding test set to evaluate the analysis branch on ASR and audio classification tasks. As Table 14 shows, GRPO

Table 9: Details of the hyper-parameters used for GRPO training.

Setting	Value
Batch Size per Device	1
Gradient Accumulation Steps	2
Learning Rate	1×10^{-6}
Temperature	1.0
Maximum Response Length	2048
Number of rollouts	8
Kullback-Leibler Coefficient	0.04

Table 10: Training stages and configurations of FactorCodec. The first two stages train the analysis branch, and the last stage trains the reconstruction branch and decoder while freezing the analysis branch.

Setting	Training data	Other configuration
Stage 1	WenetSpeech Zhang et al. [2022], MLS Pratap et al. [2020], AudioSet Gemmeke et al. [2017], IEMOCAP Busso et al. [2008], AudioCaps Kim et al. [2019], WavCaps Mei et al. [2023], LibriSpeech Panayotov et al. [2015], TUT Acoustic scenes 2017 Mesaros et al. [2017] LP-MusicCaps Doh et al. [2023], mixture of audio reasoning data	learning rate = 2e-4 LoRA rank = 64
Stage 2	LibriSpeech Panayotov et al. [2015], AudioCaps Kim et al. [2019], small-set of LP-MusicCaps Doh et al. [2023], IEMOCAP Busso et al. [2008], AudioCaps Kim et al. [2019], WavCaps Mei et al. [2023], LibriSpeech Panayotov et al. [2015], TUT Acoustic scenes 2017 Mesaros et al. [2017], mixture of audio reasoning data	learning rate = 1e-6
Stage 3	Million Song Bertin-Mahieux et al. [2011], MLS Pratap et al. [2020], AudioSet Gemmeke et al. [2017]	learning rate = 2e-4

further improves performance on audio understanding tasks and also improves the quality of detailed audio reasoning analysis. Overall, GRPO brings a significant improvement.

A.3.2 The effectiveness of multi-expert semantic features

Different audio modalities may emphasize different types of semantic information. We therefore use multiple expert encoders to extract modality-specific semantic features and apply separate VQ heads to quantize them. To demonstrate the effectiveness of this strategy, we build two baselines: (1) concatenating all semantic features and applying query-based quantization; and (2) concatenating all semantic features and applying standard RVQ. Table 21 shows the results. From the reconstruction perspective, query-based quantization performs better, while group-wise VQ yields better token modeling performance (semantic information), especially for music. We attribute this to using separate VQ layers for different semantic features. Although query-based quantization can further improve reconstruction, it requires multiple transformer encoders. Considering the additional inference cost, we do not apply query-based quantization in the reconstruction branch.

A.3.3 The effectiveness of FiLM condition

To explore whether using FiLM Perez et al. [2018] to connect both analysis branch and reconstruction branch is useful, we design an ablation study, as shown in Table 11. We can see that using FiLM to build the connection between analysis tokens and reconstruction tokens is effective. Furthermore, as Table 12 shows, introducing the FiLM module does not influence the reconstruction performance.

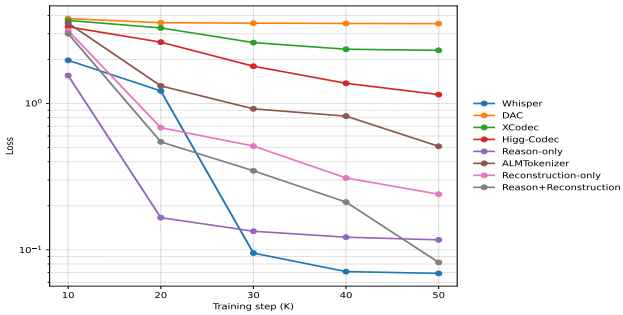

Figure 4: The training loss of different audio tokenizers for understanding tasks.

Table 11: Ablation study for the effectiveness of FiLM. Similarly, we also use the PPL across codebooks as the metric.

Model	Speech (↓)					Music (↓)				
	VQ1	VQ2	VQ3	VQ4	Avg	VQ1	VQ2	VQ3	VQ4	Avg
Analysis + Reconstruction	6.92	11.37	19.42	55.36	23.77	7.02	15.55	15.23	42.5	20.1
W/O FiLM	7.22	12.92	23.45	60.12	25.9	7.98	17.23	19.67	54.6	24.9

A.3.4 The subjective evaluation for audio tokenizer

For the subjective evaluation, we follow previous works Kreuk et al. [2022] conduct the MUSHRA test. Listeners rate quality similarity (0-100) comparing multiple versions with a labeled and hidden reference. Setup: 20 samples per category (speech, music, sound), 15 audio professionals as participants. We report both mean and CI95 scores.

A.3.5 Why Analysis Tokens Help: An Information-Theoretic Perspective

We use information-theoretic concepts to motivate role-separated tokenization and interpret the empirical results. The discussion below is not a formal claim that the learned representation is an optimal information bottleneck.

Analysis tokens as a compressed, text-oriented representation. Following the information bottleneck principle [Tishby et al., 2000], a representation R for a downstream text task Y can be motivated by the objective

$$\max_{R} \; I(R;Y) - \beta\, I(R;X), \tag{10}$$

where X is the input audio waveform and $\beta > 0$ trades off task-relevant information against compression. Standard acoustic codecs (e.g., DAC [Kumar et al., 2023], Encodec [Défossez et al., 2022]) optimize waveform reconstruction, which encourages their representations to preserve information useful for reconstruction, but does not explicitly optimize for text-grounded tasks. Our analysis tokens are instead trained with text-prediction objectives (SFT + GRPO) under a frozen text LLM head, encouraging retention of information useful for text-grounded audio tasks. The 5 Hz frame rate (vs. 25 Hz in CosyVoice 3 [Du et al., 2025]) limits the temporal rate of the analysis representation and provides a compression constraint. These choices motivate an information-bottleneck interpretation, without establishing that Eq. (10) is directly optimized.

Factorization and reconstruction-token prediction. For analysis tokens R and reconstruction tokens S, the joint entropy obeys the chain rule:

$$H(R,S) = H(R) + H(S \mid R). \tag{11}$$

The autoregressive model predicts R and then predicts S conditioned on R, corresponding to $p(R,S) = p(R)\,p(S \mid R)$. Conditioning cannot increase entropy:

$$H(S \mid R) = H(S) - I(S;R) \le H(S), \tag{12}$$

with strict inequality only when $I(S;R) > 0$. This identity alone does not guarantee that a finite model will achieve lower perplexity; that is an empirical question. In our experiments (Table 16), average reconstruction-token perplexity drops from 29.36 (Reconstruction-only) to 23.77

Table 12: The influence of FiLM module for the Reconstruction performance. PESQ includes two versions: WB and NB. AudioBox Score includes CE, CU, PC, and PQ.

Model	Speech					Sound		Music	
	PESQ	*STOI*	*UT-MOS*	*VISQOL*	*SIM*	*VISQOL*	*AudioBox Score*	*VISQOL*	*AudioBox Score*
Ours	2.36/2.93	0.85	3.91	3.94	0.97	3.10	4.12/5.06/3.58/5.96	4.03	7.51/7.68/6.12/7.87
W/O FiLM	2.39/2.92	0.85	3.90	3.92	0.97	3.14	4.21/5.12/3.49/5.92	4.01	7.48/7.62/6.14/7.78

Table 13: Speech reconstruction performance, tokenizer configurations, and nominal bitrate. BPS is computed as token rate $\times \log_2$(codebook size); values assume fixed-length codebook indices. PESQ is reported as WB/NB.

Model	Frame	RVQ	Token	Codebook size	BPS	PESQ	STOI	UTMOS	VISQOL	SIM
SQ-Codec	25	-	-	-	-	*4.28 / 4.22*	*0.96*	*4.12*	*4.81*	*0.99*
DAC-Codec Kumar et al. [2023]	50	2	100	1024	1000	2.10/2.29	0.81	3.13	3.67	0.91
EnCodec Défossez et al. [2022]	75	2	150	1024	1500	2.00/2.24	0.81	2.58	3.64	0.92
MimiCodec Défossez et al. [2024]	12.5	8	100	2048	1100	2.09/2.82	0.82	3.65	3.82	0.96
Higg-Audio Boson AI [2025]	25	4	100	1024	1000	2.20/2.90	0.78	3.90	3.84	0.96
X-Codec Ye et al. [2025a]	50	2	100	1024	1000	2.08/2.72	0.83	3.75	3.90	0.94
ALMTokenizer Yang et al. [2025]	12.5	3	37.5	2048	412.5	2.00/2.30	0.81	3.76	3.78	0.92
WavTokenizer-large-unify Ji et al. [2024]	40	1	40	4096	480	2.00/2.52	0.80	3.77	3.79	0.95
SNACodec-speech Siuzdak et al. [2024]	46	3	81	4096	972	**2.42**/2.92	0.85	3.91	**4.01**	0.97
XCodec2 Ye et al. [2025b]	50	1	50	65536	800	2.06/2.77	0.84	**3.94**	3.90	0.97
Ours	12.5	8	100	1024	1000	2.36/**2.93**	**0.85**	3.91	3.94	**0.97**

(Analysis+Reconstruction) for speech, and from 26.33 to 20.10 for music under the reported training and evaluation protocol. This result is consistent with analysis tokens providing useful context for predicting reconstruction tokens. Perplexity is a model- and training-dependent measurement, rather than a direct estimate of the entropy in Eq. (12).

