CVSSP
Centre for Vision,
Speech and Signal
Processing

UNIVERSITY OF SURREY

# Large Audio-Language Models and Applications

**Wenwu Wang**

Centre for Vision, Speech and Signal Processing (CVSSP)
&
Surrey Institute for People Centred AI

**University of Surrey, UK**

Email: w.wang@surrey.ac.uk
Web: https://personalpages.surrey.ac.uk/w.wang/

Invited talk on the ASTAR AI Symposium, Singapore, 27 January 2026

# Many thanks to…

- **Xinhao Mei**
- **Haohe Liu**
- **Xubo Liu**
- **Jinhua Liang**
- **Yi Yuan**
- **Yiming Zhang**
- **Qiuqiang Kong**
- **Yuelan Cheng**
- **Jisheng Bai**
- **Zehua Chen**
- **Xinran Liu**
- **Mark Plumbley**
- Turab Iqbal
- Jianyuan Sun
- Jian Guan
- Feiyang Xiao
- Yong Xu
- Yin Cao
- Jinzheng Zhao
- Yuxuan Wang
- Qiaoxi Zhu
- Volkan Kilic
- Zhanyu Ma
- Danilo Mandic
- Philip Jackson
- Qiang Huang
- David Frohlich
- Emily Corrigan-Kavanagh
- Marc Green
- Andres Fernandes
- Christian Kroos
- Trevor Cox
- Arshdeep Singh
- …

Funding from:
UK Engineering and Physical Sciences Research Council (EPSRC) &
British Council Newton Institutional Links Award

# Outline

- **Introduction**
- **Large audio models & large language models**
- **Large audio-language models**
  - Example models & datasets
  - Typical methods for fusing/aligning audio and language
  - Open challenges
  - Integration of audio and language models
- **Examples of LALMs for various cross-modal generation tasks**
  - Audio captioning (e.g. audio to text generation)
  - Audio question answering and reasoning
  - Text to audio generation & storytelling
  - LLMs for controllable audio editing
  - Neural audio codecs
- **Conclusions and future works**

# Audio Signal Processing

**Tasks:**

- Audio source separation
- Audio source localisation/tracking
- Audio event detection/localisation
- Audio scene classification
- Audio tagging
- Audio search and retrieval
- Audio rendering
- Audio recognition
- …

**Models:**

- Physics-based models
- Perceptually motivated models
- Data-driven models
- Hybrid models
- ….

**Data:**

- Audio-only
- Multimodal (audio, visual, texts, EEG, etc)

