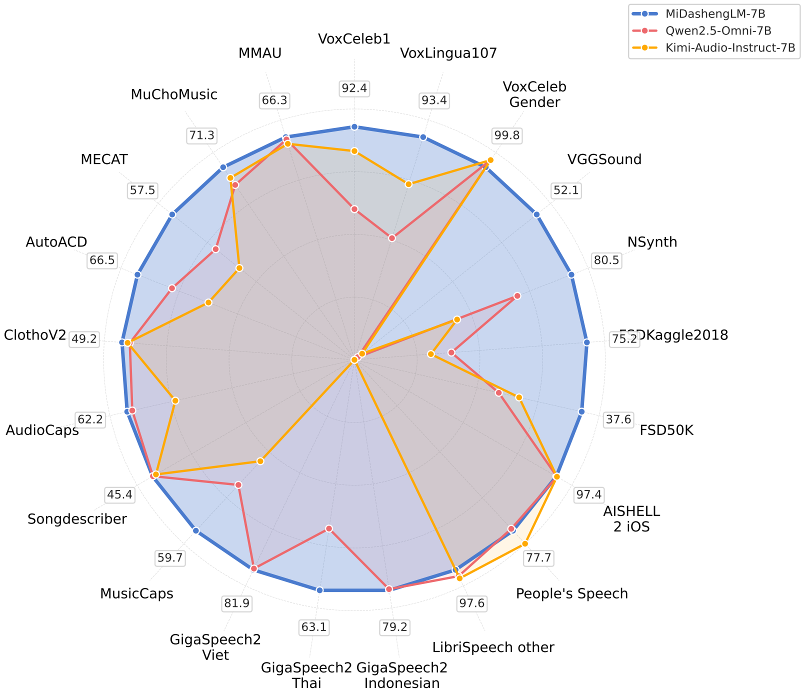
MiDashengLM: Efficient Audio Understanding with General Audio Captions

Horizon Team, MiLM Plus

Xiaomi Inc., China

Abstract

Current approaches for large audio language models (LALMs) often rely on closed data sources or proprietary models, limiting their generalization and accessibility. This paper introduces MiDashengLM, a novel open audio-language model designed for efficient and comprehensive audio understanding through the use of general audio captions using our novel ACAVCaps training dataset. MiDashengLM exclusively relies on publicly available pretraining and supervised fine-tuning (SFT) datasets, ensuring full transparency and reproducibility. At its core, MiDashengLM integrates Dasheng, an open-source audio encoder, specifically engineered to process diverse auditory information effectively. Unlike previous works primarily focused on Automatic Speech Recognition (ASR) based audio-text alignment, our strategy centers on general audio captions, fusing speech, sound and music information into one textual representation, enabling a holistic textual representation of complex audio scenes. Lastly, MiDashengLM provides an up to 4× speedup in terms of time-to-first-token (TTFT) and up to 20× higher throughput than comparable models. Checkpoints are available at and 🤗.



Technical report.

1 Introduction

Large language models (LLMs) have played a pivotal role in advancing machine learning approaches for natural language processing (NLP), demonstrating impressive capabilities in understanding the world through text. While these models can effectively interact with humans via text, the ability to understand sound remains crucial for agents to fully engage with the physical world. Large Audio-Language Models (LALMs) aim to bridge the gap between auditory and textual understanding. Within the audio domain, we identify three commonly used broad categories: speech, (environmental) sounds and music. Aligning audio with text requires a mapping between speech/sound/music and respective text. For speech the most common alignment are transcripts, while captions are used for sound and music. Transcripts can be understood as a monotonous alignment between audio and text domains. In contrast, captions are typically used for broader audio elements like sounds and music, offering a more generalized alignment, meaning they capture the overall nature or occurrence of a sound.

Current audio understanding research typically processes speech transcripts, audio captions, and music captions separately. This independent approach limits the depth and completeness of auditory scene analysis. Anther key limitation stems from existing audio captions, which often offer only superficial descriptions. For example, spoken content is frequently simplified to "somebody is speaking", ignoring semantic details. Furthermore, these datasets often fail to capture critical auditory aspects like room acoustics (e.g., reverberation) or signal quality.

To overcome these limitations, this paper proposes fusing speech transcripts, audio captions, and music captions into a single, unified general caption. Our goal is to create a holistic textual representation that jointly includes all relevant audio information, providing a more detailed and semantically rich description of the auditory environment.

1.1 Motivation

Developing a LALM requires aligning audio features with textual descriptions. Utilizing sound and music captions as a training target has been previously explored [1; 2; 3; 4] to enhance audio understanding. However, these approaches lack automatic speech recognition (ASR) capabilities, limiting their usefulness for general applications, as users expect a LALM to handle both general audio understanding and speech — not just captions. The most used alignment paradigm couples large language models (LLMs) with audio understanding through automatic speech recognition (ASR). This approach prevails for two key reasons: First, numerous high-quality off-the-shelf ASR models exist that can generate reasonably accurate transcripts automatically. Second, a substantial portion of internet audio content consists of speech-based material - including podcasts, lectures, interviews, and other spoken-word formats - making ASR an effective bridge between audio and text modalities. Several prominent works have demonstrated the effectiveness of ASR-based LALM training, such as Whisper [5], SpeechT5 [6], Universal Speech Model (USM) [7], Open Whisper-style Model (OWSM) [8] and Kimi-Audio [9]. However, we argue that ASR-based pretraining provides limited benefits for general audio-language understanding, due to the following reasons:

Inefficient Data Utilization Large-scale pretraining on million-hour long datasets typically relies on existing automated speech recognition (ASR) pipelines to generate transcripts from speech. This results in a substantial loss of potentially valuable data, as sounds like music, environmental noises, or even silent pauses are discarded. Using a general captioning approach has the benefit that any audio can be used for training, as even "noisy" audio clips could be labeled. This significantly enhances data diversity, allowing models to learn from a much wider range of acoustic information beyond just speech.

Trivial objective The training losses for ASR-based LALMs are typically low, even across different languages, suggesting that the models learn relatively little meaningful information from ASR-based data, compared to text-based training [10] (see Figure 1). We attribute this to the simplicity of speech-text alignments, where the temporal ordering of acoustic units and their corresponding text tokens follows a monotonic (left-to-right) correspondence. Thus a model only needs to establish local correspondences between spoken words and their textual counterparts, bypassing the need to understand broader (global) audio context.

Limitations of ASR-Based Pretraining Beyond Speech Content ASR-based pretraining does not focus on information other than the spoken content. This limited scope means that important speech meta-information, such as a speaker's gender, age, or emotional state, is not captured or integrated during the pretraining process. Furthermore, the pretraining methodology overlooks audio signal-specific characteristics like reverberation levels, recording quality, and environmental acoustics.

1.2 Audio caption and speech summarization

Audio captions have been the focus of extensive research [11; 12; 13]. Most datasets during the start of the audio-caption era were manually labeled [14; 11; 12; 15], but recent work has leveraged large language models (LLMs) to scale and streamline dataset creation. Notable LLM-assisted audio captioning datasets include WavCaps [16], AutoACD [17], SoundVECaps [18], AudiosetCaps [19] and FusionAudio-1.2M [20].

These works utilized LLMs in order to enhance existing audio captions by additional visual information [18], temporal information [17] or with additional CLAP filtering [19; 3]. However, we identify two key limitations in existing datasets:

Neglect of spoken language: Publicly available captioning data primarily focuses on sound/music events and their audio-visual/temporal relationships, despite speech constituting the majority of real-world audio [21]. Current audio captioning datasets can therefore be better understood as (environmental-) sound captioning datasets. Limited data diversity: Popular datasets (AudioCaps, WavCaps, AutoACD, SoundVECaps, AudiosetCaps and FusionAudio-1.2M) predominantly derive from the same audio sources (Audioset [21], VGGSound [22] and FSD50k [23]). This source overlap leads to a problematic one-to-many mapping: multiple “distinct” datasets are, in fact, derived from identical underlying audio clips, containing different textual descriptions. This redundancy adds little training audio data variation, limiting model generalization.

While audio captions have been used for LALM pretraining, existing approaches typically generate new captions through either (1) paraphrasing existing descriptions [3; 24] or (2) augmenting them with (unrelated) video context [4] using LLMs, rather than genuinely diversifying the underlying audio content.