A useful interpretation is that analysis tokens act as a learned "planning variable": the model predicts a high-level representation R before generating reconstruction tokens S conditioned on it. The measured perplexity reductions support this interpretation in our experiments, but do not establish a general reduction in long-range uncertainty.

Text supervision and optimization behavior. The analysis branch is trained with text-prediction objectives and a frozen text LLM head. This encourages the discrete representation to retain information useful for producing text-grounded outputs. The training-loss curves in Figure 4 show the observed optimization behavior for the compared systems; they do not establish that the analysis-token distribution matches the LLM's hidden-state distribution or that a particular distributional divergence is minimized. We therefore treat improved alignment as a design motivation and interpret the measured transfer results empirically.

Summary. The information-bottleneck view motivates the text-aligned analysis stream, while the entropy identity describes when conditioning can reduce reconstruction-token uncertainty. In our experiments, adding analysis tokens is associated with lower reconstruction-token perplexity, and the training-loss curves characterize optimization behavior under the tested configurations. These findings support the design empirically; they do not by themselves prove that the representation is an optimal information bottleneck or that distribution shift is reduced. The reconstruction test below probes how much waveform information the analysis tokens provide under one specified decoder and training protocol.

A.3.6 Empirical test of waveform reconstruction from analysis tokens

To test whether analysis tokens alone support waveform reconstruction under our current setup, we train a flow-based decoder with the same architecture and training protocol as the FactorCodec reconstruction decoder. Results are shown in Table 2. This experiment measures reconstruction quality for the tested decoder and protocol; it does not establish that no other decoder or training procedure could recover additional waveform information from analysis tokens.

Table 14: Ablation results of SFT and GRPO on understanding tasks. Audio CLS denotes the audio classification tasks. For Audio Reasoning, we report GPT-score in terms of relevance and fluency.

Setting	ASR	Audio CLS	Audio Reasoning (GPT-score)	
			Relevance	Fluency
SFT	10.68	36.4	5.2	7.8
SFT+GRPO	**7.64**	**40.2**	**5.7**	**8.4**

Table 15: The subjective reconstruction results using MUSHRA (comparative scoring of samples) of codec models on speech, sound and music. **Bold** for the best result. FPS denotes that the frame number in one second. TPS denotes that the token number in one second. The details of MUSHRA experimental settings can be found in Appendix A.3.4.

Models	FPS/TPS	Cookbook size	Speech (↑)	Sound (↑)	Music (↑)
MimiCodec (8 RVQ) Défossez et al. [2024]	12.5/100	2048	86.7 ± 2.1	-	-
XCodec Ye et al. [2025a]	50/100	1024	78.5 ± 4.5	72.6 ± 2.1	69.8 ± 1.9
Higgs-Audio Boson AI [2025]	25/100	1024	84.4 ± 2.6	79.2 ± 1.8	81.0 ± 1.6
Encodec Défossez et al. [2022]	75/150	1024	69.3 ± 2.4	68.5 ± 2.0	62.6 ± 2.2
DAC Kumar et al. [2023]	50/100	1024	71.3 ± 1.9	70.0 ± 1.9	63.0 ± 1.8
FactorCodec (Ours)	12.5/100	1024	$\mathbf{90.5 \pm 2.8}$	$\mathbf{80.8 \pm 2.0}$	$\mathbf{86.6 \pm 2.3}$

Under this setup, analysis-only decoding scores 1.13 PESQ-WB and 1.47 PESQ-NB, compared with 2.39/2.92 from reconstruction tokens alone. This supports a functional distinction between the two streams in the tested reconstruction pathway, while the result remains specific to the evaluated decoder and training protocol.

(i) Analysis tokens provide a compact, text-oriented interface. The low analysis-only reconstruction scores are consistent with the intended information bottleneck, but do not quantify $I(R; X_{\text{acoustic}})$ or show that it is minimized.

(ii) Reconstruction tokens support the measured synthesis quality. Reconstruction tokens alone achieve PESQ close to that of the full FactorCodec in this experiment. This indicates that the analysis stream is not required for comparable PESQ under the tested decoder, rather than proving that the streams have no overlapping information.

(iii) The FiLM ablation is consistent with a conditioning role. Removing FiLM does not materially change the reported reconstruction quality (Table 12), while it affects reconstruction-token perplexity (Table 16). These results are consistent with FiLM helping condition token learning without providing a separate waveform-decoding path; they do not by themselves identify exactly which information FiLM transmits.

A.3.7 Analysis tokens as a planning mechanism

In generation tasks, analysis tokens serve a role analogous to chain-of-thought (CoT) in text LLMs: the model first emits analysis tokens that encode a high-level perceptual plan, then conditions reconstruction-token prediction on this plan. Two pieces of evidence support this view.

Quantitative. Conditioning reconstruction-token prediction on analysis tokens significantly reduces reconstruction PPL compared to predicting reconstruction tokens directly (Figure 3, left panel; speech avg. 23.77 vs. 29.36, music avg. 20.10 vs. 26.33). This mirrors the well-known PPL reduction that CoT prompting yields for text LLMs on reasoning-heavy tasks.

Qualitative. The analysis branch's LLM head allows us to decode generated analysis tokens directly back to natural language and inspect the plan. Figure 6 illustrates this on a TTS example: given the input "please call me", the decoded plan reads as a structured perceptual audit (gender, language, speech rate, key words, emotion) that the reconstruction stream then realizes acoustically.

Table 16: LLM-based perplexity (PPL) across codebooks. We present the results from first four VQ layers. Following the common practice, lower PPL denotes better performance. Note that all of tokenizers have the same codebook size (1024).

Model	Speech					Music				
	VQ1	VQ2	VQ3	VQ4	Avg	VQ1	VQ2	VQ3	VQ4	Avg
XCodec Ye et al. [2025a]	9.22	16.31	21.73	26.36	18.40	11.77	23.60	34.47	42.3	28.5
HiggCodec Boson AI [2025]	9.12	29.49	51.32	65.92	38.46	12.83	21.34	57.82	88.1	45.02
DAC Codec Kumar et al. [2023]	13.74	66.06	116.54	163.89	90.06	25.67	68.09	99.66	133.9	81.34
Analysis-only (Ours)	**3.68**	**7.73**	**15.88**	**18.99**	**11.57**	**3.81**	**6.97**	**7.26**	**8.20**	**6.56**
Reconstruction-only (Ours)	8.69	13.99	19.75	75.02	29.36	9.19	22.94	15.94	59.3	26.33
Analysis + Reconstruction (Ours)	6.92	11.37	19.42	55.36	23.77	7.02	15.55	15.23	42.5	20.1

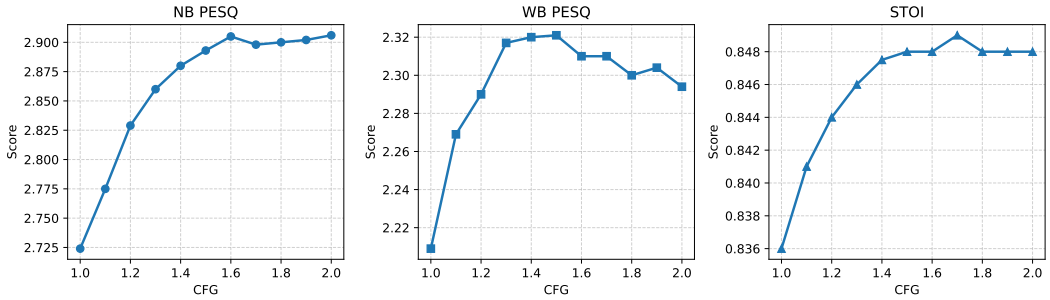

Figure 5: The influence of CFG for reconstruction performance.

A.3.8 Compared to previous semantic tokenizers

To obtain the analysis tokens, we introduce a multi-task audio understanding objective. Similar ideas of injecting supervision into a tokenizer/encoder have also been explored in CosyVoice 3 Du et al. [2025] and USTokenizer Wang et al. [2025c]. Nevertheless, our target and design are fundamentally different from prior works in several key aspects.
(1) Alignment to text LLM. CosyVoice 3 inserts an FSQ layer into an intermediate encoder layer and updates the entire encoder during training. As a result, its learned semantic tokens are not explicitly constrained to align with the latent space of a text LLM. In contrast, we *freeze* the audio encoder and learn only a lightweight linear projection that maps the analysis tokens into the latent space of the text LLM, enabling direct and stable cross-modal alignment.
(2) Modality coverage. Prior work mainly targets the speech domain, whereas our model is designed for *unified* understanding across speech, general sounds, and music.
(3) Temporal granularity. Previous approaches typically adopt finer-grained discrete representations (e.g., $\sim$25 Hz). Our analysis tokens operate at 5 Hz, yielding a much more compact representation that is better suited for high-level latent reasoning/planning rather than dense acoustic content modeling.
(4) Richer understanding targets. Beyond task-level supervision, we incorporate more detailed understanding signals, encouraging the model to capture fine-grained audio attributes and to form an explicit reasoning process over audio details.

