Bagpiper: Solving Open-Ended Audio Tasks via Rich Captions

Jinchuan Tian[*], Haoran Wang[*], Bo-Hao Su[*], Chien-yu Huang[*], Qingzheng Wang[*], Jiatong Shi[1], William Chen[1], Xun Gong[1], Siddhant Arora[1], Chin-Jou Li[1], Masao Someki[1], Takashi Maekaku[2], Keita Goto[2], Yusuke Shinohara[2], Jin Sakuma[2], Chao-Han Huck Yang[3], Shinji Watanabe[1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]LY Corporation [3]NVIDIA
jinchuat@andrew.cmu.edu

Abstract

Current audio foundation models typically rely on rigid, task-specific supervision (*e.g.*, speech recognition), addressing isolated factors of audio rather than the whole. In contrast, human processes audio holistically, seamlessly bridging raw audio waveform with abstract cognitive concepts (*e.g.*, all perception details of audio events) to execute complex tasks. Grounded in this philosophy, we introduce `Bagpiper`, an 8B audio foundation model that interprets physical audio via *rich captions*, *i.e.*, comprehensive natural language descriptions that encapsulate the critical cognitive concepts inherent in the audio. By pre-training on a massive corpus of 600B tokens, the model establishes a robust bidirectional mapping between raw audio and this high-level conceptual space. During fine-tuning, `Bagpiper` adopts a *caption-then-process* workflow, simulating an intermediate cognitive reasoning step to solve diverse tasks without knowing prior task-specific practice. Experimentally, `Bagpiper` achieves universal generation that can uniformly generate speech, sound effects, music, and their arbitrary combinations. It also maintains comparable performance with the 7B Qwen-2.5-Omni for audio understanding. To the best of our knowledge, `Bagpiper` is among the first works that achieve open-ended audio understanding and generation on speech, sound, and music. Model, data, and code will be released at `Bagpiper` Home Page.

1 Introduction

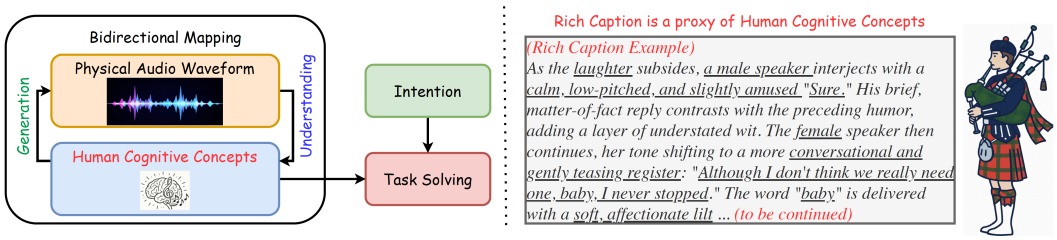

Figure 1: (Left) When solving an audio task, `Bagpiper` always materializes human cognitive concepts for the real physical audio waveforms. It produces a subsequent text answer (understanding) or audio clip (generation) based on those cognitive concepts. (Right) In this work, we take rich caption as the proxy of human cognitive concepts.

[*]Equal contribution.

Current audio foundation models are suboptimal for solving open-ended user requests. These models mainly rely on massive multi-tasking (Goel et al., 2025; Yang et al., 2023; Wang et al., 2024; Abouelenin et al., 2025; Valle et al., 2025; Radford et al., 2023; Peng et al., 2024; Maiti et al., 2024), where each task only addresses an isolated aspect of the audio. Additionally, the audio modeling (especially the generation modeling) (Du et al., 2025; Hung et al., 2024; Mousavi et al., 2025; Shi et al., 2024) often has a clear split of speech, music, and sound effects. The lack of a holistic concept at the task- and category-level creates a fundamental bottleneck for scalability: addressing the wide spectrum of audio tasks via exhaustive, domain-specific engineering is unsustainable. More importantly, it generalizes poorly to the user requests that are naturally compositional, flexible, and are thus not well defined. Thus, we argue for a paradigm shift toward a universal modeling philosophy that unifies all audio tasks and types within a single modeling paradigm.

Human intelligence perceives (understands) and produces (generates) audio in a holistic, open-ended, and multi-perspective manner (Galantucci et al., 2006), seamlessly bridging distinct categories like speech, music, and sound effects (Bregman, 1990; Patel, 2008). We propose that audio in this context is fundamentally twofold, defined by both the physical audio signals and the cognitive concepts they evoke in the human mind, such as text transcription, emotion, prosody, sound events, and music genres. Human intelligence processes audio by bridging these two planes: it converts audio signals into cognitive concepts to understand the world and, conversely, formulates responses derived from these concepts (O'Callaghan, 2009). In this work, we identify the *rich caption* (Ma et al., 2025b; Tang et al., 2025; Fiscus et al., 2006; Lu et al., 2026) as an ideal proxy for this wide range of human cognitive concepts. As illustrated in Figure 1, rich captions provide lengthy, comprehensive natural language descriptions that cover the critical cognitive concepts inherent in an audio signal, offering sufficient information to support diverse downstream open-ended task solving. Guided by this philosophy, the core objective of our audio foundation model, `Bagpiper`, is to uniformly solve audio tasks by establishing and leveraging a broad, holistic, and robust bidirectional mapping between physical audio signals and these cognitive concepts represented by rich captions. In implementation, we build this mapping during pre-training using a massive, curated dataset of 600B tokens, comprising audio-rich caption pairs that span speech, music, and sound effects. We then proceed to a fine-tuning stage where we simulate diverse open-ended request-response pairs with thinking traces, teaching the model how to infer appropriate answers from predicted rich captions (understanding) and how to translate user requests into rich captions as blueprints before physically producing audio (generation). Throughout this development pipeline, we incorporate no prior knowledge of specific audio tasks or types; nonetheless, the model demonstrates the capability to solve a diverse array of open-ended audio tasks.

Experimentally, we first validate the efficacy of our pre-training `Bagpiper`-Base, demonstrating that physical audio and rich captions can be translated bidirectionally with high fidelity across a suite of probing tasks. Regarding generative tasks, `Bagpiper` significantly outperforms prior state-of-the-art text-to-audio models, TangoFlux (Hung et al., 2024) and AudioLDM2-Large (Liu et al., 2024), in an extensive A/B testing on complex generation prompts. Beyond these quantitative metrics, we observe emergent model behaviors that transcend the scope of current evaluation protocols; we invite readers to explore these qualitative examples on `Bagpiper` Home Page. In terms of open-ended understanding, our `Bagpiper` achieves comparable capabilities with its 7B competitor Qwen-2.5-Omni (Xu et al., 2025) and AudioFlamingo3 (Goel et al., 2025), with competitive results on both AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2025). To facilitate reproducibility and future research, we commit to releasing our code, data, and trained models.

2 Bagpiper

2.1 Architecture

Like Tian et al. (2025b), our `Bagpiper` is designed to process any interleaved audio and text sequences within a unified framework. As shown in Figure 2, the backbone is initialized with the decoder-only LLM, i.e., `Qwen3-8B-Base` (Yang et al., 2025). We adopt the

established *Encoder-Adaptor-LLM* architecture (Liu et al., 2023; Goel et al., 2025) for audio input, which connects a pre-trained acoustic encoder to the LLM backbone via an MLP layer. For audio output, the model auto-regressively predicts multi-stream audio codec tokens, which are then detokenized into audio waveforms.[1] This design enables the consistent and seamless processing of multimodal inputs and outputs. For all training stages, we apply standard next-token prediction loss on target tokens only. During inference, we have separate inference configurations for text and audio generation. We additionally apply classifier-free guidance (Ho & Salimans, 2022; Hussain et al., 2025) for all audio generation tasks. No noticeable performance change is observed by varying the inference configurations. Architecture is not our major novelty. Detailed configurations are in the tables 9 and 10.

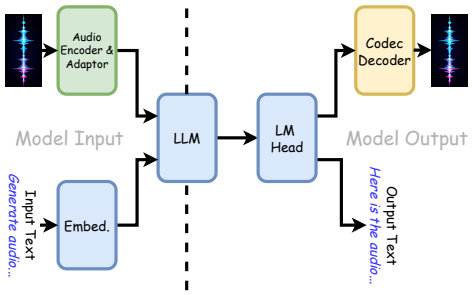

Figure 2: `Bagpiper` architecture

2.2 Pre-Training

The central premise of `Bagpiper` pre-training is that physical audio signals and cognitive concepts (i.e., rich captions) can be mutually translated with high fidelity. To realize this, our pre-training is to construct a robust bidirectional mapping between audio signals and rich captions via massive-scale learning. Additionally, it also requires preserving the text processing capabilities of the underlying LLM to support downstream complex task solving.

2.2.1 Data Curation

Collection and Labeling. To ensure universal audio processing capabilities, we aggregate a massive collection of publicly available audio data covering speech, music, and sound effects (full list in Appendix E.1). Notably, this mixture features diverse in-the-wild distributions, naturally encompassing complex compositions such as speech with background noise or musical accompaniment. We segment all audio into clips of up to 30 seconds. As in Fig.3, we generate a rich caption for every clip using the `Qwen3-Omni-30B-A3B-Captioner` (Qwen Team, 2025). This process yields a total of 422M raw audio-caption pairs.