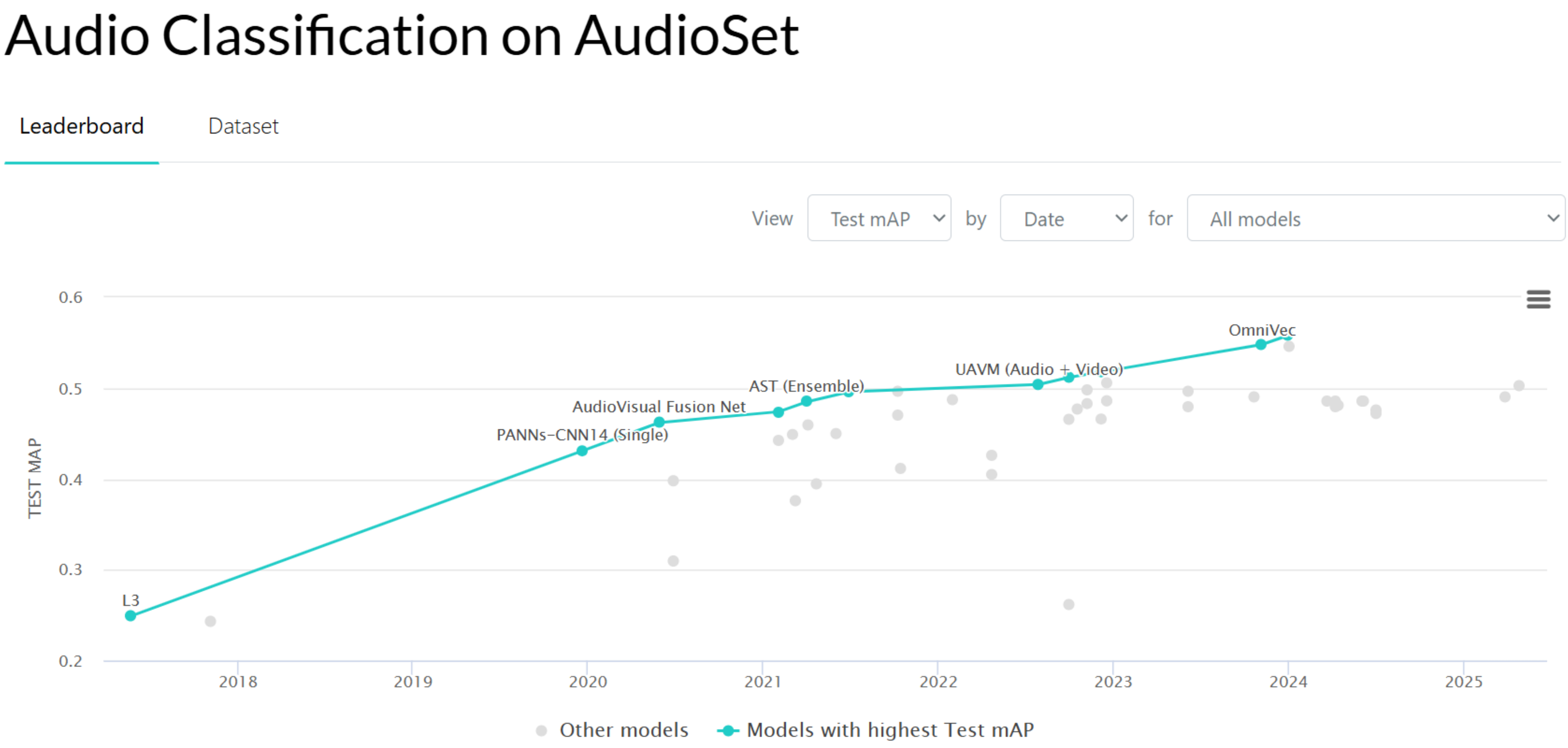
# (Large) Audio Models

Learning **general/universal audio representations** from large scale audio data shows promising performance in downstream tasks (classification, separation, retrieval, etc):

- **PANNs** (Kong, et al, 2020): large scale CNN-based audio model
- Audio2Vec (Tagliasacchi et al, 2020): sequence to sequence unsupervised model
- CLAR (Al-Tahan and Mohsenzadeh, 2021): self-supervised model
- COLA (Saeed et al, 2021): self supervised model
- BOYL-A (Niizumi et al, 2021): self supervised model
- AST (Gong et al, 2021): large scale transformer-based audio model
- ATST (Li and Li, 2022): transformer-based model
- MAE-AST (Baade et al, 2022): self-supervised model
- SSAST (Gong et al, 2022): self-supervised AST model
- Audio-MAE (Huang et al, 2022): self-supervised model
- BEATs (Chen et al, 2023): Audio pre-training with acoustic tokenizers
- **ASiT** (Atito et al, 2024): self-supervised models for general audio representation

mean Average Precision (mAP) on AudioSet: **31.4** (2017) -> **50.6** (2025)

# (Large) Audio Models



## Audio Classification on AudioSet

Leaderboard Dataset

View Test mAP by Date for All models

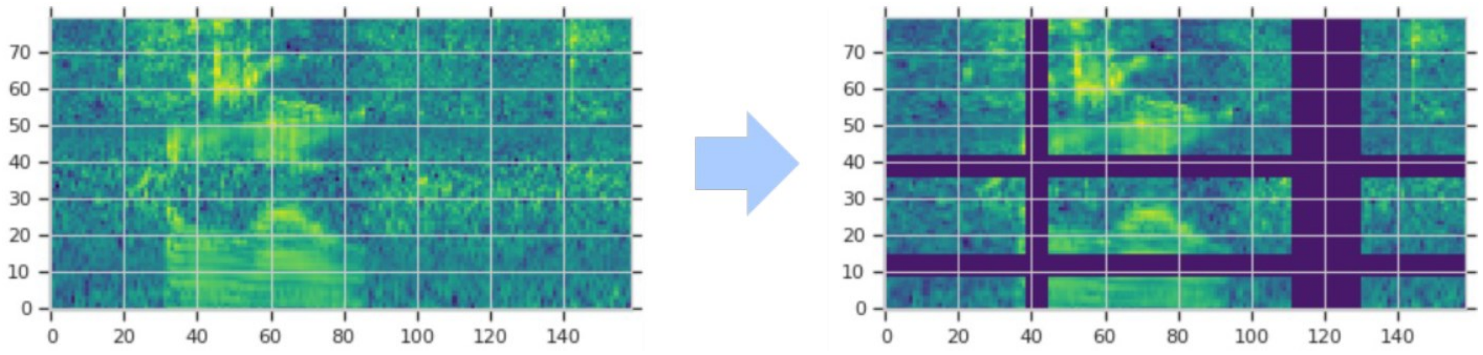
**Leaderboard**:
[https://paperswithcode.com/sota/audio-classification-on-audioset](https://paperswithcode.com/sota/audio-classification-on-audioset)
[https://www.codesota.com/audio/classification](https://www.codesota.com/audio/classification)