Furthermore, we apply the same training protocol to an alternative tokenizer (CosyVoice 3 tokenizer) on the same set of understanding tasks. As shown in Table 20, its performance is consistently worse than that of our analysis tokens, validating the advantage of our design.

Controlled supervision ablation: isolating the effect of LLM alignment. To verify that the language-aligned nature of analysis tokens, not generic semantic supervision is what makes them LLM-friendly, we conduct a controlled tokenizer ablation. We train a CosyVoice3-style tokenizer that matches the architecture of our reconstruction stream (8 RVQ layers with codebook size 1024) but is trained with the CosyVoice3-style model structure and multi-task semantic supervision objective at the native 25 Hz frame rate. This setup isolates the supervision difference from architectural confounders. We then measure (a) LM perplexity on the resulting tokens alone, and (b) LM perplexity when these tokens are concatenated with our reconstruction tokens, mirroring our factorized setup but with CosyVoice3-style supervision in place of LLM-aligned supervision.

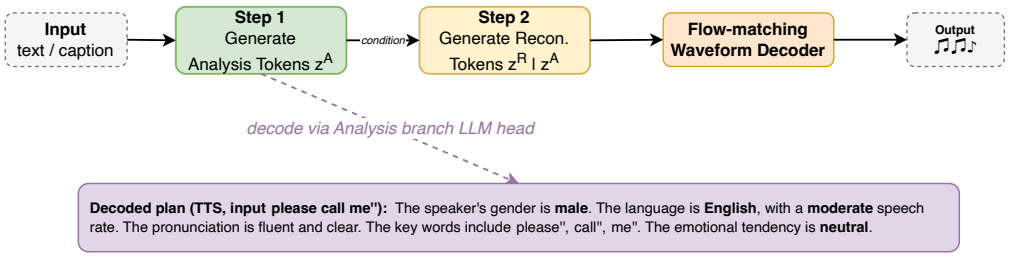

Figure 6: **Analysis tokens as an explicit plan, analogous to chain-of-thought in text LLMs.** At generation time, the model first emits analysis tokens z^A (Step 1), which can be decoded back to natural language via the analysis-branch LLM head, then predicts reconstruction tokens z^R conditioned on z^A (Step 2). The waveform is decoded from z^R alone. The decoded plan exposes the high-level perceptual structure (speaker, language, prosody, content, emotion) that guides fine-grained generation.

Results in Table 17 show two clear findings. First, the CosyVoice3-style tokens yield substantially higher LM perplexity than our analysis tokens (29.34 vs. 11.57 on speech), despite operating at a higher token rate that provides greater information capacity per second. This indicates that text-LLM alignment through our analysis branch design and SFT+GRPO is the key driver of low-perplexity tokens, not the multi-task semantic supervision used in CosyVoice3. Second, concatenating CosyVoice3-style tokens with reconstruction tokens does not reduce reconstruction-token perplexity to the level achieved by the FactorCodec factorization (37.45 vs. 23.77 average), confirming that role separation requires more than naive multi-stream concatenation: the analysis stream must itself be aligned to the LLM's latent space, and must be coupled to the reconstruction stream through explicit conditioning (Section A.3.3).

Table 17: Controlled supervision ablation. Both tokenizers use the same architecture (8 RVQ, codebook size 1024). The CosyVoice3-style baseline uses multi-task semantic supervision at 25Hz; our analysis tokens use SFT+GRPO LLM-aligned supervision at 5Hz. Lower PPL is better. We report the average PPL of first 4 VQ layers for speech and music.

Tokenizer	Standalone PPL	Concat with reconstruction PPL
CosyVoice3-style (multi-task, 25Hz)	29.34 / 30.23	37.45 / 39.16
Analysis tokens (LLM-aligned, 5Hz)	**11.57 / 6.56**	**23.77 / 20.1**

A.3.9 The influence of classifier-free guidance and decoding steps

We evaluate the effect of classifier-free guidance (CFG) and the number of flow-decoding steps for waveform reconstruction. Figure 5 shows speech reconstruction results across CFG values; we use CFG=1.5 by default. We use 10 flow steps for speech and music, and 25 steps for sound. Table 18a reports decoder RTF, while Table 18b reports the LLM-prediction and decoder components. The RTF measurements were run on an NVIDIA A100 with batch size 1 and 10-s audio inputs. For our model, the reported LLM-prediction and decoder values are 0.89 and 0.062; their arithmetic sum is 0.952. This sum is a component-based estimate, not a separately measured end-to-end RTF, so the table supports a comparison of the reported components but does not establish complete-pipeline latency.

A.3.10 Why flow matching based decoding?

We adopt a flow-based decoder for three main reasons.

First, it provides higher-fidelity reconstruction, particularly for complex audio such as sound and music, where lightweight neural codecs (e.g., GAN-based decoders) often struggle to preserve fine-grained details at low frame rates (e.g., 12.5 Hz). As shown in Table 19, the flow-based decoder consistently outperforms the GAN-based decoder (following MimiCodec Défossez et al. [2024]) across multiple reconstruction metrics.

Second, the flow-based formulation naturally supports *cross-resolution generation*, where higher-resolution audio is directly modeled and reconstructed from lower-resolution conditioning within a

Table 18: Reported Real-Time Factor (RTF) for the decoder and LLM-prediction components, measured on an NVIDIA A100 with batch size 1 and 10-s audio inputs. The arithmetic sum for UniAudio 2.0 is 0.952; it is not a separately measured end-to-end RTF.

(a) RTF of different audio tokenizers decoding.

Model	RTF ↓
Higgs-Audio	0.0059
XCodec	0.0068
UniAudio 2.0 (10 steps)	0.062

(b) LLM-prediction and waveform-decoder RTF.

Model	**LM prediction ↓**	**Waveform decoder ↓**
MiMo-Audio	1.48	**0.0071**
UniAudio 2.0	**0.89**	0.062

Table 19: Comparison of GAN-based and flow-based decoders on speech, sound, and music reconstruction. Higher is better for all metrics.

Speech	PESQ (↑)	STOI (↑)	UTMOS (↑)	VISQOL (↑)	SIM (↑)
GAN-based decoder	2.56	0.84	3.83	3.85	0.94
Flow-based decoder (Ours)	**2.93**	**0.85**	**3.91**	**3.94**	**0.97**
Sound	VISQOL (↑)	CE (↑)	CU (↑)	PC (↑)	PQ (↑)
GAN-based decoder	2.86	4.08	4.89	3.34	5.87
Flow-based decoder (Ours)	**3.10**	**4.12**	**5.06**	**3.58**	**5.96**
Music	VISQOL (↑)	CE (↑)	CU (↑)	PC (↑)	PQ (↑)
GAN-based decoder	3.90	6.82	6.59	5.94	7.34
Flow-based decoder (Ours)	**4.03**	**7.51**	**7.68**	**6.12**	**7.87**

single decoder, rather than using a separate super-resolution module. This makes it straightforward to extend to higher sampling rates or multi-channel audio (e.g., 48 kHz stereo). In our framework, the decoder predicts high-resolution SQ-Codec latents conditioned on lower-sample-rate inputs (e.g., 24 kHz MuEncoder and 16 kHz Whisper/WavLM features), providing semantic-rich guidance for high-fidelity reconstruction. This flexibility is particularly important for high-quality audio generation.

Third, while flow-based decoding introduces multi-step inference and increases decoding complexity, this trade-off is controllable. As discussed in Appendix A.3.9, high-quality reconstruction can be achieved with a relatively small number of steps. Moreover, prior work Liu et al. [2025a] shows that flow-based models can be distilled to 1–2 steps with minimal quality degradation. We leave such acceleration to future work.

A.4 Audio Tokenizer Baselines

To make a fair comparison, we classify audio tokenizers into two types: (1) speech-based tokenizers trained on speech datasets, and (2) audio-based tokenizers trained on speech, sound, and music datasets. In this study, we mainly compare against audio-based tokenizers trained on speech, sound, and music datasets. Below, we list our chosen baselines:

(1) Encodec Défossez et al. [2022], a SOTA audio codec model trained on large-scale speech, sound, and music datasets. We use the official open-sourced 24 kHz version. The frame rate is 75 Hz.

(2) DAC-Codec Kumar et al. [2023], which offers very high reconstruction performance. It is trained on large-scale speech, sound, and music datasets. The official open-sourced 24 kHz version is used. The frame-rate is 50 Hz.

(3) MimiCodec Défossez et al. [2024], a SOTA low-bitrate speech codec model trained on a large-scale speech dataset. The sampling rate is 24 kHz, the frame-rate is 12.5 Hz.

(4) X-Codec Ye et al. [2025a], a semantic-rich audio codec model trained on a large-scale speech, sound, and music datasets. The sampling rate is 16 kHz, the frame-rate is 50 Hz.

Table 20: Ablation study of different semantic tokenizer. We report the ASR and emotion recognition (ER) tasks.

Model	**ASR** (↓)	**ER** (↑)
CosyVoice3 Du et al. [2025]	32.3	34.5
Ours	**10.1**	**50.2**

Table 21: Comparison of different quantization strategies. For each strategy, we control the same VQ layers (8 VQ layers), training data, and decoder.