Audio Taxonomy Classification. As distinct audio categories (speech, music, and sound effects) exhibit divergent quality distributions and severe data imbalances, a stratified processing approach is used to keep a rough balance in the curated pre-training mixture.[2] While audio classification directly from waveforms has been studied extensively (Gemmeke et al., 2017; Kong et al., 2020; Gong et al., 2021; Chen et al., 2023b), categorizing audio via rich captions remains underexplored, and we find it empirically effective. We explicitly categorize all audio clips into speech, music, or sound effects by prompting `Qwen3-32B` to analyze their rich descriptions. This text-based strategy proves highly effective, achieving a 93% accuracy on the MMAU dataset (Sakshi et al., 2024) when benchmarked against the original ground-truth category labels.

Stratified Data Filtering and Sampling. For each audio category, we score the audio-text pairs along three dimensions: (1) Audio-Only: We use UTMOS (Saeki et al., 2022; Shi et al., 2025) for speech and audiobox-aesthetics (Tjandra et al., 2025) for music and sound effects. (2) Text-Only: We apply heuristic filters combined with an LLM-as-a-judge approach (Zheng et al., 2023), evaluating three specific rubrics (Peng et al., 2025a). (3) Audio-text alignment: We compute the semantic similarity between the audio and caption using a CLAP-based model[3] (Wu* et al., 2023). To unify these diverse metrics, we compute the percentile rank of

[1]The audio encoder and codec model are adopted from `Qwen3-Omni-30B-A3B` (Qwen Team, 2025) and `X-Codec` (Ye et al., 2025), respectively. Each audio is represented by 8 tokens, and the audio token sequences are delay-interleaved Copet et al. (2024).

[2]E.g., we observe the average aesthetic score of *music* is much higher than *sound effects* on average.

[3]Due to the context length limit of the CLAP model, we summarize them into short captions.

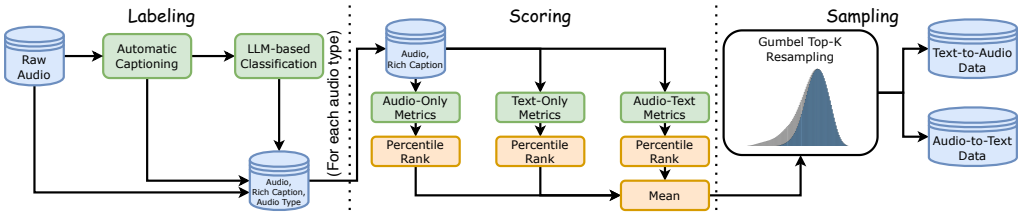

Figure 3: Pre-training data labeling and filtering workflow.

each sample for every dimension within its specific category and average these percentiles to form a final quality score. Next, to maintain data diversity while removing low-quality outliers, we employ a Gumbel Top-k resampling trick (Kool et al., 2019) to perform a soft truncation of the tail distribution. Finally, we perform the text-based de-duplication by MinHash (Broder, 1997) using a setup as in Li et al. (2024). We observed that the generated rich captions possessed high natural diversity, resulting in the removal of only $\sim$1M pairs during de-duplication. Further details are provided in Appendix E.2.

Text-Only Corpus. To preserve the text capability of the base LLM, we incorporate high-quality text-only corpora into the pre-training mixture. We specifically utilize datasets from Bercovich et al. (2025) and the mid-training data of Olmo et al. (2025), totaling 150B tokens. These datasets are yielded from harsh curation in prior works, so we skip further curation on text-only data.

Training Budget and Scheduling. Like in (Maiti et al., 2024; Olmo et al., 2025), we first allocate a total training budget of 600B tokens.[4] Following empirical observations from Tian et al. (2025b), we adopt a data ratio of 300B (Text-to-Audio) : 150B (Audio-to-Text) : 150B (Text-only). This distribution serves two purposes: first, Tian et al. (2025b) finds that audio generation (Text-to-Audio) typically requires a longer training horizon to converge than understanding; second, it perfectly balances the output modalities, ensuring a 1:1 ratio between text-output and audio-output supervision. Within the text-to-audio and audio-to-text tasks, we evenly split the budget across speech, music, and sound effects, which effectively up-samples the music and sound effects data by a factor of 3–4$\times$ to counteract the natural scarcity of these domains compared to speech.

2.2.2 *Training Strategy*

We implement a two-stage training protocol following Wu et al. (2025). Stage 1 begins with an adapter and embedding warm-up stage (2k steps), where the audio encoder and LLM backbone are frozen; only the embeddings and the adaptor are optimized. During this stage, we employ a substantially larger batch size and learning rate compared to the subsequent stage to rapidly align the modalities. Following this, stage 2 launches the full pre-training stage. Here, we unfreeze the relevant parameters and apply a linear learning rate warm-up followed by cosine decay. The loss of audio tokens is observed to be much higher than that of text tokens on average. Consequently, we use sequence packing (Krell et al., 2021) that carefully fuses samples from both sides to reduce the training loss fluctuation. Detailed training hyperparameters are provided in table 9. We refer to the resulting model as `Bagpiper`-Base.

2.3 Supervised Fine-Tuning (SFT)

In this SFT stage, we evolve `Bagpiper`-Base into a generalist problem solver. Consistent with our goal of overcoming the limitations of isolated supervision, we deliberately eschew traditional categorical labels during data simulation. Instead, we condition Bagpiper to treat rich captions as a universal semantic interface. By grounding the model's reasoning process entirely in these natural language descriptions, we enable it to handle the boundless variety of open-ended tasks described in our introduction.

[4] A token represents either a text subword or an audio frame.

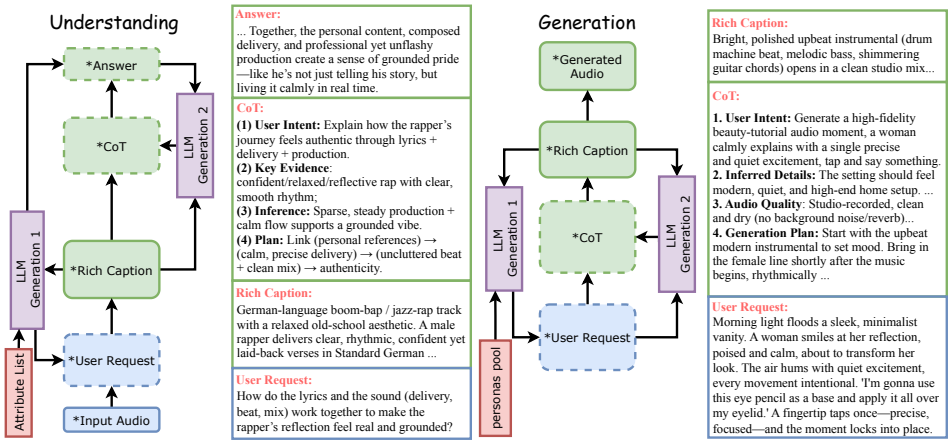

Figure 4: SFT data simulation pipeline §2.3.2 for open-ended audio task solving. Blue means model input; green means model output. Dashed boxes are simulated by prompting LLMs. Components with * are included in training sequences.

2.3.1 *Data Collection and Gemini Captioning*

We curate diverse and high-quality audio-caption pairs for both audio understanding and generation. For audio understanding, we sample ≈271k examples from a series of open datasets. We use our manual prompt to query `Gemini-3-Flash/Gemini-2.5-Pro` to obtain the rich captions, which are JSON-structured and more suitable for downstream data simulation (detailed distributions can be found in Appendix E.1). For audio generation, we continue our strategy of Gumbel Top-k sampling, but with a much lower temperature to select the highest quality data in our pre-training. We end up with 1M samples. Our audio generation data relies on the pre-training captions rather than `Gemini`-generated ones, preventing data leakage as `Gemini` is later employed as the evaluator in §3.3.2.

2.3.2 *Fine-Tuning Data Construction*

Pipeline: Caption-then-Process. Our data simulation always starts from audio-caption pairs, and uniformly enforces an explicit "thinking" pattern. For **audio understanding**, we adopt the template: *[Audio, User Request, Rich Caption, CoT, Answer]*. As shown on the left of Figure 4, we prompt a strong text LLM to first synthesize a valid *(User Request, Answer)* pair given the ground-truth rich caption, and subsequently generate a *Chain-of-Thought (CoT)* trace (Wei et al., 2022) that reasons from cognitive concepts (i.e., rich captions) to the expected answer. For **audio generation**, we adopt the template: *[User Request, CoT, Rich Caption, Audio]*. As shown in the right of Figure 4, the simulation process is similar: the LLM generates a creative *User Request* based on the rich caption, followed by a *CoT* trace that plans the audio content before outputting the caption.

Diversification. To cover more diversity of user requests, we employ targeted prompting strategies using strong open-source text LLMs.[5] For **understanding**, we sample from an attribute list of 108 attributes[6] to construct complex, multi-faceted inquiries. The LLMs are instructed to generate complex but reasonable requests with self-selected attributes. Our simulation also includes the cases where multiple related attributes are queried together to form compositional requests (Acuna et al., 2025). For **generation**, we adopt a "Persona-based" prompting strategy, selecting from 12 distinct user personas (Ge et al., 2024) to generate varying generation requests. Besides the user requests that explicitly mention the exact audio events and details, we also generate the imaginary user requests that describe a

[5]Specifically `Qwen3-235B-A22B-Instruct-2507-FP8` and `GPT-OSS-120B`.

[6]We randomly sample 10k pre-training captions, ask `Gemini-3-Pro` to analyze them, and obtain this attribute list (App. E.10).

feeling, an atmosphere, or a scene, which potentially boost the model's creativity in audio generation (Tian et al., 2025b).