In our work we rely on *general audio captions*, a novel captioning type. General audio captions can be understood as a fusion of speech summarization [25], music captions and audio captions into one.

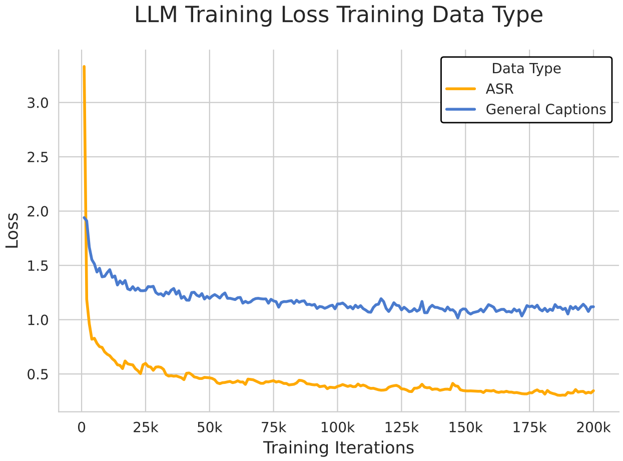


Figure 1: Training cross entropy loss (next token) curves between ASR and caption based pretraining. General captions utilize the ACAVCaps (Table 15) dataset, while ASR uses ACAV100M-Speech (Table 14). ACAV100M-Speech contains up to 90 different languages, while captions are English only.

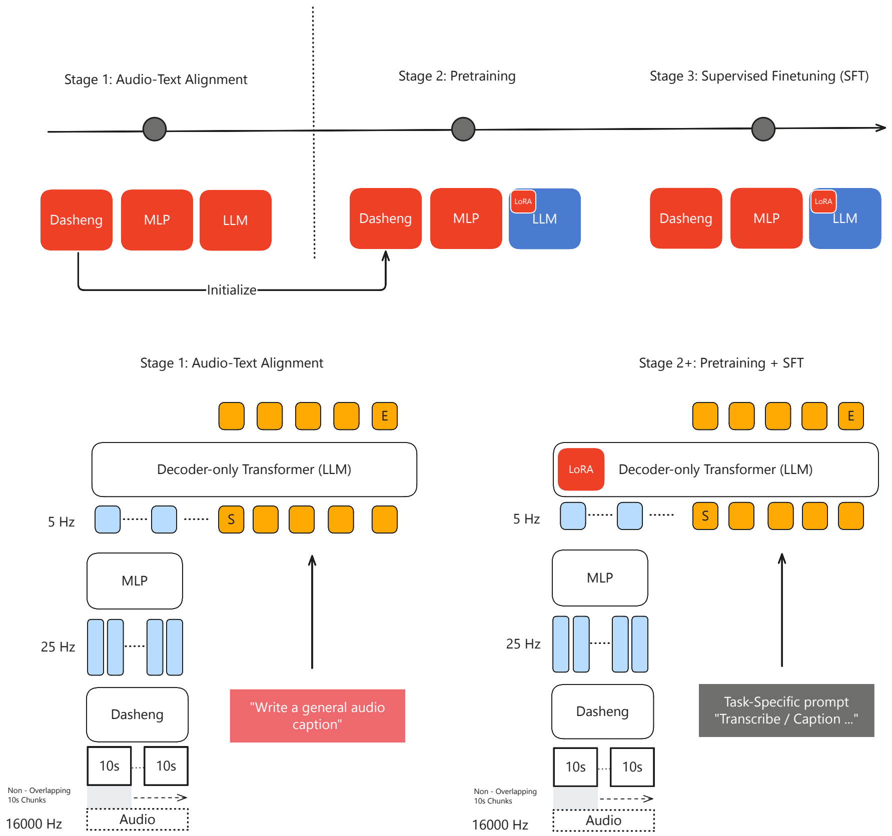

Stage 1: Audio-Text Alignment
Stage 2: Pretraining
Stage 3: Supervised Finetuning (SFT)
Dasheng
MLP
LLM
Dasheng
MLP
LoRA
LLM
Dasheng
MLP
LoRA
LLM
Initialize
Stage 1: Audio-Text Alignment
Stage 2+: Pretraining + SFT
E
Decoder-only Transformer (LLM)
5 Hz
S
MLP
25 Hz
Dasheng
"Write a general audio caption"
10s
10s
Non - Overlapping 10s Chunks
16000 Hz
Audio
E
LoRA
Decoder-only Transformer (LLM)
5 Hz
S
MLP
25 Hz
Dasheng
Task-Specific prompt "Transcribe / Caption ..."
10s
10s
Non - Overlapping 10s Chunks
16000 Hz
Audio

Figure 2: Proposed MiDashengLM framework. For all three stages, training is done with standard next-token prediction loss. Stage 1 aligns the audio encoder with the text modality, after which the audio encoder is taken and initialized for Stage 2.

2 Framework

Our proposed framework can be seen in Figure 2. The framework is a common prefix-based large language model, where features of an audio encoder are mapped into the embedding space of an LLM via a multilayer perceptron (MLP) layer. Our framework mainly differences from previous works in the following regards.

Public data Our approach only uses publicly available audio-text data for pretraining, supervised finetuning (SFT) and instruction tuning. All data sources are listed in Tables 14 to 18.

Audio-text alignment Training LALMs is generally seen as an alignment problem, that aims to map audio features into a text-based space, such that an LLM can process these audio tokens. In order to improve the training speed and performance, the vast majority of works utilize pretrained audio encoders. One of the most prevalent pre-trained model is the Whisper encoder [5], as seen in models like LTU-AS [26], Qwen-Audio [27], Qwen2-Audio [28], and Kimi-Audio [9], Mini-Omni [29], Llama-Omni [30], R1-AQA [31] and SALMONN [32]. Other audio encoders such as HuBERT [33], HTS-AT [34], AST [35] and BEATs [36] have also been utilized, often as secondary encoders to accommodate sound/music task knowledge. To the best of our knowledge, this paper is the first to propose audio-text alignment via general captions, without relying on ASR or Sound event based models. Further, we only utilize a *single* general audio encoder that is jointly capable of processing speech, sound and music.

Training efficiency Even though transformer models are fully parallelize during training, they scale quadratically with regards to the input sequence length. Since most audio data used for LALM

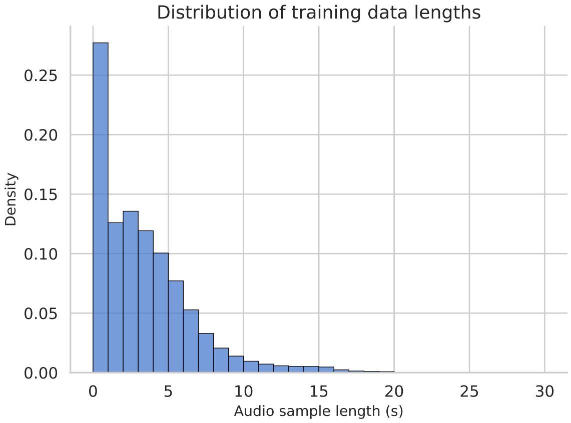


Figure 3: Histrogram plot of training data sample lengths.

training has different lengths, one requires padding in order to batch samples towards a fixed sequence length. One common way to significantly speedup training is by reducing the amount of padding by grouping samples with similar length together. However, models such as Whisper natively does not support variable sequence lengths during training or inference and pads by default all inputs to a fixed duration of 30 seconds [28]. Changing this behavior can lead to significant performance degradation [37; 38]. We plot our training dataset's sample length distribution in Figure 3. Since the majority of samples are between 1 and 10s long, padding to 30s would lead to inefficient training and inference, since the majority of encoder compute is wasted. In contrast, our audio-encoder supports variable length inputs, significantly reducing the amount of padding and improve training efficiency. More importantly, the majority of compute is done in the decoder, which benefits heavily from shorter sequences. To further boost efficiency, we aggressively downsample the audio sequence length to a low framerate of 5 Hz, to accommodate fast training and inference speeds.

3 Datasets

MiDashengLM is trained solely on publicly available datasets during its pretraining and supervised finetuning phases. All our training datasets are provided in Appendix A. We further provide information about our novel general audio caption dataset.