Model	**PESQ (WB)**	**PESQ (NB)**	**STOI**	**UT-MOS**	**ViSQOL**	**SIM**	$\mathbf{PPL}_{\text{speech}}$	$\mathbf{PPL}_{\text{music}}$
RVQ	2.12	2.76	0.83	3.84	3.78	0.95	32.8	29.2
Query-based	**2.54**	**3.07**	**0.85**	3.89	3.88	**0.97**	26.3	24.4
Group-wise VQ	2.36	2.93	**0.85**	**3.91**	**3.94**	**0.97**	**23.77**	**20.1**

(5) Higgs-Audio-tokenizer[3], a SOTA audio codec model trained on a large-scale speech, sound, and music dataset. The sampling rate is 24 kHz, the frame-rate is 25 Hz.

(6) XCodec2 Ye et al. [2025b][4], a SOTA speech codec model trained on a large-scale speech dataset. The sampling rate is 16 kHz, the frame-rate is 50 Hz with single FSQ layer.

(7) SNACodec speech version Siuzdak et al. [2024], a SOTA multi-scale speech codec model trained on a large-scale speech dataset.

(8) WavTokenizer Ji et al. [2024], a single-layer audio codec model trained on a large-scale speech, sound, and music dataset.

For the reconstruction settings in Table 13, the nominal bitrate is computed as token rate $\times \log_2$(codebook size). At 1024-entry codebooks, DAC-Codec and X-Codec (100 tokens/s), Higgs-Audio (100 tokens/s), and FactorCodec reconstruction (100 tokens/s) each use 1000 bps; EnCodec (150 tokens/s) uses 1500 bps. MimiCodec uses 1100 bps (100 tokens/s with a 2048-entry codebook). The table reports the nominal BPS for the other listed configurations as well. The comparison is at the evaluated configurations, not a multi-bitrate sweep. For RVQ baselines with adjustable depth, we reduce active layers where applicable to target a similar token rate; this does not guarantee an equal BPS.

Speech Analysis Case

The audio begins with a clear, albeit slightly muffled, sound. The speaker's gender is discernible as male. The language is English. There is only one speaker. The recording environment seems to be indoors, possibly a quiet room, but with some ambient background noise, suggesting it's not a professional studio. The speaker exhibits a moderate speech rate. There are no significant pauses indicating hesitation or thought, suggesting a natural flow. The pitch variation is relatively consistent, without dramatic shifts, implying a neutral to slightly conversational prosody. The pronunciation is generally fluent and clear, with no noticeable impediments. The audio contains the spoken phrase '00:03'. This phrase is a time indicator. The emotional tendency is neutral, and the tone is informative or declarative.

[3]https://github.com/boson-ai/higgs-audio

[4]https://huggingface.co/HKUSTAudio/xcodec2

Sound Analysis Case

The audio consists of a distinct, rhythmic pattern of impacts. These impacts are predominantly in the mid-frequency range and suggest a mechanical or physical action occurring outdoors or in a large space. There is a clear periodic repetition. The primary sound event is a repeating impact, sounding like a hammer or a similar percussive tool. The first instance occurs around 00:00:01 and repeats approximately every second, with a sharp, distinct onset. The duration of each impact is very short, less than half a second. The spectral content is broadband but with a strong emphasis on mid to high frequencies due to the impact. There's a slight sense of reverberation, suggesting the sound is not in a completely enclosed space but perhaps outdoors or in a large hall. The sound appears to be localized to a central or slightly directional source. The repetitive striking sound strongly suggests construction work, hammering, or perhaps pile driving. The consistent rhythm and sharp impacts indicate a deliberate and ongoing activity. The emotional connotation is neutral to slightly industrious, not alarming. The source interaction is minimal, with the sound of the impact itself being the dominant element, potentially masking very quiet background noises. The action appears to be a single, repeating task.

Music Analysis Case

The audio has a moderate to high fidelity, with clear instrumentation and vocals. The genre appears to be rock or hard rock, characterized by distorted electric guitars, a prominent drum beat, and aggressive vocals. The main instruments are electric guitar, bass guitar, and drums, with male vocals. The tempo is moderately fast, around 120 BPM, with a consistent and driving beat. The key appears to be A minor, and the chord progression is likely based on standard rock progressions, possibly with power chords. The meter is 4/4 time, and the rhythmic complexity is moderate, with a strong backbeat. The melody is delivered by the vocals and guitar, with a relatively narrow range in the vocals and more varied melodic lines in the guitar riffs. The timbre is dominated by distorted guitars, a solid bass tone, and punchy drums. Dynamics are generally loud, with some variation in intensity. The mood is energetic, aggressive, and defiant. The lyrics, though partially obscured by the music, seem to convey a sense of rebellion or challenging authority ('smash the machine', 'ain't no stopping'). The expressive techniques include distorted guitar tones, heavy drumming with prominent cymbal work, and powerful, somewhat raw vocals. A potential hook can be identified in the main guitar riff and vocal chorus, which is memorable and driving. The structure appears to follow a typical verse-chorus rock song format, likely with an intro, verses, choruses, and potentially a bridge or guitar solo (though not fully audible in this short clip).

B UniAudio 2.0

B.1 Model Configuration

We list the detailed configuration of UniAudio 2.0 in Table 22.

B.2 Training Data Details

We organize the training corpus into several data types, each corresponding to a class of tasks supported by UniAudio 2.0. The training data includes the following data types:

Text-only data To maintain and stabilize the language modeling capability of UniAudio 2.0, we incorporate large-scale text-only data during pre-training. Specifically, we use approximately 100B tokens from the high-quality annealed corpus of OLMo 3 [Team OLMo et al., 2025]. This data is used to maintain the text capability of the original LLM when we introduce audio-modality data.

Audio-only data To expose the model to diverse acoustic patterns beyond speech, we use large-scale unlabeled audio data from LAION-Audio-300M.[5] This dataset covers speech, environmental sounds, and music.

[5] `https://huggingface.co/datasets/laion/LAION-Audio-300M`

Table 22: UniAudio 2.0 architecture configuration.

Module	Configuration
Backbone	Init from `LLaMA-3.2-3B`
Audio understanding experts	#layers $= 3$; $d_{\text{model}} = 3072$; #heads $= 24$; $T =$ [2048]
Cross-modal experts	#layers $= 28$; $d_{\text{model}} = 3072$; #heads $= 24$; $T =$ [2048]
Audio generation experts	#layers $= 2$; $d_{\text{model}} = 3072$; #heads $= 24$; $T =$ [2048]
Audio Local Decoder	#layers $= 4$; $d_{\text{model}} = 2048$; #heads $= 32$; $T =$ [8]
Total layers	$L = 37$
Text tokenizer	`LLaMA3.2 tokenizer`
Audio tokenizer	`FactorCodec`
Audio sample rate	[24 kHz]
Analysis token rate	[5 Hz]
Reconstruction token rate	[12.5 Hz]

Speech-transcription data Paired speech and transcription data are used to construct automatic speech recognition (ASR) and text-to-speech (TTS) tasks. We collect such data from multiple sources, including Emilia [He et al., 2025], YODAS-English [Li et al., 2023], WenetSpeech [Zhang et al., 2022], and WenetSpeech-Yue [Li et al., 2025b].

Speech-caption-transcription data To support speech captioning and instruction-following TTS tasks, we use datasets that provide speech, transcription, and descriptive captions. In particular, we adopt CapSpeech-MLS dataset [Wang et al., 2025a], following prior work on instruction-based speech generation.

Audio-caption data Audio-caption pairs enable text-to-audio and text-to-music generation. We aggregate such data from multiple sources, including WavCaps Mei et al. [2023], AudioSet Gemmeke et al. [2017], the Million Song Dataset (MSD) Bertin-Mahieux et al. [2011], and YouTube-8M Abu-El-Haija et al. [2016].

Lyric-song data To model lyric recognition and lyric-to-song generation, we construct lyric–song pairs by following the preprocessing pipeline of SongGen [Liu et al., 2025b] on the MSD dataset, Free Music Archive Defferrard et al. [2017] and MTG-Jamendo Dataset Bogdanov et al. [2019].

B.3 Composed audio sequences.

To improve compositional generalization and long-context reasoning over audio, we introduce *composed audio sequences*, inspired by [Bai et al., 2023]: a long-context training sequence composed of multiple, related segments designed to encourage cross-segment compositional reasoning. We frame this as a *compositional curriculum* rather than data augmentation. The intuition is formalized as follows. Let $\mathcal{P}_{\text{single}}$ denote the standard single-task training distribution (independent input-output pairs):

$$\mathcal{P}_{\text{single}} = \{(x_i, y_i)\}_{i=1}^{N}, \tag{13}$$

which only supervises input-output mappings $f : x \mapsto y$. In contrast, composed sequences sample from a structured distribution

$$\mathcal{P}_{\text{composed}} = \{[u_1, u_2, \ldots, u_M] \mid (u_1, \ldots, u_M) \sim \mathcal{R}\}, \tag{14}$$

where $\mathcal{R}$ is one of several relational structures (decomposition, same-speaker continuation, content invariance under style change, etc.). Under the next-token objective, training on $\mathcal{P}_{\text{composed}}$ implicitly supervises the *operation* (e.g., decomposition) rather than any specific f, which is what enables compositional generalization at inference.