Automatic Quality Assurance. To ensure high data quality, we implement a rigorous filtering pipeline utilizing the LLM-as-judge framework (Zheng et al., 2023) for descriptive quality estimation (Chen et al., 2023a). We employ `Qwen3-235B-A22B-Instruct-2507-FP8` to evaluate each sample on a 5-point scale, assessing five key dimensions: user request diversity, request-response alignment, thinking trace coherence, rich caption quality, and overall training value. By retaining only those samples with an average score exceeding 3, we distill our simulated data into a final set of 845k understanding samples and 1.47M generation samples. We validate this cutoff in App. E.7. Note that multiple requests will be generated for each audio-caption pair before this filtering.[7]

3 Experiments

3.1 Pre-Training Evaluation Setup

The `Bagpiper`-Base is not for real task-solving. Instead, our evaluation design for this pre-trained model is to quantify how much critical cognitive concepts are preserved in the rich captions that can support downstream task solving. We mainly rely on LibriSpeech Test-Clean (Panayotov et al., 2015) for local, phonetic-level capability and MMAU-Mini (Sakshi et al., 2024) for other audio-related capability in general, which covers speech, music, and sound effects.

Audio Understanding (Information Retention).

Following Ma et al. (2025b), to quantify how much cognitive concepts are preserved for downstream understanding, we feed the generated caption into `Gemini-3-Flash` and request it to perform audio tasks solely based on the text, without access to the original audio. We report the Word Error Rate (WER) on LibriSpeech Test-Clean and question-answering accuracy on MMAU-Mini. High performance here indicates that the captions generated from audio successfully encapsulate the essential semantics of the audio clips for understanding.

Audio Generation (Reconstruction Fidelity). We evaluate whether the pre-trained model can physically produce the audio waveform with the audio caption generated by the original captioner (Qwen Team, 2025). We generate audio conditioned solely on the rich captions and compare the output to the ground truth audio. We report WER on LibriSpeech Test-Clean to assess phonetic preservation.[8] For general audio, we employ a model-based metric on MMAU-Mini: We use Shi et al. (2025) to compute the FAD (Kilgour et al., 2018) and audio-audio CLAP similarity Wu* et al. (2023) between the reference and generated audio, and average the number across all three categories.

Cycle Consistency. Given the unified understanding-generation nature of our model, isolated evaluation is often insufficient. We therefore follow Hori et al. (2019) to conduct a cycle consistency test to verify the robustness of the bidirectional mapping, which is similar to evaluating an auto-encoder. We perform consecutive *Audio* $\rightarrow$ *Rich Caption* $\rightarrow$ *Audio* (reconstruction) and *Rich Caption* $\rightarrow$ *Audio* $\rightarrow$ *Rich Caption* (grounding) loops using `Bagpiper`-Base. We measure the cosine similarity between the input and the final output using `Qwen3-Embedding-8B` (Yang et al., 2025) for the text domain (*Text Sim.*) and the CLAP audio encoder (Wu* et al., 2023) for the audio domain (*Audio Sim.*). High similarity confirms minimal information loss through mutual translation between the physical audio signals and cognitive concepts.

Text Capability Preservation. To verify that incorporating audio objectives does not degrade language reasoning, we evaluate `Bagpiper`-Base on standard text benchmarks: MMLU-

[7] Within each of the understanding and generation tracks, the evaluation judge participated in neither caption generation nor SFT data simulation. See App. E.6 for details.

[8] We only test the examples whose rich captions contain the fully correct text transcriptions, which is around 60% of the volume of the whole test set.

Redux (world knowledge) (Gema et al., 2025), GPQA-Diamond (general reasoning) (Rein et al., 2024), and GSM8K (mathematical reasoning) (Cobbe et al., 2021).

3.2 Pre-Training Results and Analysis

In **audio understanding**, our objective is to match the performance of our teacher model. As shown in Table 1, Bagpiper-Base (8B) achieves performance parity with the Qwen3-Captioner (30B) topline (Qwen Team, 2025). This comparison is a distillation-parity sanity check rather than evidence that Bagpiper-Base exceeds its caption teacher. This result validates the effectiveness of our distillation efforts from the captioner, demonstrating that our 8B backbone successfully learns to extract comprehensive cognitive details comparable to much larger models. Also, we find that both models show considerable hallucination on the ASR task (5.5% and 5.0%). After manually checking, the hallucination is almost all from the captioner models, not the probing LLM. This is further discussed in table 5. Overall, these probing results conclude that physical audio signals can be effectively compressed into cognitive concepts by Bagpiper-Base.

Table 1: Understanding probing results for Bagpiper-Base. Both models are NOT for real task solving; the task is solved by an external LLM using the generated rich captions.

Model	Param.	WER (↓)	MMAU-Mini (↑)
Qwen3-Captioner	30B-A3B	5.5	**71.1**
Bagpiper-Base (ours)	8B	**5.0**	69.0

For **audio generation**, to the best of our knowledge, there are no rich caption-to-audio baselines to compare with. Thus, we compare our Bagpiper-Base with specialized state-of-the-art baselines in speech (TTS; OpusLM (Tian et al., 2025a) and CosyVoice3 (Du et al., 2025)) and sound/music (TTA; TangoFlux (Hung et al., 2024) and AudioLDM2-Large (Liu et al., 2024)), respectively. Note that these baselines are limited to their standard input format (transcripts for TTS; short captions for TTA), whereas Bagpiper-Base utilizes the full rich captions. Table 2 shows that Bagpiper-Base achieves WER comparable to dedicated TTS systems and outperforms TTA baselines in fidelity. This suggests that cognitive concepts in rich captions can be accurately grounded back into physical audio signals with Bagpiper-Base.

Table 2: Generation probing results. Our model is prompted by rich captions, while others are prompted by text transcriptions or short captions.

Model	Param.	Text-to-Speech WER (↓)	Text-to-Audio FAD (↓)	Text-to-Audio CLAP (↑)
OpusLM (TTS)	8B	4.0	-	-
CosyVoice3 (TTS)	0.5B	2.9	-	-
TangoFlux (TTA)	0.5B	-	5.15	0.50
AudioLDM2-L (TTA)	1.5B	-	3.79	0.48
Bagpiper-Base (ours)	8B	**1.8**	**2.98**	**0.55**

Subsequently, on the **cycle consistency** test (Table 3), we compare Bagpiper-Base against pipelined combinations of different understanding (Qwen Team, 2025; Goel et al., 2025) and generation models (Hung et al., 2024; Liu et al., 2024). Since baseline generation models cannot process the lengthy rich captions, we summarize captions only for these baselines to fit their context windows; Bagpiper-Base receives the full captions. Bagpiper-Base significantly outperforms these composite pipelines in both audio and text similarity, confirming that a unified model establishes a far more coherent bidirectional mapping than loosely coupled systems.

Table 3: Cycle consistency probing results

Audio-to-Text	Text-to-Audio	Audio-Sim. (↑)	Text-Sim. (↑)
Qwen3-Omni	AudioLDM2-L	0.445	0.465
Qwen3-Omni	TangoFlux	0.361	0.486
AudioFlamingo3	AudioLDM2-L	0.457	0.515
AudioFlamingo3	TangoFlux	0.471	0.570
Bagpiper-Base (ours)		**0.502**	**0.840**

As demonstrated in Table 4, Bagpiper-Base maintains close performance to its initialization checkpoint (Qwen3-8B-Base) across knowledge, general reasoning, and math benchmarks. Preserving this strong textual foundation is not merely for maintenance; it directly powers our

Table 4: Text capability probing results

	MMLU-Redux (↑)	GPQA-Diamond (↑)	GSM8K (↑)
Qwen3-8B-Base	**76.1**	39.3	**89.8**
Bagpiper-Base (ours)	74.3	**41.9**	88.2

Table 5: Audio understanding results after fine-tuning. †Qwen3-Omni-Instruct is a non-size-matched reference; ‡UALM results are paper-reported rather than re-run.

Model	Size	WER (↓)	MMAU-Mini (↑)	MMAU (↑)	MMAR (↑)	AIR-chat (↑)	AudioBench (↑)
Audio Flamingo 3	8B	1.6	73.3	**76.7**	55.2	**6.80**	**70.45**
Qwen2.5-Omni	7B	1.6	71.5	65.5	56.7	6.30	68.03
Qwen3-Omni-Instruct†	30B-A3B	**1.2**	**78.0**	75.2	**70.4**	–	–
MiMo-Audio	7B	3.7	73.7	66.0	–	–	–
Voxtral-Mini	3B	1.8	61.3	57.1	–	–	–
Voxtral-Small	24B	1.5	68.8	62.2	–	–	–
UALM‡	8B	–	–	74.1	55.2	–	–
Bagpiper (ours)	8B	2.5	74.5	73.1	57.0	6.57	70.39

open-ended task solving during downstream fine-tuning. By retaining the model's inherent logic and world knowledge, `Bagpiper` can leverage complex text-based reasoning (e.g., Chain-of-Thought) to decompose and solve flexible audio instructions within the text domain and reflect it back to the audio modality.

3.3 SFT Evaluation

Unlike prior probing tests in §3.1, this SFT evaluation is to validate the fine-tuned `Bagpiper`'s efficacy in addressing real audio tasks through a unified, open-ended interface.