3.1 ACAVCaps and Multi-Expert Chain for Audio Tasks (MECAT)

As discussed in Section 1.1, previous captioning datasets are insufficient mainly due to the lack of speech understanding and their monotonous data source mainly stemming from Audioset [21], VGGSound [22] and FSD50k [23]. We identify that for our purposes, we would like a dataset that is publicly available and rich in content, containing multilingual speech, different types of music and a plethora of complex audio environments. We identify ACAV100M [39] as a plausible source dataset candidate for these purposes, since it has not been labeled for audio captioning before and contains little overlap with previously mentioned datasets.

Since ACAV100M lacks labels, we developed an efficient data curation pipeline. We began by using CED-Base [40] to predict AudioSet labels on a 2-second scale. We use this finer 2-second scale to enable our captions to capture temporal relationships. Having obtained sound event labels, we further process the data using a plethora of different audio classification models, each tailored for a specific task.

Speech Analysis: This curation task identifies spoken language, distinguishes individual speakers, segments audio by speaker (diarization), detects speech emotion, classifies speaker gender and age and infers a transcript using Whisper [5]. **Vocal Analysis:** Beyond basic speech, this curation task refines vocal emotion detection, assesses vocal health, and analyzes unique vocal characteristics like pitch and timbre. **Music Analysis:** For musical content, models classify music genre, recognize instruments, detect tempo, analyze music mood, and identify singing voices. **Environmental Acoustics:** This

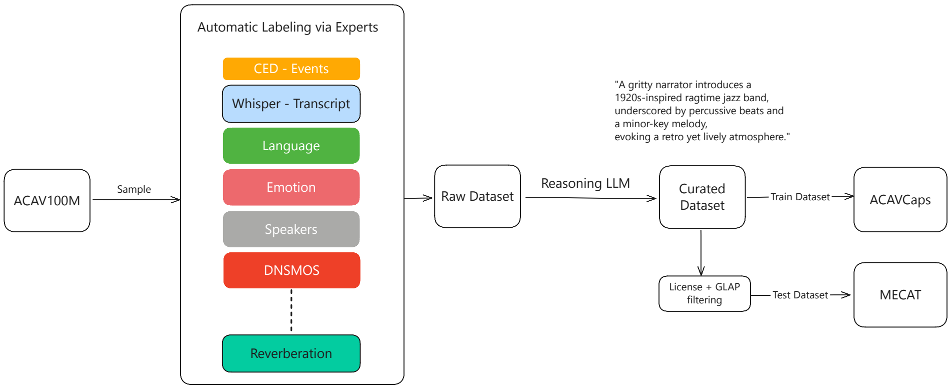

Automatic Labeling via Experts
CED - Events
Whisper - Transcript
Language
Emotion
Speakers
DNSMOS
Reverberation
ACAV100M
Sample
Raw Dataset
Reasoning LLM
"A gritty narrator introduces a 1920s-inspired ragtime jazz band, underscored by percussive beats and a minor-key melody, evoking a retro yet lively atmosphere."
Curated Dataset
Train Dataset
ACAVCaps
License + GLAP filtering
Test Dataset
MECAT

Figure 4: Our proposed data curation pipeline. We filter ACAV100M with an automatic pipeline, that predicts transcripts, sound events, sound quality and other meta information. A reasoning-LLM is then used to generate a caption from the provided meta information. The resulting curated dataset is then split into a training set (ACAVCaps) and a novel evaluation set Multi-Expert Chain for Audio Tasks (MECAT).

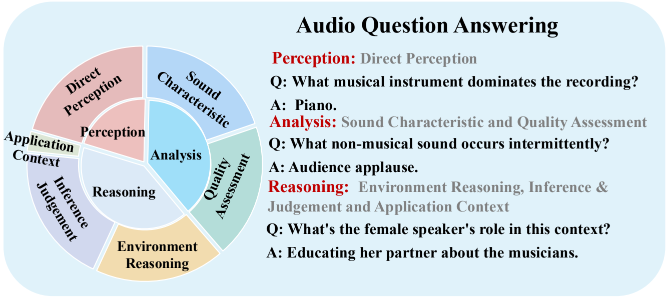
part of the pipeline categorizes the acoustic scene, assesses audio quality, analyzes reverberation, and identifies various noise types.

Category	Caption
Pure Speech	A female voice narrates a historical team competition (1966–1971) based on basketball rules, with intermittent synthetic speech modulation and variable acoustic reverberation.
Pure Sound	An outdoor scene with wind blowing, birds chirping, and a duck quacking, accompanied by significant background noise and low audio quality.
Pure Music	*"If I were a zombie, I'd want your heart, not your brain"* — A quirky electronic-pop anthem with gritty vocals, pulsing beats, and a dash of dark romance.
Mixed Music	The audio features a crowd cheering and clapping alongside electronic music with a synthesizer-driven, dark, and energetic soundscape.
Mixed Speech	A Russian voice demonstrates a synthesizer's capabilities over an experimental electronic backdrop, explaining its sound design and value in a gritty, vocal-fry tone.
Mixed Sound	A man speaks in English about entering a city and village, accompanied by the sounds of a running vehicle.

Table 1: A selection of our general audio captions generated by the proposed pipeline.

Having obtained all these labels, we prompt a reasoning LLM (DeepSeek-R1 [41]) in order to generate a short audio caption. The resulting curated audio caption dataset is then split into a train-set (ACAVCaps) and test-set (**M**ulti-**E**xpert **C**onstructed Benchmark for Fine-Grained **A**udio Understanding **T**asks, MECAT). MECAT is extracted from the curated dataset by filtering each source video by license and finally performing GLAP [42] to score the audio-text consistency. A depiction of our pipeline can be seen in Figure 4. MECAT will also be made publicly available [43]. Lastly, we segment the dataset into six respective categories according to their CED labels, which can be seen in Table 1.

Statistics about our resulting captioning training set can be seen in Table 2. Notably, LAION-Audio-300M is a dataset that focuses on speech-only captions, neglecting sounds. As we can see, our proposed dataset has a much richer vocabulary than previous approaches. There are two main reasons for this. First, since our captions summarize spoken content, the vocabulary naturally increases against other sound-event focused captions. The second reason is the multilingual nature of our source dataset, where often transcripts from a foreign language are kept in the final caption e.g., "A



Perception: Direct Perception

Q: What musical instrument dominates the recording?

A: Piano.

Analysis: Sound Characteristic and Quality Assessment

Q: What non-musical sound occurs intermittently?

A: Audience applause.

Reasoning: Environment Reasoning, Inference & Judgement and Application Context

Q: What's the female speaker's role in this context?

A: Educating her partner about the musicians.

Figure 5: Subtasks of the proposed MECAT-QA testset.

synthesized Spanish voice narrates a tense zombie confrontation: "Repentinamente... golpe varias veces" delivered with mechanical flatness amid variable reverberation and background noise."

MECAT-QA In MECAT-QA, each audio clip is paired with five question-answer pairs that span different categories and difficulty levels, resulting in over 100,000 total QA pairs. They are organized into three main cognitive categories: a) **Perception**, which consists of a single sub-category, *Direct Perception*, focusing on the direct identification and naming of audio content and events. b)**Analysis**, which is composed of two sub-categories: *Sound Characteristics*, for examining the acoustic properties of a sound (e.g., pitch), and *Quality Assessment*, for evaluating the technical fidelity of the audio (e.g., noise level). c) **Reasoning**, which covers higher-level cognitive skills and is divided into three sub-categories: *Environment Reasoning*, requiring the inference of the acoustic scene in which the sound occurs; *Inference & Judgement*, involving logical deductions and judgments based on the audio content; and *Application Context*, testing the understanding of a sound's practical purpose or scenario. A short introduction of available tasks and samples can be seen in Figure 5.

Table 2: Comparison of publicly available captioning datasets. Datasets denoted with ‡ contain multilingual captions. The number of unique words (# Vocab) and the average sentence length are displayed.