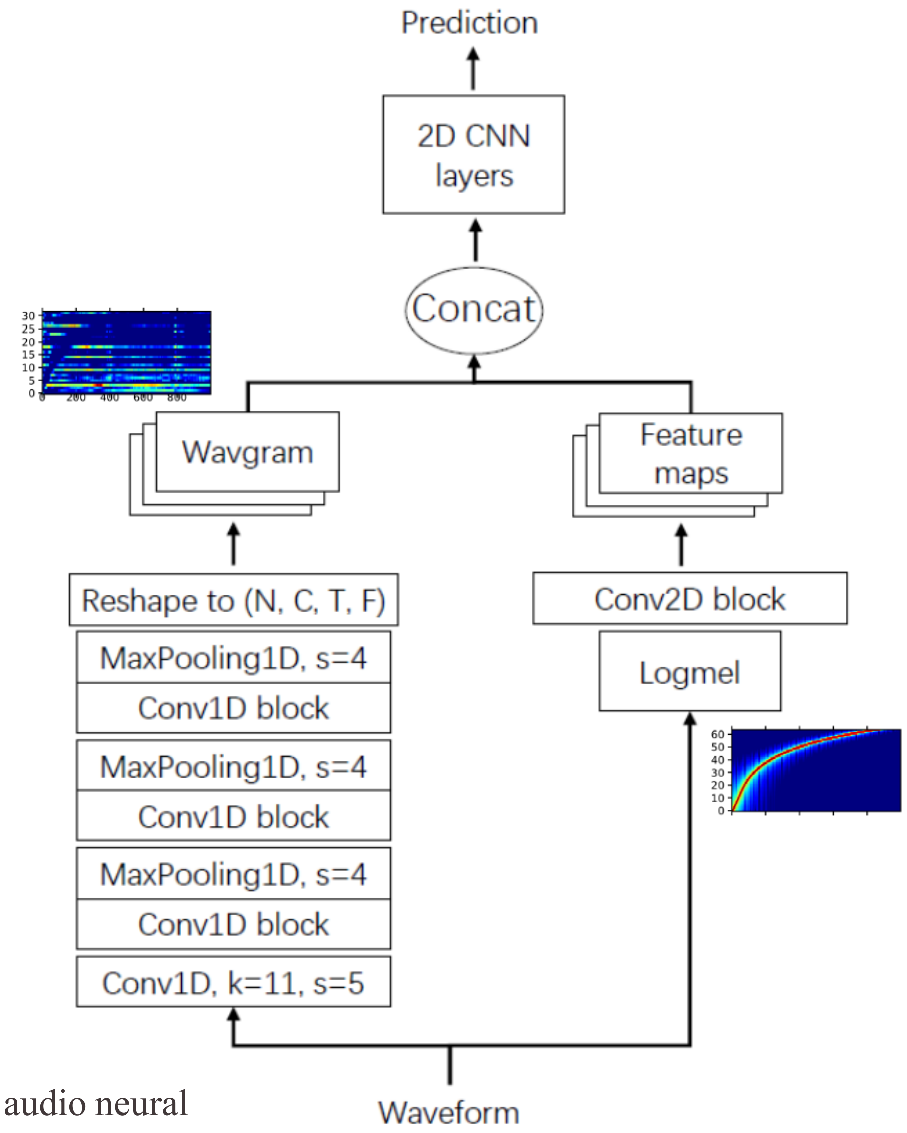
# PANNs: Large-Scale Pre-trained Audio Neural Networks