3.3.1 Audio Understanding

We compare recent systems (Goel et al., 2025; Xu et al., 2025; Qwen Team, 2025; Xiaomi LLM-Core Team et al., 2025; Liu et al., 2025; Tian et al., 2025b) on LibriSpeech test-clean, MMAU-Mini and full MMAU (Sakshi et al., 2024), and MMAR (Ma et al., 2025a); the final two columns report AIR-Bench-chat (Yang et al., 2024) and five AudioBench QA subsets (Wang et al., 2025).[9] Table 5 shows that `Bagpiper` leads MMAU-Mini among 7–8B systems with complete results and scores above Qwen2.5-Omni, Audio Flamingo 3, and UALM on MMAR. On full MMAU it trails Audio Flamingo 3, the larger Qwen3-Omni-Instruct reference, and UALM. UALM's numbers are paper-reported: with no official checkpoint released, a protocol-matched comparison is left to follow-up. Notably, `Bagpiper` learns to solve ASR and open-ended problems through the shared caption-then-process objective, without task-specific heads or prior knowledge of the individual task formats. WER falls from 5.0 before SFT to 2.5 due to more calibrated supervision signals; App. E.8 scopes caption-error effects on MMAU-Mini speech. Experiments in App. E.9 demonstrate that our audio pre-training, CoT-format supervision, and end-to-end caption-then-process design benefit overall system performance; they also show that `Bagpiper` can be strengthened by scaling up to a 30B-A3B backbone.

3.3.2 Audio Generation

Table 6: Text-to-speech results after fine-tuning.

Model	WER
AnyGPT	24.9
SpeechGPT	42.8
CosyVoice3	2.9
Bagpiper (ours)	**2.7**

We examine our model on the classical text-to-speech (TTS) task on Librispeech Test-Clean. As our model is pre-trained for a general purpose, we prompt the model to generate with *a calm, neutral voice*, plus the text transcription. As in the table 6, our generalist `Bagpiper` outperforms the dedicated TTS model on the WER, even though it has never been tuned for this task. App. E.8 shows that the intermediate rich caption preserves the requested transcription almost exactly (0.2% WER), locating the residual errors in the caption-to-audio stage. We also list AnyGPT (Zhan et al., 2024) and SpeechGPT (Zhang et al., 2023), which pioneered unified speech and audio generation with language models; their intelligibility on this specialized metric is well below the dedicated systems',

[9] The five subsets are CN-College-Listen, DREAM-TTS, Public-SG-SpeechQA, WavCaps-QA, and AudioCaps-QA. Some evaluation clips overlap with the SFT pool despite test-split exclusion; see App. E.1 for details.

so the comparison contextualizes current unified-generation TTS quality rather than diminishing these early systems. To compare against specialized TTS systems on their own terrain, App. F.2 reports a TTS-specialized `Bagpiper` variant A/B-tested against `Qwen3-TTS`, `CosyVoice3`, and `VibeVoice` under both judges and both rubrics.

We subsequently evaluate the model's capability for general audio generation, specifically its ability to synthesize arbitrary combinations of diverse audio types. To assess performance on complex instructions, we query `Claude-Opus-4.8` (not used in model development) to generate detailed prompts for each audio sample in MMAU-Mini, yielding a total of 1,000 evaluation samples. We compare `Bagpiper` against state-of-the-art baselines, including TangoFlux (Hung et al., 2024) and AudioLDM2-Large (Liu et al., 2024). To ensure a fair comparison (standard text-to-audio models may struggle with lengthy instructions), we also evaluate the baselines using condensed 15-word summaries of the original prompts (denoted as *summary*). As a matched-input control, App. F.1 additionally gives `Bagpiper` the same 15-word summary and confirms it still wins in aggregate under both judges. We conduct a side-by-side preference evaluation using `Gemini-3-Pro` and `GPT-4o` as judges, with the two systems presented in randomized order to control for position bias. As shown in Fig. 5, `Bagpiper` wins in aggregate in every comparison under both judges. The advantage is most pronounced on the *speech* subset, where `Bagpiper` synthesizes intelligible speech integrated with complex sound events or music, a capability largely absent in prior text-to-audio models. Overall, these results indicate that `Bagpiper` can faithfully follow natural-language user prompts and generate complex, multi-faceted audio.

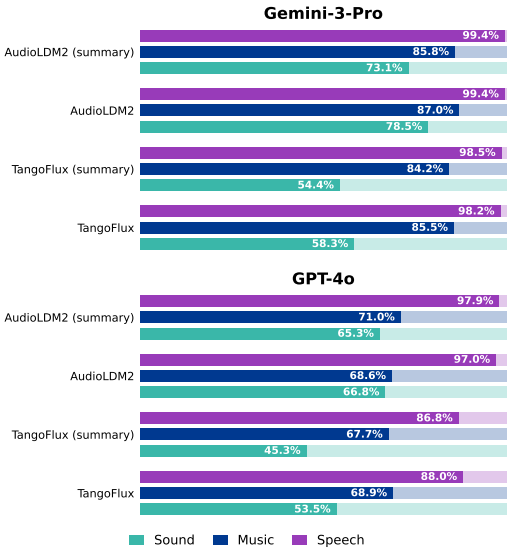

Figure 5: A/B win rates of `Bagpiper` against text-to-audio baselines under two judges, `Gemini-3-Pro` and `GPT-4o`, broken down by **Sound**, **Music**, and **Speech**. The darker region is `Bagpiper`'s win rate and the lighter region its loss.

3.4 Qualitative Observation

As an open-ended task solver, `Bagpiper` exhibits several attractive behaviors that extend beyond the scope of current evaluation protocols, particularly in the domain of audio generation. We highlight key observations below and refer readers to our `Bagpiper` Project Page for comprehensive demonstrations.

Compositional Audio Synthesis. As demonstrated in Fig.6, our model is capable of generating complex audio sequences that simultaneously incorporate speech, environment sounds, and music, within a single clip. Notably, this includes conversational features such as multi-speaker dialogue with sound effects.

Instruction Following. Leveraging its reasoning capabilities, the model processes instructions that extend beyond simple acoustic descriptions to include logic-based constraints. For instance, when prompted to *"sing 'Merry Christmas' twice"*, the model correctly produces the repetition: *"Merry Christmas, Merry Christmas"*.

World Knowledge Utilization. The model effectively leverages world knowledge derived from its textual pre-training to ground audio generation in specific contexts. For example, when prompted to *"Imagine walking through Disneyland"*, the model generates contextually appropriate audio, such as marching bands and fireworks, without needing explicit acoustic descriptions of those elements, with further discussion of these qualitative observations in App. F.3.

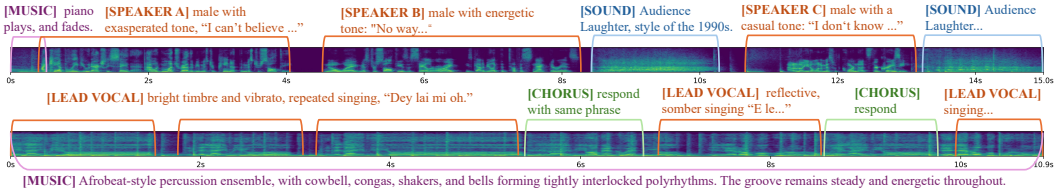

Figure 6: Spectrum of audio samples generated by `Bagpiper`. Our model is capable of generating multi-speaker dialogue with sound and music (upper) and singing voice with accompaniment (lower).

4 Related Work

We outline the closest lines below; App. B gives the comprehensive discussion.

Open-Ended Task Solving. Text foundation models attain broad generalization through massive multi-task supervision (Chung et al., 2024), but replicating this task scaling in audio is impractical, and existing audio language models still struggle with compositional instructions outside their training distribution. `Bagpiper` instead offloads open-ended reasoning to the text LLM through rich captions, needing no explicit audio-task supervision.

Rich Captions as a Modal Interface. Detailed captions have been shown to enhance individual downstream capabilities, but prior work remains task-specific, small-scale, and unidirectional. `Bagpiper` instead establishes a universal, bidirectional mapping between natural language and all audio categories.

Universal Audio Processing. While universal audio understanding has advanced, audio generation remains fragmented across models specialized for speech, sound, or music (Hung et al., 2024). `Bagpiper` synthesizes integrated acoustic scenes with arbitrary, overlapping speech, music, and sound events in a single clip.

5 Limitations

`Bagpiper` depends on an intermediate rich caption, so both understanding and generation inherit the caption model's errors, and residual ASR or captioner hallucinations can propagate downstream (App. E.8). Its attribute taxonomy and SFT data are LLM-agent-generated rather than expert-verified, and evaluation relies heavily on LLM-as-judge.

6 Conclusion

We presented `Bagpiper`, an 8B model that uses rich captions as a shared interface for audio understanding and generation across speech, sound, and music. Pre-training establishes a bidirectional mapping between audio signals and cognitive concepts expressed as rich captions, while caption-then-process SFT supports open-ended tasks. The reported results show competitive understanding and flexible mixed-audio generation, subject to the evaluation and contamination caveats above.

Acknowledgments

Parts of this work used the PSC Bridges2 system and Delta/DeltaAI system at NCSA through allocations CIS210014 and IRI120008P from the ACCESS program, supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

David Acuna, Chao-Han Huck Yang, Yuntian Deng, Jaehun Jung, Ximing Lu, Prithviraj Ammanabrolu, Hyunwoo Kim, Yuan-Hong Liao, and Yejin Choi. Long grounded thoughts: Distilling compositional visual reasoning chains at scale. *arXiv preprint arXiv:2511.05705*, 2025.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

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. *arXiv preprint arXiv:2209.03143*, 2022.

Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, 1990. doi: 10.7551/mitpress/1486.001.0001.

A.Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29, 1997. doi: 10.1109/SEQUEN.1997.666900.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and EngSiong Chng. Audio large language models can be descriptive speech quality evaluators. In *The Thirteenth International Conference on Learning Representations*, 2023a.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proc. International Conference on Machine Learning (ICML)*, pp. 5178–5193, 2023b.