Dataset	Labeling	#Vocab	Avg. Sent	Source
ClothoV2 [12]	Manual	4366	11.32	Freesound
AudioCaps [14]		4844	8.70	Audioset
MusicCaps [44]		3730	47.17	Audioset
Songdescriber [45]		1811	26.31	MTG-Jamendo
LPMusicCaps-MTT [46]	LLM	4045	25.04	MagnaTagATune
LPMusicCaps-MSD [46]		14049	37.06	MillionSoundDatabase
SoundVECaps [18]		58401	31.48	Audioset
AutoACD [17]		20491	18.47	Audioset
AudiosetCaps [19]		21783	28.13	Audioset + VGGSound
WavCaps [16]		24592	7.84	Audioset + BBC + FreeSound + SoundBible
LAION-Audio-300M [47]		451927	37.55	?
Ours‡	Reasoning-LLM	644407	22.18	ACAV100M

3.2 Training datasets and tasks

Our publicly available data sources, detailed in Appendix A, comprise approximately 1.1 million hours of data. Notably, approx. 90% of the training data originates from public ASR datasets, while the remaining datasets are significantly smaller. If not properly treated, this would lead to inadequate performance for tasks other than ASR. Data sampling can be viewed in Figure 6. For audio-text alignment, we utilize the previously introduced ACAVCaps dataset (see Section 3.1), which contains 38,000 hours of high-quality general captions. We train for three epochs on ACAVCaps to align the

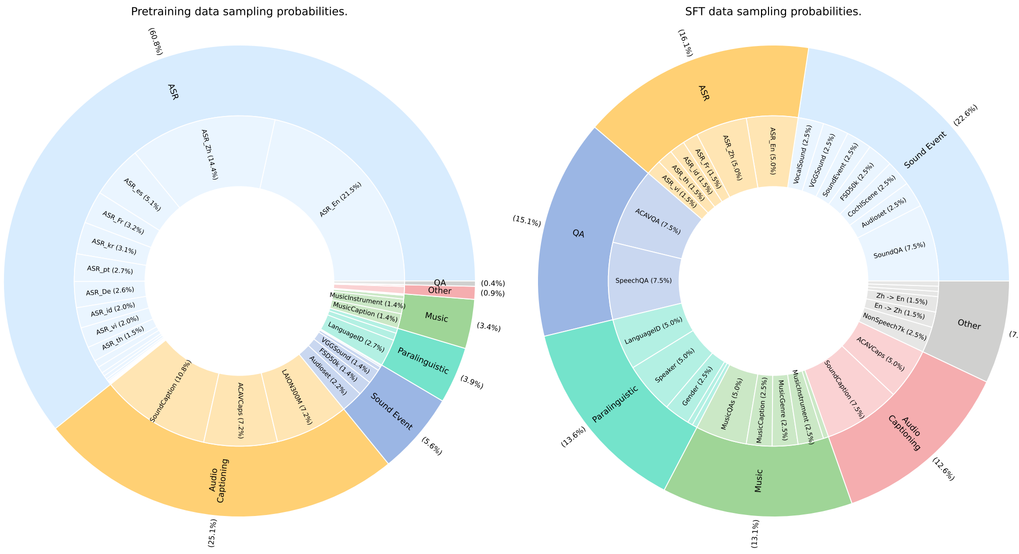


Figure 6: Pretraining and SFT sampling across datasets.

audio encoder with text. Following alignment, we pretrain MiDashengLM on the full 1.1 million hours of training data for approximately 1.4 epochs. After pretraining, we conduct supervised fine-tuning for one additional epoch on a curated subset of the pretraining data, totaling 352k hours. Further details on the datasets used can be found in Appendix A.

4 Experimental Setup

MiDashengLM is a standard Transformer-based encoder-decoder model [48], comprising a Transformer audio encoder and a text decoder. The audio encoder builds upon Dasheng-0.6B [49], a *frame-level* Vision Transformer (ViT) [50] pretrained using the Masked Autoencoder (MAE) objective [51], primarily on the ACAV100M dataset. MiDashengLM exclusively supports 16 kHz audio inputs, and all input data is automatically resampled to this sampling rate. Audio waveforms are converted into 64-dimensional mel-spectrograms, which Dasheng-0.6B processes by extracting 32 ms frame features with a 10 ms stride.

By default, Dasheng further downsamples the input features by a factor of four, producing high-level features at 40 ms intervals. As noted in Section 1.1, Dasheng supports variable-length inputs, with a maximum input length of 1008 frames (10.08 seconds). For longer inputs, we apply a non-overlapping sliding window approach, by forwarding each chunk through Dasheng, and concatenating the resulting frame-level features. The complete hyperparameter configuration is documented in Table 4, with a systematic comparison between our audio encoder and Whisper's architecture presented in Table 3. Training the full pipeline required roughly 19,200 GPU hours, or 10 days on 80 GPUs.

Audio-text alignment Pretraining for our Dasheng-based audio encoder is done via the masked-autoencoder (MAE) objective, which learns high-level audio features in a latent space. However, a major difference between Whisper and our proposed Dasheng based encoder is that Whisper has been aligned with textual data (ASR). Thus the first step of our MiDashengLM aligns the audio encoder with textual data. For this alignment stage, we employ the ACAVCaps dataset, performing end-to-end fine-tuning of both the audio encoder and text decoder components. Following alignment, we extract the trained audio encoder for initialization in subsequent pretraining and SFT phases. During model development, we empirically evaluated two alternative approaches: (1) integration with a frozen large language model (LLM) and (2) low-rank adaptation (LoRA [52]). However, both approaches yielded unsatisfactory audio encoder performance. Audio-text alignment ran with an effective batch-size of 256 on 8 GPUs for one day.

Table 3: Audio encoder differences between our proposed model and the more common Whisper-Large v3.

	Whisper-Large v3	Ours
Parameters	637.7M	630.3M
Pretraining data size	5M	270k
Training Objective	ASR	General captions
Context	30s	10s
Known pretraining data?	✗	✓ [39]
Open train code?	✗	✓ [49]
Open weight?	✓	✓

Text Decoder The text decoder is initialized using Qwen2.5-Omni-3B [53], a publicly available pretrained language model. For both pretraining and supervised fine-tuning phases, we employ LoRA to enhance parameter efficiency. The training objective minimizes the standard cross-entropy loss:

$$\mathcal{L}_{ce} = -\log P(x_t|x_{1:t-1}, A),$$

where x_t is the current text token, $x_{1:t-1}$ represents the past text tokens, and A denotes the audio features.

Training All training procedures incorporate a linear learning rate warm-up spanning the initial 1,000 iterations, during which the learning rate increases from zero to the target value. Subsequently, the learning rate follows a cosine decay schedule, progressively decreasing to 10% of its maximum value by training completion. Notable differences between pretraining and SFT include: (1) a reduced learning rate during SFT, and (2) the expansion of trainable parameters influenced by LoRA. The training hyperparameters are provided in Table 4. Here “all-linear” modifies all projection layers within the decoder using LoRA, while “q,v” exclusively adapts the query and value matrices within the self-attention layers.

Table 4: Decoder Hyper Parameters for MiDashengLM-7B and Training Configuration.

	Stage	
Parameter	Pretrain	SFT
Decoder-Size	7B	
Optimizer	AdamW8bit	
LoRA rank	8	
LoRA alpha	32	
LoRA dropout	0.1	
Audio-token framerate	5 Hz	
Learning rate	1e-4	1e-5
Weight decay	0.01	0.1
LoRA target	q,v	all-linear
Batchsize	10	8

5 Results

We evaluate performance on each dataset’s designated standard test/evaluation split.

5.1 Audio encoder performance

To evaluate our audio-text alignment framework trained with general audio captions, we compare the resulting audio encoder against Whisper-Large V3. We employ the X-Ares benchmark [54], which evaluates frozen encoder embeddings through a lightweight MLP layer across three core audio domains: speech, music, and (environmental) sound.

Table 5: Performance Comparison between our proposed captioning pretrained (Dasheng) model and Whisper-Large V3 (Whisper) using the X-Ares benchmark. For all metrics, higher is better and the best results are visualized in boldface.