We instantiate five sequence types, each targeting a distinct compositional skill. We summarize the data proportion of each type in Table 24

(1) Long-form audio segmentation (*long-context dependency*). We split each long-form recording (speech conversation, environmental recording, or song) into 2–8 segments, ensuring the token

sequence does not exceed 2048. Sources: LAION-Audio-300M, LibriLight Kahn et al. [2020], WavCaps, MSD, DailyTalk Lee et al. [2023], Expresso Nguyen et al. [2023].

(2) Same-speaker speech-text interleaving (*speaker-identity tracking*). For a fixed speaker, we sample multiple speech-transcription pairs and concatenate them with a fixed (speech, text) ordering:

$$[s_1, t_1, s_2, t_2, \ldots, s_K, t_K], \quad s_i, t_i \sim \mathcal{D}_{\text{speaker}=k}. \tag{15}$$

Sources: MLS Pratap et al. [2020], LibriSpeech Panayotov et al. [2015].

(3) Audio/music-caption interleaving (*caption-audio alignment*). Alternating audio and caption segments of the form $[a_1, c_1, a_2, c_2, \ldots]$, sampled from AudioSet, WavCaps, AudioCaps, and MSD.

(4) Mixture-clean triples (*source decomposition*). For two audio clips a, b drawn from a pre-defined pool $\mathcal{A}$ (speech from MLS, LibriSpeech, WenetSpeech; sound from AudioSet, WavCaps; music from MSD), we form a mixture $c = a + b$ and arrange them as a composed sequence:

$$[\sigma(a, b, c)], \quad \sigma \in \{\text{permutations}\}, \quad c = a + b. \tag{16}$$

We additionally concatenate multiple triples to form longer composed sequences such as $\{a_1, b_1, c_1, a_2, b_2, c_2, \ldots\}$, exposing the model to multiple independent decomposition instances within a single context. This structure naturally supports in-context learning at inference: the model learns to apply the decomposition operation across multiple examples within a sequence, mirroring the few-shot prompt format used at evaluation time. We additionally use demucs Rouard et al. [2023] to separate music into vocals and accompaniment for diversity. Crucially, the model is never told "this is denoising": it only sees the structural pattern $c = a + b$ under next-token prediction.

(5) Content-consistent style-varied speech (*content-style disentanglement*). Following InstructSpeech Huang et al. [2024], we use a TTS model to synthesize K versions of the same text under different prosodic styles (pitch, volume, speed, emotion):

$$[s_1, s_2, \ldots, s_K], \quad \text{Content}(s_i) = \text{Content}(s_j), \quad \text{Style}(s_i) \neq \text{Style}(s_j). \tag{17}$$

This forces the model to recognize content and style as independent attributes.

Held-out evaluation data and compositional transfer. The evaluation examples are held out from the training data: no evaluation audio examples are duplicated in the training sets used for FactorCodec or UniAudio 2.0. Table 23 lists the evaluation datasets and the training sequence types that provide related compositional skills. Thus, the generalization claim concerns applying or recombining skills on held-out examples; it does not imply that every underlying operation is absent from training. For instance, Stage 4 contains mixture-clean triples, while the few-shot denoising evaluation uses held-out WHAM examples.

Table 23: Held-out evaluation data and related training skills. Evaluation examples are disjoint from training data; the listed training sequence types provide related skills that can be applied or recombined at evaluation.

Eval task	Required skill(s)	Training source of skill(s)	Eval data (disjoint)
FS speech denoising	Decomposition	Type 4: $\{a, b, a+b\}$	WHAM Wichern et al. [2019]
FS voice conversion	Speaker tracking + disentanglement	Types 2 + 5	VCTK Veaux et al. [2017]
FS sound classification	Decomposition + Alignment	Types 3 + 4	TUT 2017 Mesaros et al. [2017]
FS emotion classification	Style disentanglement	Type 5	ESD Zhou et al. [2022]
A-I-TTS (zero-shot)	Speaker tracking + instruct TTS	Types 2 + 3 + 5	CapSpeech-PT Wang et al. [2025a]
Speech-S (zero-shot)	Alignment + Composition	Types 3 (+ Stage 3 TTS)	CapSpeech-SEDB Wang et al. [2025a]

What is novel at inference. At inference, the model must (i) recognize from a small number of in-context demonstrations which compositional operation is being requested, (ii) apply this operation to a novel query, and (iii) generalize across structural variations not seen in pre-training (e.g., a specific WHAM noise type, a specific VCTK speaker). The few-shot/zero-shot framing thus reflects how the model accesses pre-induced operations at inference, not whether the operations themselves are entirely novel.

Compositional vs. fundamentally novel generalization. We claim compositional generalization, not generalization to fundamentally novel operations. Tasks requiring operations not represented in

Table 24: Composition of Stage 4 mid-training data (approximately 20B audio tokens in total). Stage 4 mixes a replay of Stage 3 task data (to preserve seen-task performance) with newly constructed composed audio sequences (to induce compositional skills). Percentages within composed sequences reflect the relative coverage of each compositional skill.

Component	Primary data sources	% of Stage 4
Stage 3 task replay (50%)		
ASR / TTS / Caption / etc.	Replay of Stage 3 mixture	50%
Composed audio sequences (50%)		
1. Long-form segmentation	LAION-Audio-300M, LibriLight, MSD	20%
2. Same-speaker speech-text	MLS, LibriSpeech	10%
3. Audio/music-caption	AudioSet, WavCaps, AudioCaps, MSD	9%
4. Mixture-clean triples	Combinations from speech / sound / music sources	8%
5. Content-consistent style-varied	TTS-synthesized from text corpora	3%

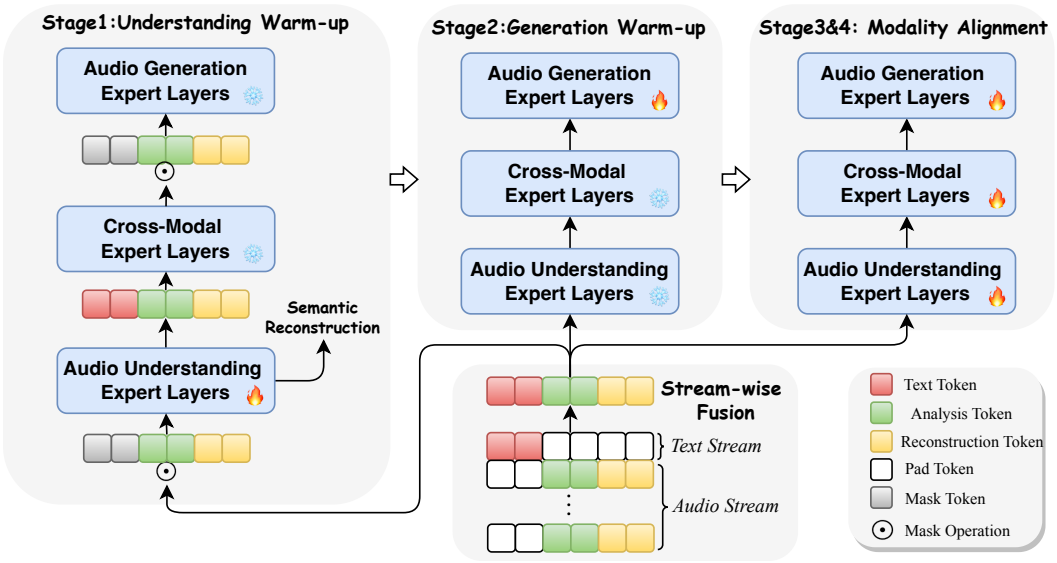

Figure 7: Overview of the multi-stage pre-training strategy of UniAudio 2.0. The model is trained in a progressive manner, including an understanding warm-up stage, a generation warm-up stage, and a modality alignment stage, with different expert layers gradually unfrozen. For simplicity, the audio generation expert layers include both the generation expert and the local decoder.

any composed sequence type (e.g., speaker diarization with explicit temporal segmentation) are out of scope and discussed in Appendix E. This is consistent with how compositional generalization is defined in the in-context learning literature for language [Brown et al., 2020] and vision models [Bai et al., 2023].

B.4 The details of multiple stage training

As Figure 7 shows, the training process of UniAudio 2.0 includes four stages. We list the configuration of the four training stages in Table 25. For all four stages, we use 64 NVIDIA H100 GPUs to train the model. Below, we describe each stage in detail.

Stage 1: Audio understanding warm-up In the first stage, we focus on initializing the audio understanding experts. We train the model using a subset of audio understanding tasks, while freezing all other components. To encourage the understanding experts to encode rich semantic information, we introduce an auxiliary semantic distillation objective. Following FactorCodec (Section 3), a lightweight decoder is attached to reconstruct continuous semantic features extracted from frozen WavLM and music SSL encoders. The overall objective includes a reconstruction loss and an LM text

Table 25: Four-stage training recipe of UniAudio 2.0. "Trainable" indicates the modules updated at each stage, and "Ctx" denotes the maximum context length used during training. LR denotes the learning rate. The text-only tokens are used only in Stages 3 and 4. Audio-token budgets count global temporal positions, computed as $17.5 \cdot D_s$ for an audio duration of D_s seconds. The stage-wise budgets sum to 76B tokens; the de-duplicated audio corpus contains 60B unique tokens. Audio clips may be reused across tasks and stages.