Shunian Chen, Xinyuan Xie, Zheshu Chen, Liyan Zhao, Owen Lee, Zhan Su, Qilin Sun, and Benyou Wang. Fusionaudio-1.2 m: Towards fine-grained audio captioning with multimodal contextual fusion. *arXiv preprint arXiv:2506.01111*, 2025a.

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

Iwona Christop, Mateusz Czyżnikiewicz, Paweł Skórzewski, Łukasz Bondaruk, Jakub Kubiak, Marcin Lewandowski, and Marek Kubis. A benchmark for audio reasoning capabilities of multimodal large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 953–983, 2026.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets. *arXiv preprint arXiv:2503.04713*, 2025.

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.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 682–689. IEEE, 2024.

Jonathan Fiscus, Jerome Ajot, Martial Michel, and John Garofolo. The rich transcription 2006 spring meeting recognition evaluation. In *Proceedings of the Rich Transcription Spring Meeting Recognition Evaluation (RT-06S)*. NIST, May 2006.

Bruno Galantucci, Carol A Fowler, and Michael T Turvey. The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, 13(3):361–377, 2006.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, 2025.

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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, et al. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*, 2025.

Sang gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. ETTA: Elucidating the design space of text-to-audio models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Z9BaaWDE0r.

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

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proc. Interspeech*, 2021.

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.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, Ming Yang, and Jun Zhou. M2-omni: Advancing omni-MLLM for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Takaaki Hori, Ramon Astudillo, Tomoki Hayashi, Yu Zhang, Shinji Watanabe, and Jonathan Le Roux. Cycle-consistency training for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6271–6275. IEEE, 2019.

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

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. AudioGPT: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23802–23804, 2024.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Shehzeen Samarah Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Roy Fejgin, Mikyas T Desta, Rafael Valle, and Jason Li. Koel-tts: Enhancing llm based speech generation with preference alignment and classifier free guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21230–21245, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

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

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Mario Michael Krell, Matej Kosec, Sergio P Perez, and Andrew Fitzgibbon. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv preprint arXiv:2107.02027*, 2021.

Kumar et al. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025.

Max WY Lam, Yijin Xing, Weiya You, Jingcheng Wu, Zongyu Yin, Fuqiang Jiang, Hangyu Liu, Feng Liu, Xingda Li, Wei-Tsung Lu, et al. Analyzable chain-of-musical-thought prompting for high-fidelity music generation. *arXiv preprint arXiv:2503.19611*, 2025.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt,

Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024.

Alexander H. Liu et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, et al. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025a.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025b.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13326–13330. IEEE, 2024.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, et al. Discrete audio tokens: More than a survey! *arXiv preprint arXiv:2506.10274*, 2025.

Casey O'Callaghan. Auditory perception. *Stanford Encyclopedia of Philosophy*, 2009.

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.

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)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Aniruddh D. Patel. *Music, Language, and the Brain*. Oxford University Press, New York, 2008. ISBN 9780195123753.

Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. Dataman: Data manager for pre-training large language models. *arXiv preprint arXiv:2502.19363*, 2025a.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*, 2024.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, et al. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*, 2025b.

Qwen Team. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

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, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

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

Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, et al. Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. In *2024 IEEE Spoken language technology workshop (SLT)*, pp. 562–569. IEEE, 2024.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, et al. Versa: A versatile evaluation toolkit for speech, audio, and music. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 191–209, 2025.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. *arXiv preprint arXiv:2410.10676*, 2024.

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. *arXiv preprint arXiv:2310.13289*, 2023.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Jinchuan Tian, William Chen, Yifan Peng, Jiatong Shi, Siddhant Arora, Shikhar Bharadwaj, Takashi Maekaku, Yusuke Shinohara, Keita Goto, Xiang Yue, et al. Opuslm: A family of open unified speech language models. *arXiv preprint arXiv:2506.17611*, 2025a.

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

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lyu, Wei Xue, and Yike Guo. Audio-Omni: Extending multi-modal understanding to versatile audio generation and editing. *arXiv preprint arXiv:2604.10708*, 2026.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, Joao Felipe Santos, Shuqi Dai, Siddharth Gururani, Aya Aljafari, et al. Fugatto 1: Foundational generative audio transformer opus 1. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025.

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3355–3364, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 53366–53397, 2024.

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

Xiaomi LLM-Core Team et al. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. UniAudio 2.0: A unified audio language model with text-aligned factorized audio tokenization. *arXiv preprint arXiv:2602.04683*, 2026.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025.

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

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

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Xi Wang, Sheng Zhao, and Lei Xie. Cosyaudio: Improving audio generation with confidence scores and synthetic captions. *arXiv preprint arXiv:2501.16761*, 2025.

A Reproducibility Statement

To support reproduction, we will release Bagpiper-Base and the post-SFT checkpoint, the pre-training rich captions, and our training and evaluation recipes and scripts.

B Related Work (Extended)

Open-Ended Task Solving. The paradigm of open-ended task solving has matured significantly in the text domain (Hurst et al., 2024; Yang et al., 2025; Olmo et al., 2025; Grattafiori et al., 2024), where generalization is initially achieved through massive multi-task supervision. As the milestone, text foundational models like FLAN-T5 (Chung et al., 2024) enumerate over 1,800 distinct text tasks to facilitate zero-shot generalization on unseen user requests. However, replicating this "task scaling" strategy in the audio domain is prohibitively expensive and practically infeasible due to the massive human labor and scarcity of well-defined, diverse enough audio tasks. Consequently, existing audio language models that attempt naive multi-tasking (Abouelenin et al., 2025; Wang et al., 2024; Yang et al., 2023; Tang et al., 2023) still struggle with flexible, compositional instructions that fall outside their training distribution, as in §3.3.2. In contrast, Bagpiper bypasses the need for massive task enumeration. By establishing rich captions as a universal semantic proxy, we effectively offload the burden of open-ended reasoning to the text capabilities of the LLM, enabling the solution of audio tasks without explicit audio-domain task supervision.

Rich Captions as a Modal Interface. The utilization of automated, detailed captions has gained traction in recent research. Emerging works have demonstrated that descriptive captions can significantly enhance specific downstream capabilities, such as in music understanding (Ghosh et al., 2025), text-to-speech (Diwan et al., 2025), and text-to-audio/music generation (Zhu et al., 2025; Chen et al., 2025a; Sun et al., 2024), with the research scope narrowed to task-specific objectives, relatively small scale, and unidirectional. Our rich-caption interface also follows a line of contrastive and generative audio-captioning models (Wu* et al., 2023; Ma et al., 2025b; Goel et al., 2025). To our knowledge, Bagpiper is among the first frameworks to establish a universal mapping between natural language and all audio categories, and subsequently becomes a firm foundation for downstream open-ended processing.

Universal Audio Processing. While the pursuit of universal audio understanding has yielded significant progress (Abouelenin et al., 2025; Tang et al., 2023; Goel et al., 2025), the landscape of audio generation remains largely fragmented. Current state-of-the-art generation models are typically confined to specialized verticals, optimizing exclusively for speech (Peng et al., 2025b; Ju et al., 2024; Eskimez et al., 2024; Chen et al., 2025b), environmental sound (gil Lee et al., 2025; Hung et al., 2024; Huang et al., 2023), or music (Copet et al., 2024). Even hybrid domains like singing voice synthesis, which inherently blend vocals and accompaniment (Yuan et al., 2025; Lam et al., 2025), remain restricted to their specific modality. Although pioneering works like Yang et al. (2023) have attempted to cover the full auditory spectrum, they rely on rigid and pre-defined control tokens, extending a broader line of joint audio–text language models (Borsos et al., 2022; Rubenstein et al., 2023; Zhan et al., 2024; Wu et al., 2024; Guo et al., 2025). Among unified models, UALM (Tian et al., 2025b) is closest in objective, likewise unifying audio understanding and generation in one model, while agentic systems such as AudioGPT (Huang et al., 2024) reach comparable breadth by orchestrating external audio tools around an LLM. Concurrent efforts (Yang et al., 2026; Tian et al., 2026) pursue similar unification. In contrast, Bagpiper is capable of synthesizing integrated acoustic scenes containing arbitrary, overlapping combinations of speech, music, and sound events within a single clip. On the evaluation side, MMAU-Pro (Kumar et al., 2025) broadens audio-understanding assessment to spatial, multi-audio, and long-form inputs, and ART (Christop et al., 2026) is a complementary benchmark targeting audio-reasoning capabilities alongside AIR-Bench, AudioBench, and MMAU.

C Rich Caption Examples

Rich Caption Example for Music

The audio clip begins abruptly, launching into a high-energy, distorted electric guitar riff characteristic of heavy metal or hard rock, immediately accompanied by a powerful, driving drum beat. The guitar, panned hard left and treated with heavy distortion and reverb, delivers a rapid, repetitive motif that anchors the track's intensity. The drums, centered in the mix, feature a punchy kick and snare with a prominent crash cymbal, while the bass guitar sits low and centered, reinforcing the riff's rhythmic foundation. The overall sound is dense, loud, and saturated with analog-style distortion, suggesting a vintage or intentionally lo-fi production style.
At the three-second mark, a male vocalist-delivering a strained, aggressive, and raspy performance–shouts, "Sold my car!" in a non-rhotic accent reminiscent of British or Australian English. His voice, heavily processed with reverb and distortion, is panned slightly to the right and stands out against the thick instrumental backdrop. The shout is forceful and emotionally charged, with the final syllable ("car") trailing off and blending into the reverberant mix. No other vocals, harmonies, or spoken words are present.
Immediately following the vocal line, the instruments maintain their relentless energy, with the guitar and drums continuing their unyielding rhythm. The audio ends abruptly, cutting off mid-beat without any fade-out or resolution, leaving the listener with a sense of unresolved intensity.
This audio clip features a raw, high-energy rock/metal excerpt with a distorted guitar riff, driving drums, and a shouted male vocal line. The production is intentionally lo-fi, evoking the sound of 1980s underground or DIY recordings, and the lyrics "Sold my car!" are delivered with aggressive emotion. The clip is likely an excerpt from a demo or independent release, designed to convey urgency, rebellion, and personal upheaval, and is devoid of additional spoken content or environmental noise."