Domain	Dataset	Ours	Whisper	Ours vs. Whisper
Speech	LibriCount	61.9	**64.4**	-3.9
	LibriSpeech-100h	85.4	**90.0**	-5.1
	LibriSpeech-MF	**98.5**	94.9	+3.8
	VoxLingua33	92.3	**97.4**	-5.2
	Speech Commands V1	97.4	**97.7**	-0.3
	CREMA-D	**77.0**	71.3	+8.0
	Fluent Speech Commands	**98.1**	97.8	+0.3
	RAVDESS	**76.1**	68.5	+11.1
	Vocal Imitation	**31.2**	29.3	+6.5
	VocalSound	**93.2**	91.5	+1.9
	VoxCeleb1	**73.3**	24.8	+195.6
Sound	ASV2015	**99.3**	97.9	+1.4
	Clotho	**5.8**	3.1	+87.1
	DESED	**53.7**	22.6	+137.6
	ESC-50	**94.3**	62.5	+50.9
	FSD50k	**55.5**	32.0	+73.4
	FSD18-Kaggle	**82.2**	49.6	+65.7
	UrbanSound 8k	**87.9**	75.7	+16.1
Music	Free Music Archive Small	**67.2**	58.9	+14.1
	GTZAN Genre	**88.6**	71.8	+23.4
	MAESTRO	**54.5**	0.0	+∞
	NSynth-Instruments	**72.2**	63.5	+13.7

As shown in Table 5, our Dasheng-based encoder demonstrates strong performance across diverse audio classification tasks. Comparative analysis reveals that while Whisper-Large v3 achieves superior results on 4 of 22 tasks, our encoder outperforms Whisper on the remaining 18 tasks. Whisper outperforms our proposed encoder on tasks such as automatic speech recognition (ASR) by 5% WER, speaker counting (LibriCount), spoken language recognition (VoxLingua33) and keyword spotting (Speech Commands V1). All of those tasks are strictly speech-related. On the other hand our proposed audio encoder outperforms Whisper-Large v3 on the majority of environment, music and sound classification tasks. Largest gains are achieved for speaker recognition (VoxCeleb1, + 195%), domestic sound event classification (DESED, + 137 %) and Audio-text retrieval (Clotho, + 87%). These results demonstrate that audio-text alignment through general audio captions represents an effective approach for high-performance general-purpose audio understanding.

5.2 Traditional dataset Benchmarks

Table 6: Comparison between the proposed MiDashengLM and baseline models.

Parameter	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Encoder	Dasheng-based	Whisper-based	Whisper-based
Decoder Parameters	7B	7B	7B
Audio-token framerate ↓	**5 Hz**	25 Hz	12.5 Hz
Audio-text alignment	General caption	ASR	ASR
Capable of ASR ?	✓	✓	✓
Known pretraining data ?	✓	✗	✗

We evaluate our proposed MiDashengLM on common benchmarks against two strong baselines: Qwen2.5-Omni [53] and Kimi-Audio-Instruct [9]. Note that we exclusively compare with general audio understanding models that are capable of captioning as well as spoken language understanding in order to compare fairly, since there exist work solely optimized for captions only [4; 2]. A short overview about the models can be seen in Table 6. For all subsequent results in tables and figures, we explicitly indicate decoder sizes using the following nomenclature: Qwen2.5-Omni-7B (Qwen2.5-Omni), Kimi-Audio-Instruct-7B (Kimi-Audio-Instruct) and MiDashengLM-7B (MiDashengLM).

5.2.1 Audio captioning results

Results for audio captioning can be seen in Table 7, where we select FENSE [55] as our primary audio caption metric. For both music and audio (sound) captioning datasets, MiDashengLM outperforms consistently the baseline models. The performance gains are particularly significant for general audio, with our model substantially outperforming baselines on AutoACD, while showing more modest improvements on music-specific benchmarks.

Table 7: Results for traditional music and audio captioning datasets. All results represent FENSE, where higher is better and best is in bold.

Domain	Dataset	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Music	MusicCaps	**59.71**	43.71	35.43
	Songdescriber	**45.39**	45.31	44.63
Sound	AudioCaps	**62.18**	60.79	49.00
	ClothoV2	**49.20**	47.55	48.01
	AutoACD	**66.52**	55.93	44.76

5.2.2 MECAT

Unlike traditional captioning datasets, MECAT provides a comprehensive evaluation framework across nine distinct domains: short captions, long captions, and pure/mixed categories of speech, sound, and music, along with environmental captions. This benchmark requires domain-specific caption generation—for instance, environmental captions must exclude spoken content, while pure-speech outputs should focus exclusively on verbal elements. As shown in Table 8, our results align with findings from standard audio captioning benchmarks (Table 7). From these results we observe that Kimi-Audio-Instruct performs poorly for captioning tasks. Further, MiDashengLM, benefiting from its general captioning capabilities, surpassed the baselines by a significant margin.

Table 8: Model Performance Comparison on MECAT. All results represent FENSE, where higher is better and best is in bold.

Task	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Content Long	**60.11**	48.34	40.83
Content Short	**61.38**	45.29	45.72
Pure Speech	**50.69**	37.27	25.57
Pure Sound	**53.78**	46.60	35.75
Pure Music	**66.17**	50.68	39.54
Mixed Speech	**51.06**	37.43	27.12
Mixed Sound	**32.40**	32.07	19.44
Mixed Music	**59.50**	34.71	16.18
Environment	**51.38**	47.84	16.66
Overall	**57.53**	43.80	36.32

5.2.3 Audio and paralinguistic classification

We next evaluate our approach on paralinguistic tasks, with results detailed in Table 9. Note that we directly test the model's capabilities of each respective dataset, while other reports such as Kimi-

Audio prompt the model with a choice of available labels. For speaker verification (VoxCeleb1), we introduce a novel evaluation protocol that presents utterance pairs (same or different speakers) for binary classification. We combine pairs of utterances - either from the same speaker or different speakers - and task the model with determining whether the two utterances originate from the same speaker or different speakers. Performance across the ten tested tasks implicate that MiDashengLM outperforms baselines for speaker verification (VoxCeleb1), Language identification (VoxLingua107), Sound classification (VGGSound, FSD50k) and Music classification (NSynth, FMA).

Table 9: Results for audio classification and paralinguistic benchmarks. Best in bold.

Dataset	Metric	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
VoxCeleb1	ACC ↑	**92.36**	59.71	82.72
VoxLingua107		**93.41**	51.03	73.65
VoxCeleb-Gender		96.12	99.82	**99.69**
VGGSound		**52.11**	0.97	2.20
Cochlscene		**74.06**	23.88	18.34
NSynth		**80.52**	60.45	38.09
FMA		63.73	**66.77**	27.91
FSDKaggle2018		**75.25**	31.38	24.75
AudioSet	mAP ↑	**8.86**	6.48	3.47
FSD50K		**37.58**	23.87	27.23

5.2.4 Automatic speech recognition

We assess ASR performance across all models using standard public benchmarks (see Table 10). We would like to point out that audio-token framerate significantly impacts ASR performance, with higher rates improving performance at the expense of computational efficiency (Table 6). These results align with our earlier findings in Table 5, demonstrating that our encoder continues to trail the closed-source Whisper model - the audio encoder employed by both baseline systems. Since MiDashengLM is a captioning model first and foremost, it's ASR performance suffers against the baselines on the traditional LibriSpeech dataset. However, performance on larger test-sets such as People's Speech outperforms the Qwen2.5-Omni baseline. Kimi-Audio performs best overall on English and Mandarin speech recognition, which is likely stemming from its large pretraining using English and Chinese ASR data. However, MiDashengLM and Qwen2.5-Omni are both capable of ASR on different languages such as Indonesian, Vietnamese and Thai. This suggests our encoder, despite no speech-specific training, develops surprisingly robust multilingual capabilities.

Table 10: Results for common ASR benchmarks. Results denoted with ">100" represent unsupported language, where the corresponding model only outputs English. All results represent WER/CER, where lower is better and the best result is displayed in bold.

Dataset	Language	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
LibriSpeech test-clean	English	3.7	1.7	**1.3**
LibriSpeech test-other		6.2	3.4	**2.4**
People's Speech		27.8	28.6	**22.3**
AISHELL2 Mic	Chinese	3.2	**2.5**	2.7
AISHELL2 iOS		2.9	**2.6**	**2.6**
AISHELL2 Android		3.1	2.7	**2.6**
GigaSpeech 2	Indonesian	**20.8**	21.2	>100
	Thai	**36.9**	53.8	>100
	Viet	**18.1**	18.6	>100

5.3 Question answering results

Question answering (QA) performance results are presented in Table 11. On closed QA benchmarks (MMAU [56] and MuChoMusic [57]), MiDashengLM achieves superior performance with accuracies of 71.35% and 66.30%, respectively, outperforming all baseline models. This advantage extends to open QA tasks (MusicQA, AudioCaps-QA), where MiDashengLM maintains its leading position while Kimi-Audio-Instruct demonstrates the weakest performance, which is consistent with earlier captioning benchmark observations.