Wavegram-Logmel-CNN for AudioSet tagging

- Time-domain (“Wavegram”), plus
- Log mel spectrogram

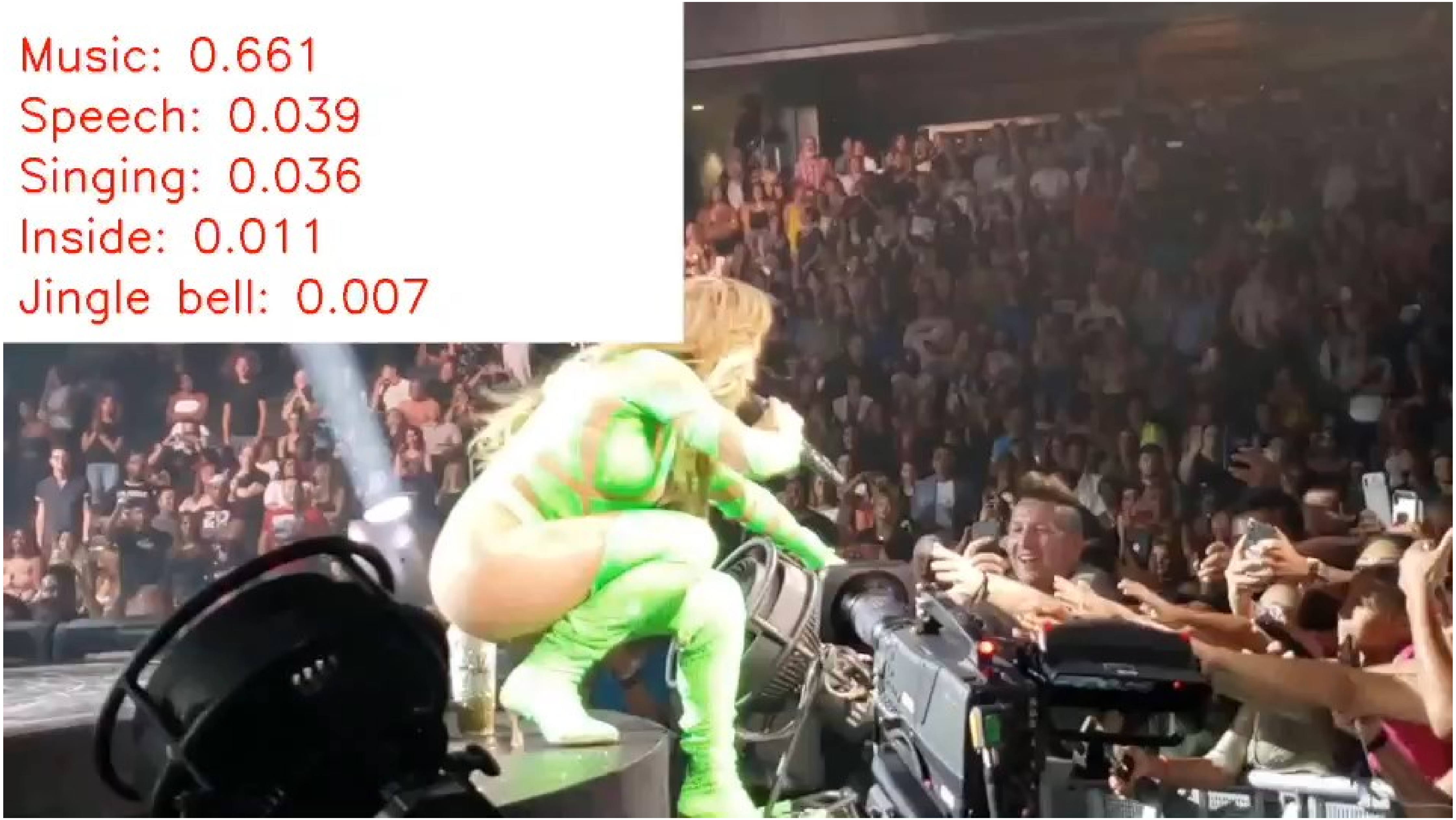
Data augmentation, e.g. use SpecAugment: randomly mask time and frequency stripes of log mel spectrogram





Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: large-scale pretrained audio neural networks for audio pattern recognition", *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2020.