Rich Caption Example for Sound Effects

The audio begins with the immersive, high-fidelity sound of a steady stream of water cascading over rocks in a natural outdoor setting. The water's movement is characterized by a continuous, low-frequency gurgle interwoven with midrange bubbling and a lively, high-frequency hiss, all centered in the stereo field and closely miked. This creates a richly detailed and enveloping sonic texture, evoking the presence of a small stream or brook flowing through a rocky bed, with no extraneous environmental sounds or human activity. The recording remains uninterrupted, maintaining its clarity and presence throughout, until it is abruptly and completely cut off at the end with no residual noise or fade-out.
This audio presents a pure, uninterrupted sample of natural water sounds, captured with professional equipment in a pristine outdoor environment. The absence of speech, music, or any cultural or geographic markers, along with the sudden termination, suggests a utilitarian purpose–likely as a sound effect for media production, relaxation, or ambient use. The high-quality, immersive recording is designed to evoke tranquility and natural serenity, serving as a versatile and context-free sonic backdrop.

Rich Caption Example for Speech

The audio begins in a large, open urban environment, likely a city street or plaza in the Netherlands, characterized by a persistent low-frequency rumble from traffic and a faint, high-frequency hiss typical of a mobile recording device. Immediately, a sharp, metallic two-tone car horn sounds nearby, lasting about a second. The horn's clarity and proximity suggest the microphone is close to the source, and the lack of reverberation confirms an outdoor, non-enclosed setting.
As the horn fades, a female voice emerges in the background, speaking Dutch with a clear, neutral tone. She says, "Nee, maar ik ben..." ("No, but I am..."), her words partially obscured by the ongoing traffic noise. The male speaker, positioned closer to the microphone, responds succinctly with "Ja" ("Yes"), his voice calm and matter-of-fact. Both voices are slightly muffled and distant, indicating their location within the larger ambient soundscape.
The female speaker continues with a casual remark: "Laat maar even zien hoe het is dan." ("Just let me see how it is then."), delivered in a friendly, conversational manner. The male speaker replies in a relaxed, explanatory tone, "Ik vind het leuk om geluiden op te nemen." ("I like to record sounds."), directly addressing the recording's purpose. Their interaction is informal and natural, reflecting an everyday exchange in a Dutch urban context.
Throughout the clip, the background remains dominated by traffic, with no evidence of additional people, music, or environmental sounds. The conversation is unhurried and the speakers' voices are neither urgent nor emotional, indicating a casual, unscripted moment. The audio concludes abruptly, with the recording ending mid-phrase, suggesting intentional capture or editing.
In summary, this recording presents a slice of urban Dutch life: a brief, friendly exchange between two individuals, set against the ambient noise of city traffic. The male speaker explains his interest in capturing sounds, while the female speaker engages in casual conversation. The audio's technical quality, conversational flow, and environmental context reflect a spontaneous, everyday interaction in a public urban space, with no music or extraneous events present.

D Supervised Fine-Tuning (SFT) Examples

Figure 7: **SFT example (audio-to-text).** Representative audio understanding samples synthesized by our Caption-then-Process pipeline (*[Audio, User Request, Rich Caption, CoT, Answer]*).

Figure 8: **SFT example (text-to-audio).** Representative audio generation samples synthesized by our simulation pipeline (*[User Request, CoT, Rich Caption, Audio]*).

E Implementation Details

E.1 Data Collection

Warm-up and Pre-training Stage: we jointly train on a mixture of audio–text paired datasets and text-only datasets. The audio-text paired datasets include: *OWSM v4 captions*[10], *LAION-Audio-300M*[11], *Clotho-AQA*[12], *clotho*, *Emilia*[13], *LAION captioned AI music snippets*[14], *LAION in-the-wild sound events*[15], *AudioCaps*[16], *AudioSet*[17], *WavCaps*[18], *FMA*[19], *YouTube-8M*[20], *LAION-DISCO-12M*[21] *YODAS*[22] . For text-only data, we use a broad *Dolma3 ingredient1*[23] mixture (domain-balanced CommonCrawl high-quality subsets, plus code/math/reasoning/STEM-centric subsets) together with dialogue-style instruction corpora (*llama-nemotron*[24], *Olmo3*[25]).

Supervised Fine-tuning Stage: we supervise multi-turn user requests using dialogue-style instruction data that contains both audio-to-text understanding and text-to-audio generation samples (examples are shown in App. D).

The audio-to-text understanding datasets include: *OWSM v4 (200k subset)*, *Librispeech (train-clean-100)*, *Common Voice Corpus 15*, *IEMOCAP*, *MELD*, *SLUE (SQA, NEL, DAC)*, *VocCeleb1*, *Fake-or-Real*, *Free Music Archive (FMA)* , *MUSIC-AVQA*, *MusicCaps*, *NSynth*, *AudioCaps*, *AudioGrounding*, *AVQA*, *Clotho-AQA*, *CochlScene*, *2025 DCASE AudioQA* , and *SoundDescs* , *TUT-acoustic-scenes-2017, and VocalSound*. We construct the SFT training set by excluding all test splits from our evaluation benchmarks. Test-split exclusion does not prevent the same underlying audio from entering through another parent corpus. Using audfprint[26], we find that 12.8% of AudioBench and 3.2% of AIR-Bench evaluation clips overlap with the SFT pool. The current scores, particularly AudioBench, should therefore be interpreted with this caveat. The overall training split is formed by randomly sampling all valid samples. The resulting data composition is summarized in Table 7, showing broad coverage across speech, music, and general sound domains. The distribution of three domains is also presented in Figure 9.

To standardize input length, when a source corpus contains clips longer than 30 seconds, we extract a 30-second segment by truncating from a uniformly random start time within the clip. In contrast, for multimodal datasets such as AVQA, where audio is tightly aligned with video content, random truncation may break semantic coherence. We therefore discard videos longer than 30 seconds and retain only those shorter than 30 seconds long. After this filtering, 50 AVQA videos remain; applying our per-corpus random sampling then yields 7 AVQA samples in the final SFT dataset. Besides, in our SLUE dataset, we only include the training set in SQA, NEL and DAC splits, since all audios in the TED split are all longer than 30 seconds.

[10] https://huggingface.co/datasets/espnet/yodas_owsmv4
[11] https://huggingface.co/datasets/laion/LAION-Audio-300M
[12] https://huggingface.co/datasets/lmms-lab/ClothoAQA
[13] https://huggingface.co/datasets/amphion/Emilia-Dataset
[14] https://huggingface.co/datasets/laion/captioned-ai-music-snippets
[15] https://huggingface.co/datasets/laion/in-the-wild-sound-events
[16] https://huggingface.co/datasets/d0rj/audiocaps
[17] https://huggingface.co/datasets/agkphysics/AudioSet
[18] https://huggingface.co/datasets/cvssp/WavCaps
[19] https://huggingface.co/datasets/benjamin-paine/free-music-archive-full
[20] https://huggingface.co/datasets/leee99/yt8m-h264
[21] https://huggingface.co/datasets/laion/LAION-DISCO-12M
[22] https://huggingface.co/datasets/espnet/yodas
[23] https://huggingface.co/datasets/allenai/dolma3_pool
[24] https://huggingface.co/datasets/nvidia/Llama-Nemotron-Post-Training-Dataset
[25] https://huggingface.co/collections/allenai/olmo-3
[26] https://github.com/dpwe/audfprint

Table 7: Dataset distribution for the SFT stage captioned by Gemini_2.5_Pro and Gemini_3_Flash, grouped by domain. "–" indicates the dataset is not present for that model.

Domain	Dataset	Gemini_2.5_Pro	Gemini_3_Flash	All sample number
Speech	Common Voice Corpus 15	9996	23775	33771
	IEMOCAP	7869	–	7869
	Librispeech (train-clean-100)	10000	4368	14368
	MELD	9984	–	9984
	SLUE	–	12779	12779
	VoxCeleb1	10000	81144	91144
	Fake-or-Real	9994	–	9994
	Speech subtotal	**57843**	**122066**	**179909**
Music	FMA	1813	–	1813
	MUSIC-AVQA	50	–	50
	MusicCaps	2644	–	2644
	NSynth	9993	10045	20038
	Music subtotal	**14500**	**10045**	**24545**
Sounds	AudioCaps	10000	8793	18793
	AudioGrounding	3528	–	3528
	AVQA	–	7	7
	Clotho-AQA	1169	–	1169
	CochlScene	9993	–	9993
	2025_DCASE_AudioQA	9314	–	9314
	SoundDescs	5629	–	5629
	TUT-acoustic-scenes-2017	3508	–	3508
	VocalSound	9996	4482	14478
	Sounds subtotal	**53137**	**13282**	**66419**
Sum		**125480**	**145393**	**270873**

For text-to-audio generation data, we use a subset of all the pre-training data. Specifically, we use Gumbel Top-k sampling to sample 1M samples from the pre-training distribution, with the temperature of 0.03.