Stage	Goal	Task	Audio Data	Steps	LR	Ctx	Trainable
1	Understanding warm-up	ASR, Audio Caption Music Caption Lyric recognization	3B	50k	2e-4	1024	Understanding experts
2	Generation warm-up	TTS, text-to-audio, song/music generation	3B	50k	2e-4	1024	Generation experts + audio local decoder
3	Audio–text pre-training	Audio understanding and generation, text-only, audio-only	50B	500k	2e-4	1024	All parameters
4	Mid-training	subset of stage 3 mixture, new composed audio sequences	20B	300k	1e-4	2048	All parameters

loss. After training, we discard the decoder. Formally, the training objective in this stage is defined as

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{LM}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}, \tag{18}$$

where $\mathcal{L}_{\text{LM}}$ denotes the language modeling loss on text outputs, and $\mathcal{L}_{\text{rec}}$ is the reconstruction loss for distilling continuous semantic features. Specifically, the reconstruction loss is given by

$$\mathcal{L}_{\text{rec}} = \|D(h) - z_{\text{SSL}}\|_2^2, \tag{19}$$

Stage 2: Audio generation warm-up. In this stage, we update the audio generation expert and the local audio decoder. We train the model on a subset of audio generation tasks while keeping the understanding and cross-modal experts fixed. The training objective is defined as

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{AR}}, \tag{20}$$

where $\mathcal{L}_{\text{AR}}$ denotes the weighted autoregressive prediction loss over multi-stream audio tokens.

Weighted autoregressive loss over multi-stream tokens. Our FactorCodec produces $L{=}8$ token streams per audio frame. Let $s_t^{(\ell)}$ denote the token from stream $\ell \in \{1, \ldots, L\}$ at time step $t \in \{1, \ldots, T\}$. We define the stream-weighted autoregressive loss as

$$\mathcal{L}_{\text{AR}} = -\sum_{t=1}^{T} \sum_{\ell=1}^{L} w_\ell \, \log p_\theta \Big(s_t^{(\ell)} \mid x, \, s_{<t}^{(1:L)} \Big), \tag{21}$$

where x denotes the conditioning input (e.g., a text prompt), $s_{<t}^{(1:L)}$ denotes all token streams before time step t. In our implementation, we set

$$\mathbf{w} = \Big[\tfrac{2}{8}, \, \tfrac{2}{8}, \, \tfrac{2}{8}, \, \tfrac{1}{8}, \, \tfrac{1}{8}, \, \tfrac{1}{8}, \, \tfrac{1}{8}, \, \tfrac{1}{8} \Big]. \tag{22}$$

Stage 3: Audio-text pre-training. We jointly update all model parameters on a mixture of audio understanding tasks, audio generation tasks, text-only data, and audio-only data. This stage aligns the two modalities under a unified autoregressive objective. We use a maximum context length of 1024 tokens in this stage. The overall training loss is

$$\mathcal{L}_{\text{AR}} = \lambda_{\text{text}} \mathcal{L}_{\text{text}} + \lambda_{\text{audio}} \, \mathcal{L}_{\text{audio}}, \tag{23}$$

where $\lambda_{\text{text}} = 1.6$ and $\lambda_{\text{audio}} = 1$. Here, $\mathcal{L}_{\text{audio}}$ is instantiated as the weighted token-level loss in Eq. (21). This setting is based on early experiments, which helps preserve the LLM's text capability and improves audio understanding performance.

Stage 4: Audio-text mid-training In the final stage, we aim to extend the effective context length and enhance generalization to unseen tasks. We continue training on a subset of the Stage 3 pre-training data, augmented with our constructed composed audio sequences data. This stage encourages the model to model longer and more complex audio-text sequences and improves robustness across diverse task settings. We use a maximum context length of 2048 tokens in this stage. The training objective remains identical to that in Stage 3

B.5 The details of evaluation data and evaluation metrics

UniAudio 2.0 is a multi-task audio foundation model, which supports multiple audio-related tasks. For each task, we follow the commonly used evaluation benchmark and metrics. In the following, we present the details for each task.

B.5.1 Seen tasks

ASR For the ASR task, we follow previous works Xu et al. [2025a], Zhang et al. [2025], Li et al. [2025b], and choose LibriSpeech-test-clean, LibriSpeech-test-other, SEED-TTS-Eval-EN, SEED-TTS-Eval-ZH, and WSYue-ASR-eval Li et al. [2025b] as the evaluation benchmark. For English datasets, we use WER as the metric. For Chinese and Cantonese, we use CER as the metric.

TTS For the TTS task, we use LibriSpeech-test-clean, SEED-TTS-Eval-EN, SEED-TTS-Eval-ZH, WSYue-TTS-eval Li et al. [2025b] as the benchmark. For evaluation, we use Whisper-large-v3 to evaluate the English speech performance, Paraformer-zh is used for Chinese speech, and SenseVoice-s-Yue is used for Cantonese speech. Furthermore, we also use DNS-MOS to evaluate the speech quality.

Instruct TTS For the Instruct TTS task, we follow the setting of CapSpeech Wang et al. [2025a] and use WER, style accuracy, and UTMOSv2 as evaluation metrics. Although our model is not trained on Chinese Instruct TTS, we show that it performs well on Chinese instruction TTS by following InstructTTS Yang et al. [2023b] and using the same evaluation dataset, refer to our Demo Pages.

Audio Caption For the audio caption task, we follow Audio Flamingo 3 Goel et al. [2025] and use CIDER as the evaluation metric. Furthermore, we also use GPT-score to evaluate the model's prediction. The prompt is shown in Box B.5.2.

Music Caption Similarly, for the music caption task, we use MusicCaps test set. The same evaluation metrics are used as the audio caption task.

Audio Generation For the text-to-audio task, we follow the setting of Stable Audio Open Evans et al. [2025], using FD, KL, and CLAP score as the metric.

Music Generation For the text-to-music task, we follow MusicGenCopet et al. [2023], using FAD, KL, and CLAP-score as the metrics.

Song Generation For the song generation task, we follow SongGen Liu et al. [2025b], using their benchmark and evaluation metrics: WER and AudioBox Score Tjandra et al. [2025].

Lyric Recognition For the lyric recognition task, we use the benchmark from SongGen.

B.5.2 compositional generalization

Few-shot speech denoising task We build the few-shot speech denoising evaluation set based on LibriTTS-test-clean and WHAM noise Wichern et al. [2019] to build the mixture-clean pairs. For the evaluation, we follow Mimo-AudioZhang et al. [2025], using PESQ, STOI, WER, DNS-MOS as the metrics.

Few-shot voice conversion We build the few-shot voice conversion evaluation set based on VCTK Veaux et al. [2017] dataset. For the evaluation, we follow Mimo-Audio, using WER, Speaker Similarity (SIM), and DNS-MOS as the metrics.

Few-shot emotion classification We build the few-shot emotion classification evaluation set based ESD Zhou et al. [2022] dataset. We use both English and Chinese splits to build the evaluation set.

Few-shot sound event classification We build the few-shot sound event classification evaluation set based on TUT acoustic scenes 2017 Mesaros et al. [2017].

Text understanding We follow the standard zero-shot evaluation setting for text LLMs and use the MMLU dataset Hendrycks et al. [2020] as the benchmark. We note that some previous work Tian et al. [2025a] uses a few-shot setting to evaluate text understanding ability; in this study, we directly evaluate its zero-shot ability.

Speech-to-speech/text question answer During our training stage, we do not add the speech conversation data. Instead, we build a lot of composed audio sequences without explicit instruction. Thus, we view the speech-to-speech/text question answer as the zero-shot task. We follow LLAMA-Omni Fang et al. [2024], and use the InstructS2S-Eval as the benchmark. Following LLAMA-Omni, GPT-score is used as the metric. We use the same prompt to evaluate the performance as LLAMA-Omni.

Dysarthric Speech Recognition Dysarthric speech recognition is very similar to the ASR task, but it asks the model to recognize dysarthric speech. Each utterance only includes one word. We also include it as a zero-shot setting, since the model was not trained on such data. For Dysarthric Speech Recognition, we use the UASpeech Kim et al. [2008] dataset with four selected speakers (M12, F02, M16, F04), following CoLM-DSR Chen et al. [2024] – 50 utterances per speaker.

Audio prompt and caption guided TTS We define a new task: using audio prompt to provide the timbre and the caption to guide the speaking style. We use speaker similarity, style accuracy, WER, UTMOSv2 as the metrics. For Audio-prompt-and-caption guided TTS, we use CapSpeech-PT-test Wang et al. [2025a], matching utterances to same-speaker prompts by speaker ID (filtering utterances without a matching speaker, yielding 2266 valid samples from 2800).

Speech-sound generation We define a new task: asking the model to generate speech and corresponding sound event. The input includes sound event tag and speech content. We use WER, CLAP score, and UTMOSv2 as the metrics. For speech-sound generation, we use CapSpeech_SEDB test set, synthesized from LibriTTS-test, VGGSound, and ESC-50 Piczak [2015].