# PANNs: Demo



Music: 0.661

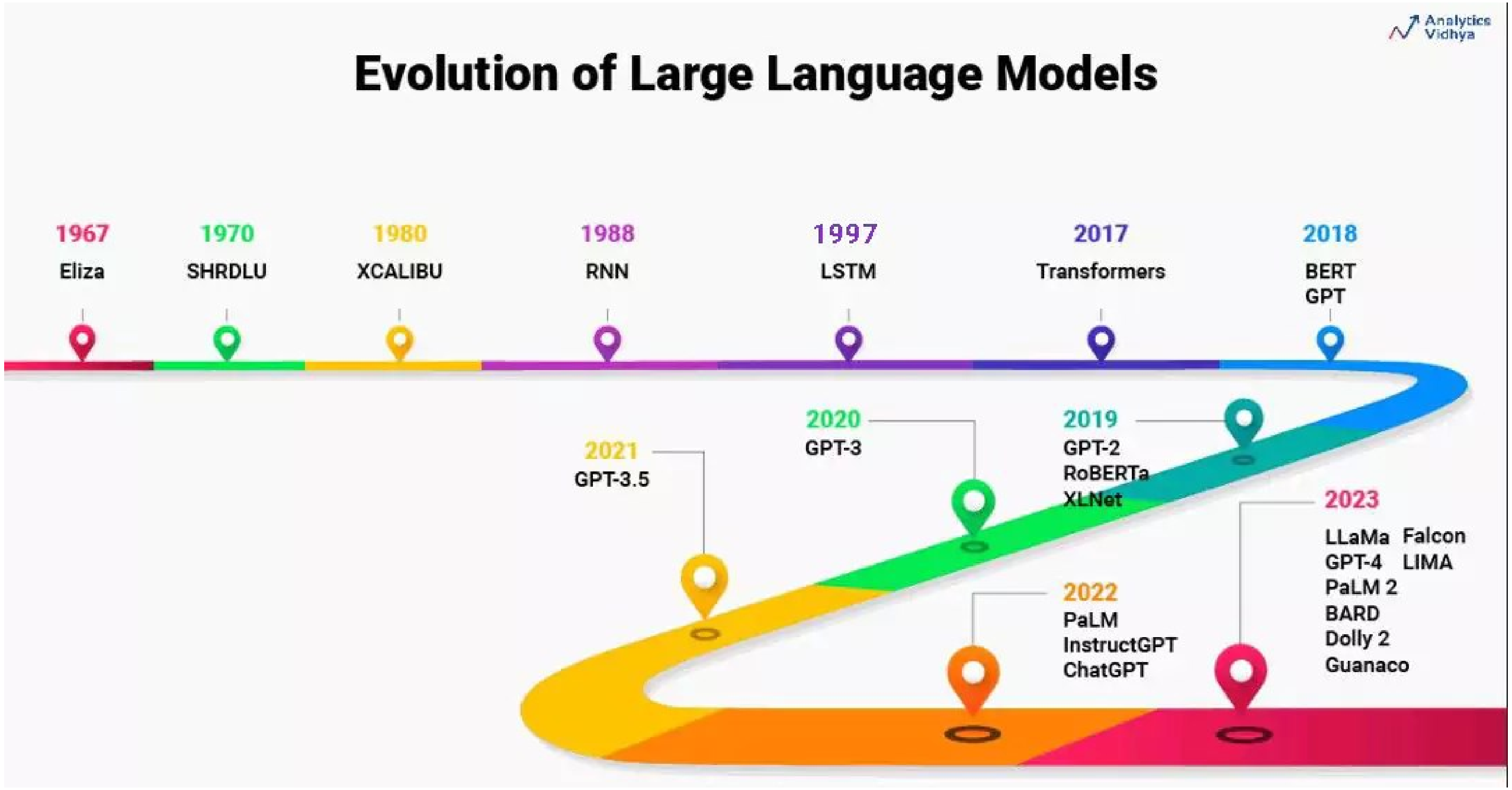
Speech: 0.039

Singing: 0.036

Inside: 0.011

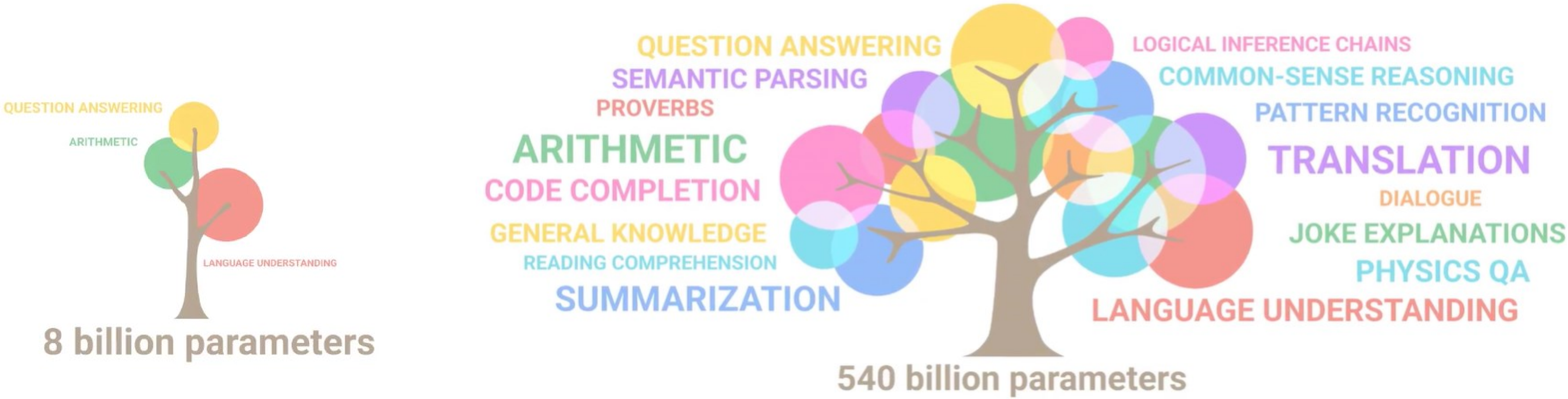
Jingle bell: 0.007

# Large Language Models (LLMs)



Courtesy to Aravind Pai: https://www.analyticsvidhya.com/blog/2023/07/beginners-guide-to-build-large-language-models-from-scratch/

# Large Language Models (LLMs)



Figures from Ryan O'Connor: https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/

Wei et al, “Emergent abilities of large language models,” *Transactions on Machine Learning Research*, 2022.

# Large Audio-Language Models: Why?

- Leverage knowledge within LLMs to address the limitations of audio models