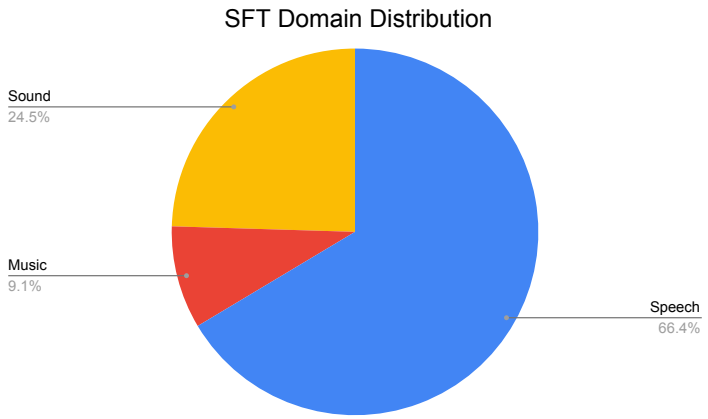

Figure 9: SFT domain distribution for speech, sound and music.

E.2 Data Curation

We provide additional details for data curation, as mentioned in §2.2.

Audio. For perceptual quality, we use UTMOS for speech and the AudioBox aesthetic score for non-speech audio.

text. We have rule-based filtering that excludes: (1) labeled incomplete in rich captioner captioning; (2) contain non-Latin characters; (3) character repetition for 5 times or above; (4) word repetition for 3 times or above; (5) 5-gram repetition for 10% and above; (6) unique word ratio for 30% or below; (7) average word length lower than 2 or larger than 15; (8) number of tokens larger than 800 or smaller than 200.

We additionally prompt the `Qwen3-32B` to: (1) find the audio type, a.k.a., speech, music, sound effects; (2) find if the text is purely in English, and exclude it if not; (3) find the text is centralized to audio content, and exclude it if not; (4) score the audio intelligibility, with a scale of 1–5; (5) score the audio complexity, with a scale of 1–5; (6) score the audio diversity, with a scale of 1–5;

Audio-Text Alignment. We use CLAP model to test the audio-text alignment. As the CLAP model would only accept text up to 77 tokens, we use `Qwen3-32B` to summarize it into short captions.

Gumbel Top-K sampling. For audio understanding, we use a temperature of 0.3/0.3/0.3 and discard 20%/20%/42% examples for sound effects/music/speech. For audio generation, we use a temperature of 0.1/0.1/0.1 and discard 20%/20%/28% examples for sound effects/music/speech.

E.3 Model Details

The model details discussed in Sec.2.1, are shown in Table 8

Table 8: Bagpiper Model configuration.

Component	Model Configuration
Text backbone	Qwen/Qwen3-8B-Base
Discrete audio tokenizer	X-Codec (hubert-general)
Continuous audio encoder	Qwen/Qwen3-Omni-30B-A3B-Instruct-Encoder

E.4 Training Details (3-stages)

We train the model in three stages: (1) warmup, (2) large-scale pretraining, and (3) supervised fine-tuning (SFT). Across all stages, we use packed batching (`batchfy_method=pack`), mixed-modality data, and DeepSpeed for distributed optimization. All stages use `dtype=bfloat16` with activation checkpointing enabled. The pre-training is complete on 80 NVIVIDA GH200 GPUs.

Table 9: Training configuration across the three stages. The pipeline progresses from connector alignment to full backbone pre-training, concluding with supervised fine-tuning (SFT).

	Warmup	Pre-train	SFT
Frozen Modules	Audio Encoder, Transformer Body	Audio Encoder	Audio Encoder
Batch Size (Token)	4.8M	1.2M	1.2M
Optimizer	AdamW	AdamW	AdamW
LR Schedule	Constant	Cosine Decay	Constant
Warmup Steps	200	5k	-
Steps	2k	480k	12.5k
Peak LR	5×10^{-4}	1×10^{-4}	1×10^{-5}

E.5 Inference Configuration

We employ modality-specific decoding with `bfloat16` precision and a greedy search ($N = 1$), other details shown in Table 10. `Bagpiper` has around 8B parameters and is served in BF16, so the model weights take roughly 16GB of GPU memory. In practice, a GPU with around 24GB memory or above can run inference smoothly, leaving space for activations and decoding buffers.

Table 10: Inference configuration with modality-specific decoding hyper-parameters.

Item	Temperature	Top-k	Classifier-Free Guidance (CFG)	max-decode-steps
Text decoding	0.6	20	-	2048
Audio decoding	0.8	20	3	2048

E.6 Model Roles

Table 11 lists every model used in Bagpiper, grouped by the role it plays. We separate three kinds of role: the *components* that constitute the model itself, the *data-pipeline* models used to caption, classify, simulate, and filter training data, and the *evaluation* models used to score results. The pipeline is organised so that no evaluation judge contributes to the training data it later scores. In the understanding track, rich captions come from Gemini, the SFT data is simulated and filtered by Qwen3-235B-A22B and GPT-OSS-120B, and evaluation uses verifiable metrics together with GPT-4o. In the generation track, rich captions come from the Qwen3-Omni-Captioner, the simulators are the same, and the A/B test is judged by Gemini-3-Pro and GPT-4o. In each track the judges are therefore disjoint from the models that produced the training data. The pre-training probe is likewise disjoint: pre-training captions come from the Qwen3-Omni-Captioner, while the probing LLM is Gemini-3-Flash. The only model appearing on both sides of the table is Gemini-3-Pro, which produced the attribute taxonomy used to diversify *understanding* prompts and judges *generation* outputs; these are different tracks, so it never scores data it helped create.

Table 11: Models used in Bagpiper and their roles.

Role	Model	Where
Model components		
LLM backbone	Qwen3-8B-Base	§2.1
Continuous audio encoder	Qwen3-Omni-30B-A3B-Instruct encoder	§2.1
Discrete audio codec	X-Codec	§2.1
Pre-training data pipeline		
Rich captioner	Qwen3-Omni-30B-A3B-Captioner	§2.2
Audio taxonomy classifier	Qwen3-32B	§2.2
Text-quality rubric judge	Qwen3-32B	§E.2
Caption summariser (for CLAP)	Qwen3-32B	§E.2
Audio-quality scorers	UTMOS; audiobox-aesthetics	§2.2
Audio–text alignment scorer	CLAP	§2.2
SFT data pipeline		
Rich captioner (understanding)	Gemini-3-Flash; Gemini-2.5-Pro	§2.3
Rich captioner (generation)	Qwen3-Omni-30B-A3B-Captioner	§2.3
Attribute taxonomy extraction	Gemini-3-Pro	§2.3.2
Request / CoT simulator	Qwen3-235B-A22B-Instruct-2507-FP8; GPT-OSS-120B	§2.3.2
Training-data quality filter	Qwen3-235B-A22B-Instruct-2507-FP8	§2.3.2
Evaluation		
Pre-training probing LLM	Gemini-3-Flash	§3.1
Understanding judge	GPT-4o	§3.3.1
Generation A/B judge	Gemini-3-Pro; GPT-4o	§3.3.2

E.7 SFT Filter Validation

The SFT quality filter of §2.3.2 retains samples whose average score over the five rubric dimensions exceeds 3. To check that this threshold is not cutting into the central mass of the score distribution, Table 12 reports the quartiles of the per-example average score together with the resulting retention rate. Under the `Qwen3-235B-A22B` filter the first quartile coincides with the threshold ($p_{25} = 3.00$) and about 75% of the pool is retained, so the cutoff removes the lowest quartile rather than samples near the midpoint. As a cross-family check we scored the same pool with `GPT-OSS-120B`; its scores saturate near the top of the scale and it retains the entire pool, giving little discrimination on this fine-grained rubric. This is why `Qwen3-235B-A22B` is used as the filter in the main pipeline.

Table 12: Validation of the SFT filter cutoff. Quartiles are over per-example average scores across the five rubric dimensions; retention is the fraction of the pool kept at the "average > 3" threshold.

Judge	Avg. score (p_{25} / p_{50} / p_{75})	Retained	Discrimination
`Qwen3-235B-A22B`	3.00 / 3.19 / 3.38	75.0%	informative
`GPT-OSS-120B`	4.75 / 4.88 / 4.94	100%	saturated

E.8 Caption-Error Propagation

We examine whether errors in the intermediate rich captions produced by the fine-tuned `Bagpiper` of §3.3.1 and §3.3.2 propagate to its downstream predictions. For understanding, we evaluate the 333-example speech subset of MMAU-Mini. `Bagpiper` first generates rich captions from audio; `Qwen3-235B-A22B-FP8` then revises these captions using pseudo-ground-truth transcriptions produced by `Qwen3-ASR-1.7B`, after which `Bagpiper` performs the final reasoning from the revised captions. Correcting the transcriptions raises accuracy only from 72.4% to 73.0%, suggesting that ASR errors have limited impact on this subset. For generation, we evaluate 2,620 LibriSpeech test-clean examples. We use `Qwen3-235B-A22B-FP8` to extract the transcription from each intermediate rich caption and compare it with the text in the original user request. The resulting WER is 0.2%, showing that the rich-caption step preserves the requested text with high fidelity in this TTS-style setting; most remaining errors therefore arise in the subsequent caption-to-audio stage.

Table 13: Propagation of errors through the intermediate rich caption.