Prompt: Audio Caption Evaluation

```
You are an expert evaluator for audio captioning.

Given:
- PRED (model caption)
- GT (human reference caption)

Evaluate PRED against GT and output ONLY a valid JSON object.

Scoring (integers 0-10):
- relevance: whether PRED describes the same sound event(s) as GT
- fluency: whether PRED covers key details without grammaticality
Also provide:
- overall: integer 0-10 (holistic score)
- brief_reason: <= 60 words, concise explanation focusing on
    mismatches
```

B.6 The details of experimental results

Due to the page limit, we only choose some representative works in Table 3. In this part, we will compare with more baseline models on different tasks.

B.6.1 ASR performance comparison

Table 30 shows the ASR performance comparison on Chinese and English benchmarks. Table 31 shows the Cantonese ASR performance. We compare with SenseVoice-Yue Li et al. [2025b], Whisper-large-v3 Radford et al. [2023], and Qwen2.5-Omni.

Table 26: Architecture ablation on speech domain data (Emila, MLS, LibriSpeech, and WenetSpeech), using only text-only, ASR, and TTS tasks. This experiment evaluates the effectiveness of the proposed understanding experts (under-expert), semantic reconstruction loss (se-rec), and generation experts (gen-expert). We conduct these experiments with 1B cross-modal experts. We evaluate the ASR and TTS tasks on Seed-TTS EN/ZH benchmarks. For the text ability, we conduct zero-shot evaluation on MMLU benchmark.

Settings	**ASR** *WER/CER* (↓)	**TTS** *WER/CER* (↓)	**MMLU** *Acc.* (↑)
Full-1B	**3.0 / 3.4**	**4.0 / 4.8**	**31.0**
w/o under-expert and se-rec	3.9 / 4.4	4.2 / 5.1	27.4
w/o se-rec loss	3.8 / 4.4	4.0 / 4.9	28.8
w/o gen-expert	3.2 / 3.6	4.3 / 5.8	29.2

Table 27: Ablation study investigating whether build unified models bring benefits. We conduct experiments on speech domain data (Emila, MLS, LibriSpeech, and WenetSpeech). For fair comparison, we control the same epoch is trained for each settings. We evaluate the ASR and TTS tasks on Seed-TTS EN/ZH benchmarks.

Settings	**ASR** *WER/CER* (↓)	**TTS** *WER/CER* (↓)
1B-ASR only	3.8 / 4.2	-
1B-TTS only	-	**3.7 / 4.1**
Unified 1B with text	**3.0 / 3.4**	4.0 / 4.8
Unified 1B without text	3.3 / 4.0	3.8 / 4.3
3B-ASR only	3.2 / 2.9	-
3B-TTS only	-	3.3 / **3.6**
Unified 3B with text	**2.5 / 2.1**	3.4 / 3.6
Unified 3B without text	3.1 / 2.6	**3.2** / 3.7

B.6.2 TTS performance comparison

Table 32 shows the TTS performance on Seed-TTS-eval (EN/ZH), LibriSpeech-clean (EN), and WenetSpeech-Yue-TTS (Cantonese) benchmarks.

B.6.3 Audio/music caption performance comparison

Table 28 shows the performance comparison on audio and music caption benchmarks.

B.6.4 Comparison with Gemini and Kimi-Audio

Table 33 shows the performance comparison between our UniAudio 2.0 and Gemini 2.5 Flash and Kimi-Audio KimiTeam et al. [2025] on TTS, ASR, and song recognition tasks. We can see that our model achieves comparable performance with these strong baselines.

C Related works

Due to space constraints in the main paper, we provide an extended discussion of related work here, covering audio understanding models, audio tokenizers, unified audio foundation models, and the information-theoretic and chain-of-thought literature that grounds our design choices.

Audio understanding models. Audio LLMs that focus on understanding typically couple a frozen audio encoder—HuBERT [Hsu et al., 2021, Lee et al., 2021], WavLM [Chen et al., 2022], Whisper [Radford et al., 2023], or wav2vec 2.0 [Baevski et al., 2020, 2022]—with a pre-trained text LLM through cross-attention or a learned projection. SALMONN [Tang et al., 2024], Qwen2-Audio [Chu et al., 2024], Qwen2.5-Omni [Xu et al., 2025a], and Audio Flamingo [Goel et al., 2025] are representative of this paradigm. These systems achieve strong perception and reasoning on speech, sound, and

Table 28: Captioning results on audio and music.

Model	Audio Caption			Music Caption	
	CIDEr	Relevance	Fluency	Relevance	Fluency
SALMON 13B Tang et al. [2024]	0.355	5.41	8.06	3.24	7.92
Music Flamingo Ghosh et al. [2025a]	–	–	–	4.74	8.19
Audio Flamingo Kong et al. [2024]	0.51	5.25	8.01	–	–
Audio Flamingo 2 Ghosh et al. [2025b]	0.58	5.71	8.13	–	–
Audio Flamingo 3 Goel et al. [2025]	0.79	6.34	8.24	5.77	8.24
Qwen2.5-Omni-7B Xu et al. [2025a]	0.39	5.61	8.21	5.33	7.90
UniAudio 2.0 (Ours)	0.69	5.51	8.31	5.14	8.00

Table 29: More compositional generalization results across tasks. Metrics: S2S denotes speech-to-speech/text instruction-following, we report S2S/S2T GPT-score; DSR reports WER (%).

Task	Model	Score
S2S	LLAMA-Omni	**3.47 / 3.99**
	SpeechGPT	2.19 / 2.98
	Ours	2.16 / 3.66
DSR	Whisper-DSR Xiong et al. [2019]	27.5
	Ours	**19.4**

music tasks, but their continuous-feature interface makes scalable autoregressive generation difficult: generation typically requires either a separate diffusion or flow-matching head [Yan et al., 2025], or a hybrid stream of discrete codec tokens bolted onto the same backbone [KimiTeam et al., 2025].

Audio tokenizers. Discrete audio tokenizers fall into two families with distinct design objectives. *Acoustic codecs*: SoundStream [Zeghidour et al., 2022], EnCodec [Défossez et al., 2022], DAC [Kumar et al., 2023], HiFi-Codec [Yang et al., 2023a], WavTokenizer [Ji et al., 2024] are trained primarily for waveform reconstruction with VQ or RVQ, producing tokens that are LM-friendly but encode mostly low-level acoustic detail. *Semantic tokenizers*: SpeechTokenizer [Zhang et al., 2023b], X-Codec [Ye et al., 2025a], XYCodec [Gong et al., 2025], Mimi [Défossez et al., 2024], CosyVoice [Du et al., 2024] and CosyVoice 3 [Du et al., 2025], ALMTokenizer [Yang et al., 2025], inject SSL features or task-level supervision into the codec, narrowing the understanding gap. Crucially, all of these systems output a single token group whose levels are jointly trained for reconstruction, so the tokens must retain prosody and speaker detail in addition to language content. FactorCodec departs from both families by producing two functionally separated token groups with distinct training objectives, where analysis tokens are not invertible to waveforms (Section 3).

Unified audio foundation models. Recent systems aim to handle understanding and generation in a single model. AudioPaLM [Rubenstein et al., 2023], SpeechGPT [Zhang et al., 2023a], SpiRit-LM [Nguyen et al., 2025], GLM-4-Voice [Zeng et al., 2024], Llama-Omni [Fang et al., 2024], Kimi-Audio [KimiTeam et al., 2025], MiMo-Audio [Zhang et al., 2025], and Qwen2.5-Omni [Xu et al., 2025a] occupy different points in the design space: some are purely speech-focused, some hybrid (continuous input + discrete output, e.g., Kimi-Audio). UniAudio 2.0 differs in three ways: (i) the interface is purely discrete on both sides; (ii) the discrete interface is itself *factorized* into role-separated streams; and (iii) the backbone uses functional layer specialization rather than a uniform transformer.

Information bottleneck and representation learning. The information-theoretic argument behind FactorCodec (Appendix A.3.5) is grounded in the information-bottleneck (IB) principle [Tishby et al., 2000], which formulates representation learning as a trade-off between sufficiency for the task variable and minimality with respect to the input. Our analysis branch can be viewed as an instance of this principle: text-understanding objectives identify the task variable, and the lack of a waveform-decoder path enforces compression away from acoustic detail.

Table 30: ASR performance on LibriSpeech and Seed-TTS benchmarks. Note that we do not find a useful ASR instruction for Step-Audio-chat, thus we directly use the official reported results on Librispeech benchmark.

Model	LibriSpeech-clean	Libri-other	Seed-TTS ZH	Seed-TTS EN
Mimo-Audio-Instruct-7B Zhang et al. [2025]	3.50	35.43	7.0	29.8
Qwen2.5-Omni-7B Xu et al. [2025a]	3.92	5.52	1.3	2.89
Step-Audio-chat-3B Huang et al. [2025]	3.11	8.44	–	–
Whisper-large-v3 Radford et al. [2023]	1.81	3.55	6.8	1.47
UniAudio 2.0 (Ours)	2.71	6.33	2.6	2.14

Table 31: ASR results on WenetSpeech-Yue ASR Benchmark.