Table 11: Results for question-answering datasets. For all results higher is better and best result are in bold.

Dataset	Subset	Metric	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
MuChoMusic [57]		ACC ↑	**71.35**	64.79	67.40
MMAU [56]	Sound	ACC ↑	68.47	67.87	**74.17**
	Music		66.77	**69.16**	61.08
	Speech		**63.66**	59.76	57.66
	Average		**66.30**	65.60	64.30
MusicQA [58]		FENSE ↑	**62.35**	60.60	40.00
AudioCaps-QA [59]			**54.31**	53.28	47.34

5.3.1 MECAT-QA

Lastly, we evaluate MiDashengLM on our proposed MECAT-QA dataset, a part of the publicly available MECAT benchmark [43]. The dataset is a open QA dataset, which we evaluate using FENSE. As the results in Table 12 show, our proposed MiDashengLM outperforms the baselines by a significant margin on the MECAT-QA dataset.

Table 12: Results for MECAT-QA. Results represent FENSE, where higher is better and best result are in bold.

Task	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Direct Perception	**65.89**	49.65	37.45
Sound Characteristics	**62.10**	43.81	32.48
Quality Assessment	**61.76**	40.47	19.24
Environment Reasoning	**63.02**	44.09	37.53
Inference & Judgement	**59.57**	42.50	38.83
Application Context	**60.12**	41.92	33.82
Average	**62.08**	43.74	33.22

5.4 Inference speed

A key advantage of MiDashengLM lies in its computational efficiency, encompassing both training speed (discussed in Section 1.1) and inference performance. In this experiments, we compare MiDashengLM with Qwen25-Omni-7B, as they utilizie the same text decoder backbone. We provide results in regards to Time to first token (TTFT) latency and theoretical computation Giga Multiply-Add Operations per Second (GMACs), where results are displayed in Figure 7. As shown in Figure 7, MiDashengLM achieves significantly lower TTFT than the baseline. We observe a speed improvement of up to 4× (160ms vs. 40ms) in regards to TTFT. Further throughput analysis in Table 13 reveals a 3.2× speedup at comparable batch sizes and an overall potential speedup of 20.2× with larger batches. These improvements stem from the better support for variable length inputs provided by Dasheng, as well as the optimized 5 Hz audio feature processing.

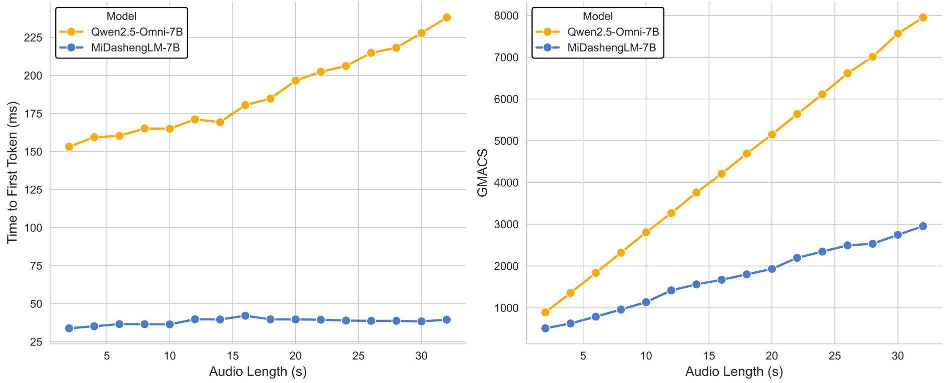


Figure 7: Time to first token (TTFT) and Giga Multiply-Add Operations per Second (GMACs) comparison between MiDashengLM-7B and Qwen2.5-Omni-7B.

Table 13: Throughput (samples/s) speed Comparison of MiDashengLM-7B and Qwen2.5-Omni-7B. Evaluation is done on a GPU with 80GB memory using bfloat16 for activations and parameters. All audio inputs are 30s long and output lengths are fixed to 100 tokens. OOM represents out of memory.

Batch Size	MiDashengLM 7B	Qwen2.5-Omni 7B	Speedup
1	0.65	0.45	1.4×
4	2.42	1.21	2.0×
8	4.67	1.44	3.2×
16	8.93	OOM	6.2×
32	14.36		10.0×
64	19.54		13.6×
128	24.26		16.8×
512	29.04		20.2×

6 Conclusion

We present MiDashengLM, an efficient large audio language model (LALM) that advances the state of general audio understanding through several key innovations. First, we introduce a novel training paradigm using general audio captioning, enabled by our newly created ACAVCaps dataset and MECAT evaluation benchmark. This framework facilitates effective audio-text alignment, as demonstrated by our pretrained Dasheng-based encoder outperforming Whisper-Large V3 on 18 of 22 tasks in the X-Ares benchmark evaluation. Notably, MiDashengLM achieves its strong performance while maintaining remarkable efficiency. Trained exclusively on publicly available audio-text data, our model competes favorably against closed-source/closed-data alternatives (Qwen2.5-Omni and Kimi-Audio) across multiple domains including audio captioning, closed question answering, open question answering, sound event detection, and paralinguistic tasks. The model's computational advantages are particularly significant, delivering up to 20.2× faster inference speeds and up to 4 × reduced time-to-first-token latency compared to baseline approaches.

References

[1] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, “Pengi: An audio language model for audio tasks,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[2] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, “Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities,” *arXiv preprint arXiv:2402.01831*, 2024.

[3] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, “Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities,” *arXiv preprint arXiv:2406.11768*, 2024.

[4] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, “Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities,” *arXiv preprint arXiv:2503.03983*, 2025.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[6] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, “Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing,” *arXiv preprint arXiv:2110.07205*, 2021.

[7] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, “Google usm: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.

[8] Y. Peng, S. Muhammad, Y. Sudo, W. Chen, J. Tian, C.-J. Lin, and S. Watanabe, “Owsm v4: Improving open whisper-style speech models via data scaling and cleaning,” *arXiv preprint arXiv:2506.00338*, 2025.

[9] M. AI, “Kimi-audio technical report,” *arXiv preprint arXiv:2504.18425*, 2025. [Online]. Available: https://arxiv.org/abs/2504.18425

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

[11] M. Wu, H. Dinkel, and K. Yu, “Audio caption: Listen and tell,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 830–834.

[12] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: An audio captioning dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[13] K. Drossos, S. Adavanne, and T. Virtanen, “Automated audio captioning with recurrent neural networks,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 374–378.

[14] C. D. Kim, B. Kim, H. Lee, and G. Kim, “Audiocaps: Generating captions for audios in the wild,” in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:174799768

[15] I. Martin and A. Mesaros, “Diversity and bias in audio captioning datasets,” in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 90–94.

[16] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, “Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research,” *arXiv preprint arXiv:2303.17395*, 2023.

[17] L. Sun, X. Xu, M. Wu, and W. Xie, “Auto-acd: A large-scale dataset for audio-language representation learning,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5025–5034.

[18] Y. Yuan, D. Jia, X. Zhuang, Y. Chen, Z. Liu, Z. Chen, Y. Wang, Y. Wang, X. Liu, X. Kang *et al.*, “Sound-vecaps: Improving audio generation with visual enhanced captions,” *arXiv preprint arXiv:2407.04416*, 2024.

[19] J. Bai, H. Liu, M. Wang, D. Shi, W. Wang, M. D. Plumbley, W.-S. Gan, and J. Chen, “Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models,” *arXiv preprint arXiv:2411.18953*, 2024.

[20] S. Chen, X. Xie, Z. Chen, L. Zhao, O. Lee, Z. Su, Q. Sun, and B. Wang, “Fusionaudio-1.2 m: Towards fine-grained audio captioning with multimodal contextual fusion,” *arXiv preprint arXiv:2506.01111*, 2025.

[21] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[22] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[23] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “Fsd50k: an open dataset of human-labeled sound events,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[24] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle *et al.*, “Audio flamingo 3: Advancing audio intelligence with fully open large audio language models,” *arXiv preprint arXiv:2507.08128*, 2025.