Analysis	Data	Result
ASR error $\rightarrow$ understanding	MMAU-Mini speech (333)	72.4% $\rightarrow$ 73.0% accuracy
Caption error $\rightarrow$ generation	LibriSpeech test-clean (2,620)	0.2% WER

E.9 Audio-Understanding Ablations

Table 14 consolidates the controlled ablations reported during rebuttal. Applying the same SFT recipe directly to `Qwen3-8B-Base` without audio pre-training degrades all four understanding metrics and, in our qualitative evaluation, does not produce intelligible audio. The CoT comparison isolates the contribution of the caption-plus-reasoning training format under this setup. The backbone-scale comparison holds the training recipe and 50k-step budget fixed; neither row is the final checkpoint. For the pipeline baseline, a frozen `Qwen3-Omni-30B-A3B-Captioner` receives audio without a text prompt, then an instruction-tuned `Qwen3-8B` receives only the resulting caption, question, and answer options. The text LLM has no access to the audio. Despite using this larger perception frontend, the pipeline trails the end-to-end 8B `Bagpiper` by 11.3, 11.9, and 0.7 points on MMAU-Mini, MMAU, and MMAR, respectively.

Table 14: Audio-understanding ablations. Scale variants are evaluated after 50k training steps. "–" denotes an unreported metric.

Ablation	Variant	WER (↓)	MMAU-Mini (↑)	MMAU (↑)	MMAR (↑)
Pre-training	with pre-training	2.5	74.5	73.1	57.0
	SFT-only	4.1	70.2	66.7	49.3
CoT format	with CoT	–	74.5	73.1	57.0
	without CoT	–	71.0	67.0	52.2
Backbone scale @50k steps	`Bagpiper` (Qwen3-8B)	–	68.0	–	–
	`Bagpiper`-30B-A3B	–	74.3	–	–
Inference system	`Bagpiper` (end-to-end)	–	74.5	73.1	57.0
	Captioner + Qwen3-8B	–	63.2	61.2	56.3

E.10 Attribute Taxonomy

For completeness, we list the full set of 108 understanding attributes used to diversify SFT prompts (§2.3), grouped into seven super-categories. We report only the attribute names; no per-attribute definitions or examples are used in the pipeline.

Acoustic & Sound Perception (24). Sound Source Identification; Sound Event Recognition; Environmental Sound Classification; Speech Content Transcription; Music Instrument Identification; Voice Identification; Animal Sound Recognition; Mechanical/Machine Sound Recognition; Natural Sound Recognition; Human Non-Speech Sounds; Musical Genre Classification; Speech vs Non-Speech Detection; Foreground/Background Sound Separation; Sound Quality Assessment; Audio Authenticity Detection; Multi-source Sound Recognition; Sound Intensity/Volume Analysis; Sound Frequency Characteristics; Audio Channel Analysis; Audio Temporal Quality; Ambient Sound Interpretation; Sound Effect Recognition; Musical Texture Analysis; Acoustic Scene Classification.

Speaker & Speech Attributes (18). Speaker Count Estimation; Speaker Gender Identification; Speaker Age Estimation; Speaker Emotion Recognition; Speaker Accent/Dialect Detection; Speaker Native Language Inference; Speaker Role/Occupation Inference; Speaker Identity Verification; Speaker Relationship Inference; Speech Pace/Rate Analysis; Speech Clarity/Articulation; Speech Fluency Analysis; Speech Formality Level; Phonemic Stress Pattern Analysis; Phonological Sequence Decoding; Multi-Speaker Role Mapping; Conversation Turn-Taking Analysis; Speech Act Classification.

Emotion & Sentiment (12). Emotion State Summarization; Emotional Tone Interpretation; Emotion Flip/Change Detection; Sarcasm/Irony Detection; Sentiment Analysis; Speaker Intention Inference; Dissonant Emotion Interpretation; Genuine vs Fake Emotion Detection; Urgency/Importance Detection; Mood Induction Analysis; Threatening vs Non-threatening Classification; Excitement/Enthusiasm Level.

Temporal & Rhythmic (14). Event Counting; Event Duration Estimation; Event Ordering/Sequencing; Temporal Event Reasoning; Sound Onset/Offset Detection; Rhythm and Tempo Understanding; Time Signature Detection; Beat Counting; Temporal Alignment Analysis; Event Frequency Estimation; Duration Comparison; Sequence Prediction; Temporal Anomaly Detection; Event Timeline Construction.

Spatial & Environmental (10). Spatial Sound Localization; Indoor/Outdoor Environment Detection; Room/Space Type Inference; Distance/Proximity Estimation; Sound Movement Direction; Environmental Perception and Reasoning; Background Environment Inference; Acoustic Space Characteristics; Reverberation/Echo Analysis; Multi-location Sound Mapping.

Content & Semantic (18). Factual Information Extraction; Conversational Fact Retrieval; Key Highlight Extraction; Topic/Subject Identification; Main Idea Summarization; Detail Extraction; Lyrical Content Analysis; Spoken Word Comprehension; Dialogue Understanding; Narrative Structure Analysis; Argumentation Analysis; Question Answering from Speech; Instruction Comprehension; Named Entity Recognition; Numerical Information Extraction; Relationship Extraction; Causal Reasoning; Inference/Implication.

Knowledge & Cultural (12). Cultural Context Recognition; Music Theory Knowledge; Professional Domain Knowledge; Historical/Cultural Event Association; Language/Dialect Culture; Musical Genre Cultural Context; Social Convention Understanding; Domain-Specific Terminology; Geographic/Regional Inference; Era/Period Detection; Style/Artistic Movement Recognition; Cross-Cultural Comparison.

F Additional Generation Evaluations

F.1 Matched-Input Control for Open-Ended Generation

In the main-text A/B test (Fig. 5), only the baselines receive the condensed 15-word instruction while Bagpiper receives the full prompt. To isolate the model contribution from this input-length difference, we additionally give Bagpiper the same 15-word summary (Bagpiper-summary) and re-run the evaluation under both judges (Table 15), comparing it both against the full-prompt baselines and against the matched *summary* baselines. Bagpiper-summary still wins in aggregate against every baseline under both judges; the margins shrink (most on *sound* against TangoFlux-summary under GPT-4o, where the per-domain rate dips below 50%), while the aggregate win rate stays above 50% against every baseline, indicating the advantage is largely not an input-length artefact. Since Bagpiper can still apply its caption-then-process reasoning under short prompts, this is a practical matched-input control rather than a complete ablation of the reasoning step.

Table 15: Matched-input control: A/B win rates (%) of Bagpiper-summary (given the same 15-word instruction as the baselines), broken down by **Sp**eech / **So**und / **Mu**sic / **All** under two judges.

	Gemini-3-Pro				GPT-4o			
Bagpiper-summary vs	Sp	So	Mu	All	Sp	So	Mu	All
AudioLDM2-Large	99.1	72.2	81.9	84.4	94.9	56.5	65.9	72.5
AudioLDM2-Large-summary	99.7	74.0	84.0	85.9	96.4	63.4	70.4	76.8
TangoFlux	97.9	54.7	81.3	78.0	87.1	46.5	65.6	66.4
TangoFlux-summary	98.5	47.7	81.3	75.9	85.6	40.5	67.7	64.6

F.2 TTS-Specialized Variant

Figure 5 measures open-ended instruction-following rather than dedicated text-to-speech quality. To compare against specialized systems on their own terrain, we fine-tune a TTS-specialized Bagpiper variant and run side-by-side preference tests on LibriSpeech test-clean against Qwen3-TTS, CosyVoice3, and VibeVoice, judged separately on *overall quality* and *prosody/naturalness* under both Gemini-3-Pro and GPT-4o (Table 16). The variant is preferred over both CosyVoice3 and VibeVoice under both judges and both rubrics, and is competitive with Qwen3-TTS: the only sub-50% cell is GPT-4o prosody/naturalness against Qwen3-TTS (48.7%). In the pooled pairwise ranking the ordering is consistent across judges and rubrics: Bagpiper > VibeVoice > CosyVoice3.

Table 16: A/B win rates (%) of the TTS-specialized Bagpiper variant on LibriSpeech test-clean, judged separately on overall quality and prosody/naturalness under two judges. Reported separately from Fig. 5.

Bagpiper-TTS vs	Rubric	Gemini-3-Pro	GPT-4o
Qwen3-TTS	Overall quality	71.3	57.5
	Prosody / naturalness	72.0	48.7
CosyVoice3	Overall quality	84.0	65.7
	Prosody / naturalness	85.7	57.7
VibeVoice	Overall quality	61.7	55.3
	Prosody / naturalness	58.3	51.3

F.3 Stage-Attributed Qualitative Study

We attribute the qualitative behaviors of §3.4 to the training stage that produces each. **Base (Qwen3-8B).** Before audio training the base model is not an audio generator at all: it has no

mechanism to emit audio tokens and produces no audio output. **Pre-trained (`Bagpiper-Base`).** Audio pre-training endows the model with compositional audio-token generation; intelligible, well-formed audio begins to emerge at roughly 100k pre-training steps. **Post-SFT (`Bagpiper`).** Supervised fine-tuning makes this generative substrate controllable by open-ended natural-language requests, yielding the compositional synthesis, logic-constrained instruction following, and world-knowledge grounding observed in §3.4. The generative capability is therefore attributable to pre-training and its controllability to SFT.

G LLM Usage Disclosure

In compliance with the COLM 2026 LLM Usage Policy, we disclose that Large Language Models (LLMs)—both locally deployed and accessed via remote APIs—were utilized for data curation, simulation, and verification tasks as detailed in §2. During manuscript preparation, LLMs provided minor assistance limited to grammar refinement and rephrasing. We explicitly state that LLMs were NOT used to originate core research ideas, draft original technical prose, or generate experimental plots.