Model	Long sentence	Short sentence
Whisper-large-v3 Radford et al. [2023]	36.8	31.5
SenseVoice-Yue Li et al. [2025b]	15.7	6.0
Qwen2.5-Omni-7B Xu et al. [2025a]	23.5	31.0
UniAudio 2.0 (Ours)	12.1	7.7

Chain-of-thought and explicit planning. The role analogy between analysis tokens and chain-of-thought [Wei et al., 2022] in text LLMs is structural, not literal: in both cases, the model first emits an explicit, decodable representation of intermediate structure before committing to fine-grained outputs, and conditioning the output on this intermediate representation lowers per-step entropy. Similar idea also be used in audio domain, such as MusicCoT [Lam et al., 2025], Omni-R1 [Rouditchenko et al., 2025] . Unlike text CoT, our analysis tokens are not produced as natural-language sequences but as discrete codes that admit natural-language decoding via the analysis branch's LLM head, providing the same interpretability benefit at a fraction of the sequence length.

D Reproducibility Details

Architectural details, training recipes, and evaluation protocols are documented in Appendices A and B. All training and evaluation data are public datasets, cited throughout. The source code for inference, evaluation, and the audio tokenizers is available under the MIT License at `https://github.com/yangdongchao/UniAudio2`; pretrained checkpoints are available at `https://huggingface.co/Dongchao/UniAudio2_ckpt`, and audio demos are available at `https://dongchaoyang.top/UniAudio2Demo/`. The complete pre-training code will be released with the camera-ready materials.

E Limitation

While UniAudio 2.0 demonstrates strong performance across a wide range of audio understanding and generation tasks, several limitations remain.

(1) Decoding efficiency. Generation is currently offline: the model predicts the complete analysis-token segment before the reconstruction-token segment, and waveform decoding follows token generation. The flow-based decoder also introduces multi-step inference and increases decoding latency. We do not claim streaming-causal waveform generation; chunk-wise generation would require additional causal or chunk-causal components and training. Further improving decoding efficiency (e.g., via few-step or distilled decoding) remains an important direction. Refer to Appendix A.3.10.

(2) Generalization to unseen tasks. Although UniAudio 2.0 shows promising few-shot and zero-shot generalization, its capability is still constrained by the diversity and coverage of training data. In particular, current zero-shot performance primarily reflects *compositional generalization*, where new tasks are solved by recombining or adapting skills learned during training. For example, speech-sound generation can be viewed as a composition of TTS and sound generation. Tasks that require fundamentally new capabilities (e.g., speaker diarization with explicit temporal segmentation) are not

Table 32: TTS results on Seed-TTS (EN/ZH), LibriSpeech-clean, and Cantonese TTS. We report the WER and DNS-MOS score (WER / DNS-MOS). The CosyVoice-Yue is the official checkpoint from WenetSpeech-Yue Li et al. [2025b].

Model	Seed-TTS-EN	Seed-TTS-ZH	LibriSpeech-clean	WenetSpeech-Yue
MiMo-Audio-7B-Instruct	5.37 / 3.27	1.93 / 3.34	4.74 / 3.35	-
Qwen2.5-Omni-7B	**3.10** / 3.62	**1.21** / 3.65	4.28 / 3.72	**12.2** / 3.36
CosyVoice-Yue	-	-	-	13.9 / 3.34
UniAudio 2.0 (Ours)	3.63 / **3.80**	2.30 / **3.82**	**3.46** / **3.88**	12.5 / **3.41**

Table 33: Comparison with hybrid continuous-discrete models Kimi-Audio and Gemini 2.5-flash on overlapping tasks. WER↓ on LibriSpeech-test-clean. Kimi-Audio uses a hybrid 12.5 Hz continuous Whisper encoder + 12.5 Hz discrete codec; UniAudio 2.0 is purely discrete.

Model	TTS WER ↓	ASR WER ↓	Lyric WER ↓
Kimi-Audio (7B, hybrid)	**2.17**	**1.49**	249.3 (fails)
Gemini 2.5-flash	**2.17**	2.56	**10.35**
UniAudio 2.0 (3B, discrete)	3.46	2.7	28.57

yet well supported. Expanding the diversity of training tasks and supervision remains an important direction for improving generalization.

(3) Model scaling. Due to computational constraints, we only study 1B and 3B model variants. The scaling behavior of UniAudio 2.0 to larger models (e.g., 7B and beyond) is not fully explored, and remains an important direction for future work.

(4) Data imbalance across domains. UniAudio 2.0 currently performs better on speech than on sound and music, partly due to the larger availability of speech data. Improving the scale and diversity of sound and music datasets is expected to further enhance performance.

(5) Limited post-training exploration. This work focuses primarily on pre-training design (e.g., tokenizer and unified architecture). We do not extensively explore post-training strategies such as better multi-task SFT to balance the performance of each task or reinforcement learning to improve the instruction following ability, which could further improve performance.

(6) Coverage of comparisons. Comparisons are not exhaustive due to the broad task scope and limited availability of some models; different architectures may offer complementary strengths.

Table 34: Ablation studies on multi-stage training, layer specialization, and model size. FS-Sound denotes the few-shot sound classification tasks. A-T-TTS denotes the Audio prompt-caption guided TTS. N/A denotes the model cannot get meaningful results.

Setting	InstructTTS (WER/ Style-Acc / UTMOS)	FS-Sound (ACC)	A-T-TTS (SIM / Style-Acc / WER / UTMOS)	Song Rec (WER)	Song Gen (WER/CE/PQ)
UniAudio 2.0 (Ours 3B)	7.3/42.3/3.38	59.8	0.89 / 32.62 / 11.57 / 2.87	28.57	36.5 / 6.87 / 7.62
w/o Stage 4	8.8/35.2/3.41	N/A	N/A	29.7	37.4 / 6.8 / 7.6
& w/o Experts	9.7/22.9/3.21	N/A	N/A	34.3	42.5 / 6.68 / 7.55
1B	7.9/31.1/3.16	48.7	0.81 / 29.5 / 13.4 / 2.68	32.8	40.3 / 6.42 / 7.04

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state three contributions (FactorCodec, UniAudio 2.0 architecture with functional layer specialization, and extensive evaluation). All claims are supported by experiments in Section 5 and the appendix. Limitations are discussed in Appendix E.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E discusses six limitations, including flow-decoder latency, limited scaling exploration (only 1B and 3B), speech-dominant data imbalance, and the absence of post-training strategies.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The information-theoretic argument (conditioning on analysis tokens reduces reconstruction-token entropy) is presented in Section A.3.5 with the standard entropy chain rule. The argument relies on standard information-theoretic identities and does not require additional assumptions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Inference, evaluation, and tokenizer code are publicly available under the MIT License at `https://github.com/yangdongchao/UniAudio2`; pretrained checkpoints are linked in the reproducibility details. The complete pre-training code will be released with the camera-ready materials. Training configurations and data sources are documented in the appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository provides inference, evaluation, and tokenizer code, and pretrained checkpoints are available. The complete pre-training code will be released with the camera-ready materials. Training configurations, dataset sources, and evaluation details are documented in the appendix.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Training details including learning rates, batch sizes, number of GPUs, and training stages are provided in Appendix B.4 and Table 25. GRPO hyperparameters are listed in Table 9. Evaluation data, benchmarks, and metrics are described in Appendix B.5.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the MUSHRA subjective evaluation, we report mean $\pm$ 95% confidence intervals (Table 15).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: FactorCodec is trained on 8 NVIDIA A100 GPUs (Appendix A.2). UniAudio 2.0 is trained on 64 NVIDIA H100 GPUs (Appendix B.4).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The Impact Statement discusses potential misuse of audio generation, including voice imitation and copyright infringement. It recommends consent, respect for applicable rights, and disclosure of synthetic audio; this work does not evaluate technical safeguards.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Impact Statement discusses potential benefits in audio accessibility, speech interfaces, and creative production, as well as risks of voice imitation, misleading audio, and copyright or consent violations.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not describe technical safeguards that have been implemented and evaluated for responsible model release. The Impact Statement provides deployment recommendations, but these are not safeguards put in place by this work.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and pre-trained models used are cited with their original publications. The training data sources are listed in Appendix B.2 with full citations. We use publicly available datasets (LibriSpeech, AudioSet, MLS, etc.) and pre-trained models (Whisper, WavLM, LLaMA) in accordance with their respective licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces FactorCodec and UniAudio 2.0. Architecture details, training configurations, and evaluation protocols are documented in the main text and appendix. Inference, evaluation, and tokenizer code are available in the public MIT-licensed repository; pretrained checkpoints are released, and the complete pre-training code will be released with the camera-ready materials.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The MUSHRA subjective evaluation (Table 15) was conducted with 15 trained audio professionals who scored 20 samples per domain (speech, sound, music) on a 0–100 scale using the MUSHRA-without-hidden-anchor protocol. The evaluation details are described in the experiments section. Participants were compensated at standard professional rates.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The human evaluation involves only standard audio listening tests (MUSHRA) by trained professionals. This type of evaluation poses no risk to participants and does not typically require IRB approval.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are integral to our method in three ways: (1) LLaMA 3.2 3B serves as the cross-modal expert backbone of UniAudio 2.0 and as the decoder head for the analysis branch of FactorCodec; (2) LLaMA 3.1-Instruct 8B is used as the reward judge during GRPO training of the analysis branch; (3) Qwen3-Omni and Gemini 2.5 Pro were used to generate the mixture of audio reasoning data for FactorCodec training (Appendix A.2). All three usages are described in the method sections.