[25] F. Retkowski, M. Züfle, A. Sudmann, D. Pfau, J. Niehues, and A. Waibel, “From speech to summary: A comprehensive survey of speech summarization,” *arXiv preprint arXiv:2504.08024*, 2025.

[26] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, “Joint audio and speech understanding,” in *International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://arxiv.org/abs/2309.14405

[27] J. Bai, X. Chen, Y. Zhou, C.-C. Liu, Y. Chen, S. Huang, K. Chen, J.-F. Li, H. Lin, H. Zhou, L. Yang, Z. Li, Y. Wang, J. Lin, Y.-H. Zheng, Y. Chen, C. Zhang, X. Lu, X. Xu, X. Zhao, W. Han, C. Wang, Y. Hu, J. Lu, H. Chen, P. Lv, W. Liu, W. Dai, and M. Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2403.02422*, 2024. [Online]. Available: https://qwen-audio.github.io/Qwen-Audio/

[28] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, “Qwen2-audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024.

[29] Z. Xie and C. Wu, “Mini-omni: Language models can hear, talk while thinking in streaming,” *arXiv preprint arXiv:2408.16725*, 2024.

[30] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, “Llama-omni: Seamless speech interaction with large language models,” *arXiv preprint arXiv:2409.06666*, 2024.

[31] G. Li, J. Liu, H. Dinkel, Y. Niu, J. Zhang, and J. Luan, “Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering,” *arXiv preprint arXiv:2503.11197*, 2025.

[32] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, “Salmonn: Towards generic hearing abilities for large language models,” *arXiv preprint arXiv:2310.13289*, 2023.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[34] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[35] Y. Gong, Y.-T. Chen, H. Xu, Y.-S. Liu, M. Cui, Z. Yu, J.-L. Qin, M.-M. Zhang, L.-B. Zhou, W.-X. Lin, H.-M. Zhou, L.-F. Wang, and C.-Y. Xu, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01040*, 2021. [Online]. Available: https://arxiv.org/abs/2104.01040

[36] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[37] N. Jeffries, E. King, M. Kudlur, G. Nicholson, J. Wang, and P. Warden, "Moonshine: Speech recognition for live transcription and voice commands," *arXiv preprint arXiv:2410.15608*, 2024.

[38] A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Muddireddy *et al.*, "Voxtral," *arXiv preprint arXiv:2507.13264*, 2025.

[39] S. Lee, J. Chung, Y. Yu, G. Kim, T. Breuel, G. Chechik, and Y. Song, "Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 274–10 284.

[40] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 291–295.

[41] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[42] H. Dinkel, Z. Yan, T. Wang, Y. Wang, X. Sun, Y. Niu, J. Liu, G. Li, J. Zhang, and J. Luan, "Glap: General contrastive audio-text pretraining across domains and languages," *arXiv preprint arXiv:2506.11350*, 2025.

[43] Y. Niu, "Mecat: A multi-experts constructed benchmark for fine-grained audio understanding tasks," 2025.

[44] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[45] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos *et al.*, "The song describer dataset: a corpus of audio captions for music-and-language evaluation," *arXiv preprint arXiv:2311.10057*, 2023.

[46] S. Doh, K. Choi, J. Lee, and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," *arXiv preprint arXiv:2307.16372*, 2023.

[47] HuggingFace, "Laion-audio-300m," accessed: 2025-07-10. [Online]. Available: huggingface.co/datasets/laion/LAION-Audio-300M

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[49] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Interspeech 2024*, 2024, pp. 547–551.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[51] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[52] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[53] J. Bai, X. Chen, X. Lu, J. Lin, Y. Chen, J.-F. Li, Y. Wang, X. Xu, C. Zhang, Y.-H. Zheng, C. Wang, W. Han, J. Lu, H. Chen, P. Lv, W. Liu, W. Dai, and M. Zhou, "Qwen2.5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025. [Online]. Available: https://arxiv.org/abs/2503.20215

[54] J. Zhang, H. Dinkel, Y. Niu, C. Liu, S. Cheng, A. Zhao, and J. Luan, "X-ares: A comprehensive framework for assessing audio encoder performance," *arXiv preprint arXiv:2505.16369*, 2025.

[55] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 981–985.

[56] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," 2024. [Online]. Available: https://arxiv.org/abs/2410.19168

[57] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, "Muchomusic: Evaluating music understanding in multimodal audio-language models," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[58] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 286–290, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261064622

[59] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "Audiobench: A universal benchmark for audio large language models," *arXiv preprint arXiv:2406.16020*, 2024.

[60] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2015, pp. 5206–5210.

[61] W. Kang, X. Yang, Z. Yao, F. Kuang, Y. Yang, L. Guo, L. Lin, and D. Povey, "Libriheavy: A 50,000 hours asr corpus with punctuation casing and context," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2024, pp. 10 991–10 995.

[62] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[63] Y. Yang, Z. Song, J. Zhuo, M. Cui, J. Li, B. Yang, Y. Du, Z. Ma, X. Liu, Z. Wang *et al.*, "Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement," *arXiv preprint arXiv:2406.11546*, 2024.

[64] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 6182–6186.

[65] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, "Yodas: Youtube-oriented dataset for audio and speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).* IEEE, 2023, pp. 1–8.

[66] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[67] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA).* IEEE, 2017, pp. 1–5.

[68] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[69] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *ArXiv*, vol. abs/2010.11567, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225039887

[70] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[71] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[72] M. Mazumder, S. Chitlangia, C. Banbury, Y. Kang, J. M. Ciro, K. Achorn, D. Galvez, M. Sabini, P. Mattson, D. Kanter *et al.*, "Multilingual spoken words corpus," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[73] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi *et al.*, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in *2024 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2024, pp. 885–890.

[74] C. Wang, A. Wu, J. Gu, and J. Pino, "Covost 2 and massively multilingual speech translation." in *Interspeech*, vol. 2021, 2021, pp. 2247–2251.

[75] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2023, pp. 798–805.

[76] S. Li, Y. You, X. Wang, Z. Tian, K. Ding, and G. Wan, "Msr-86k: An evolving, multilingual corpus with 86,300 hours of transcribed audio for speech recognition research," *arXiv preprint arXiv:2406.18301*, 2024.

[77] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, 2019, pp. 2012–2017.

[78] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[79] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman *et al.*, "Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv preprint arXiv:2104.02014*, 2021.

[80] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *arXiv preprint arXiv:2111.09344*, 2021.

[81] Z. Tang, D. Wang, Y. Xu, J. Sun, X. Lei, S. Zhao, C. Wen, X. Tan, C. Xie, S. Zhou *et al.*, "Kespeech: An open source speech dataset of mandarin and its eight subdialects," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[82] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th ISMIR Conference, p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[83] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. Channing Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 366–370.

[84] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018.

[85] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[86] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, "Arca23k: An audio dataset for investigating open-set label noise," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[87] P. Primus, F. Schmid, and G. Widmer, "Tacos: Temporally-aligned audio captions for language-audio pretraining," *arXiv preprint arXiv:2505.07609*, 2025.

[88] I.-Y. Jeong and J. Park, "Cochlscene: Acquisition of acoustic scene data using crowdsourcing," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 17–21.

[89] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdset: A multi-task benchmark for classification in computational bioacoustics," in *NeurIPS Datasets and Benchmarks Track*, 2023.

[90] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," in *Language resources and evaluation*, vol. 42, no. 4, 2008, pp. 335–359.

[91] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[92] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, p. e0250173, 2021.

[93] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018.

[94] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[95] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Commun.*, vol. 137, pp. 1–18, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235254736

[96] Y. Gong, J. Yu, and J. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 151–155.

[97] M. M. Rashid, G. Li, and C. Du, "Nonspeech7k dataset: Classification and analysis of human non-speech sound," *IET Signal Processing*, vol. 17, no. 6, p. e12233, 2023.

[98] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[99] G. Sinisetty, P. Ruban, O. Dymov, and M. Ravanelli, "Commonlanguage," Jun. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5036977

[100] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the English accents in the British isles," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6532–6541. [Online]. Available: https://aclanthology.org/2020.lrec-1.804

[101] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*, 2017, pp. 2616–2620.

[102] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[103] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[104] Y. Lin, X. Qin, G. Zhao, M. Cheng, N. Jiang, H. Wu, and M. Li, "Voxblink: A large scale speaker verification dataset on camera," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 271–10 275.

[105] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," *arXiv preprint arXiv:2407.11510*, 2024.

[106] I. Yakovlev, A. Okhotnikov, N. Torgashov, R. Makarov, Y. Voevodin, and K. Simonchik, "Voxtube: a multilingual speaker recognition dataset," in *Proc. Interspeech*, 2023, pp. 2238–2242.

[107] F. Stöter, S. Chakrabarty, E. A. P. Habets, and B. Edler, "Libricount, a dataset for speaker count estimation (version v1.0.0)," https://doi.org/10.5281/zenodo.1216072, Apr. 2018, accessed on YYYY-MM-DD. [Online]. Available: https://doi.org/10.5281/zenodo.1216072

[108] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[109] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Proc. Interspeech 2021*, 2021.

[110] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[111] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset." in *ISMIR*, 2018, pp. 468–474.

[112] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," 2017.

[113] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine learning for music discovery workshop, international conference on machine learning (ICML 2019)*, 2019, pp. 1–3.

[114] G. Bandiera, O. R. Picas, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, “Good-sounds. org: A framework to explore goodness in instrumental sounds.” in *ISMIR*, 2016, pp. 414–419.

[115] X. Gong, Y. Zhu, H. Zhu, and H. Wei, “Chmusic: a traditional chinese music dataset for evaluation of instrument recognition,” in *Proceedings of the 4th international conference on big data technologies*, 2021, pp. 184–189.

[116] X. Liang, Z. Li, J. Liu, W. Li, J. Zhu, and B. Han, “Constructing a multimedia chinese musical instrument database,” in *Proceedings of the 6th Conference on Sound and Music Technology (CSMT)*. Singapore: Springer Singapore, 2019, pp. 53–60.

[117] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu, “Avqa: A dataset for audio-visual question answering on videos,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3480–3491.

[118] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, “Clotho-aqa: A crowdsourced dataset for audio question answering,” 08 2022, pp. 1140–1144.

[119] P. Pandey, R. V. Swaminathan, K. V. V. Girish, A. Sen, J. Xie, G. Strimel, and A. Schwarz, “Sift-50m: A large-scale multilingual dataset for speech instruction fine-tuning,” 2025. [Online]. Available: https://www.amazon.science/publications/sift-50m-a-large-scale-multilingual-dataset-for-speech-instruction-fine-tuning

A Data sources

A.1 Speech datasets

Table 14: Speech training data. The notation † leverages Whisper to generate automatic transcripts by the authors. The column "SFT ?" indicates whether the dataset is used for supervised finetuning. By default all data is used for pretraining.

Data	Task	Length (h)	SFT ?
LibriSpeech [60]	ASR	960	✓
LibriHeavy [61]	ASR	50,000	✗
GigaSpeech [62]	ASR	10,000	✓
GigaSpeech 2 [63]	ASR	30,000	✓
WeNetSpeech [64]	ASR	10,000	✓
YODAS [65]	ASR	320,000	✗
CommonVoice-17.0 [66]	ASR	5,000	✓
AISHELL-1 [67]	ASR	100	✓
AISHELL-2 [68]	ASR	1,000	✓
AISHELL-3 [69]	ASR	70	✓
LJSpeech-1.1 [70]	ASR	37	✗
LibriTTS [71]	ASR	585	✗
MultiLingualSpokenWords [72]	KWS	5,000	✗
Emilia [73]	ASR	101,000	✓
CovoST-v2 [74]	S2TT	2,880	✓
Fleurs [75]	S2TT	1,224	✗
MSR-86K [76]	ASR, LangID	86,000	✓
ACAV100M-Speech† [39]	ASR	55,754	✗
Must-C [77]	ASR, S2TT	1,000	✓
MLS [78]	ASR	50,000	✗
SpgiSpeech [79]	ASR	5,000	✗
People's Speech [80]	ASR	30,000	✗
KeSpeech [81]	ASR	1,400	✓
LAION-Audio-300M [47]	Caption	230,000	✗
Total		997,010	258,410

A.2 Sound and general audio datasets

Table 15: General Sound and Audio Datasets. ACAVCaps is utilized for audio-text alignment. The column "SFT ?" indicates whether the dataset is used for supervised finetuning. By default all data is used for pretraining.

Dataset	Task	Length (h)	SFT ?
FSD50k [82]	Sound Event	77	✓
AudioSet [21]		5,200	✓
AudioSet-strong [83]		220	✗
VGGSound [22]		540	✓
FSDKaggle2018 [84]		20	✓
FSDKaggle2019 [85]		100	✓
ARCA23k [86]		120	✗
AutoACD [17]	Audio (Sound) Caption	5,200	✓
AudioSetCaps [19]		6,000	✓
SoundVECaps [18]		5,000	✓
WavCaps [16]		7,567	✓
Audiocaps [14]		100	✓
Clothov2 [12]		17	✓
TACOS [87]		98	✓
CochlScene [88]	SoundScape	500	✓
BirdSet [89]		7,000	✗
ACAVCaps	General Caption	38,662	✓
Total		76,421	69,081

A.3 Speech and paralinguistic datasets

Table 16: Speech and sound paralinguistic datasets. The column "SFT ?" indicates whether the dataset is used for supervised finetuning. By default all data is used for pretraining.

Dataset	**Task**	**Length (hours)**	SFT ?
IEMOCAP [90]	Emotion	8	✓
Meld [91]		12	✓
SUBESCO [92]		9	✗
RAVDESS-Speech [93]		2	✗
RAVDESS-Song [93]		1	✗
CREMA-D [94]		4	✗
ESD [95]		29	✗
VocalSound [96]	Vocal Sound classification	20	✓
NonSpeech7k [97]		3	✓
VoxLingua107 [98]	Language Identification	7,200	✓
CommonLanguage [99]		45	✓
YLACombe [100]		5	✗
VoxCeleb1 [101]	Speaker verification	76	✓
CNCeleb [102]	Speaker verification	2,100	✓
VoxCeleb2 [103]	Speaker age Speaker verification Gender classification	1,000	✓
VoxBlink1 [104]	Speaker verification	1,300	✓
VoxBlink2 [105]	Speaker verification	2,600	✓
VoxTube [106]	Speaker verification Language Identification Gender classification	5,200	✓
LibriCount [107]	Speaker counting	8	✓
FluentSpeechCommands [108]	Intent Classification	17	✗
speechocean762 [109]	Gender Speaker age	5	✗
ASVSpoof5 [110]	Spoof detection	603	✗
Total		20,247	19,572

A.4 Music Datasets

Table 17: Music-Related Datasets Overview. The column "SFT ?" indicates whether the dataset is used for supervised finetuning. By default all data is used for pretraining.

Dataset	Task	Length (h)	SFT ?
MusicCaps [44]	Music Caption	15	✓
Songdescriber [45]		23	✓
LPMusicCaps-MTT [46]		18	✓
LPMusicCaps-MSD [46]		1,000	✓
VocalSet [111]	Singing style identification	10	✗
FreeMusicArchive [112]	Genre recognition	610	✓
MTG-Jamendo [113]	Instrument classification Genre recognition	3,768	✓
NSynth [112]	Instrument classification	360	✓
GoodSounds [114]		28	✓
chMusic [115]		1	✓
CTIS [116]		1	✓
Total		5,824	5,814

A.5 Question Answering Datasets

Table 18: Question answering datasets used in this work. Datasets denoted with † have been modified from their original dataset by using an LLM to change captions into question-answer pairs. We display the number of questions and answers in each dataset as # QA. The column "SFT ?" indicates whether the dataset is used for supervised finetuning. By default only AVQA, MusicQA ad ClothoAQA are used during pretraining.

Dataset	Task	# QA	SFT ?
AVQA [117]	Environment QA	36,114	✓
ClothoAQA [118]		6175	✓
TACOS† [87]		40,019	✓
MusicQA [58]	Music QA	112,878	✓
SIFT-50M [119] (closed)	Speech QA	21,430,000	✓
ACAV-QA†	General QA	24,371	✓

B Contributors

Contributors are listed in Alphabetical order.

Heinrich Dinkel
Gang Li
Jizhong Liu
Jian Luan
Yadong Niu
Xingwei Sun
Tianzi Wang
Qiyang Xiao
Junbo Zhang
Jiahao Zhou