  - Zero-shot or few-shot classification
- Explore homogeneity across tasks with LLMs
  - Use LLMs as an agent to solve multi-task problems
- Extend the capabilities of audio models for new tasks
  - Extending from audio classification to audio captioning/question answering & reasoning
  - Extending from audio generation to storytelling & controllable editing
  - Extending from audio source separation to language queried audio source separation

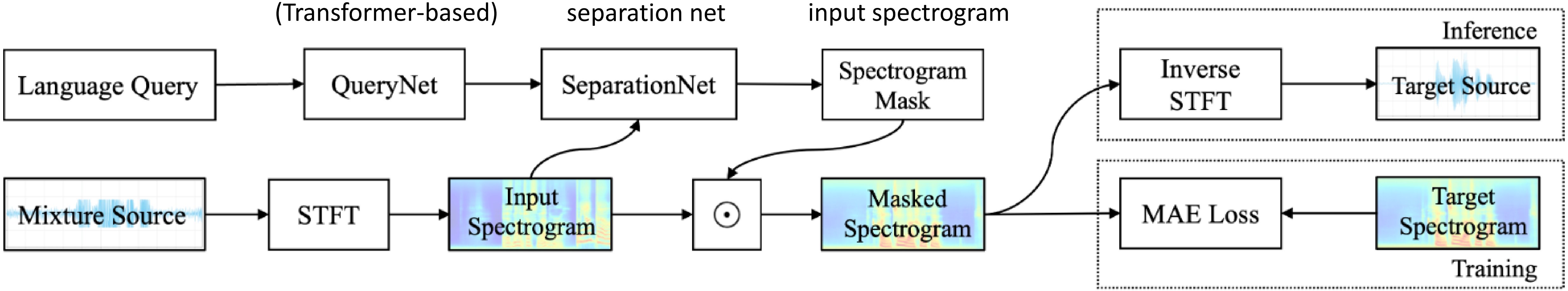
# An Example: Language-Queried Audio Source Separation

## Use language query to extract target source

BERT-mini (Transformer-based)

ResUNet-based separation net

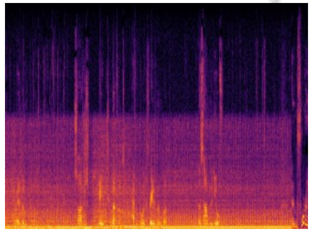
Same shape as input spectrogram



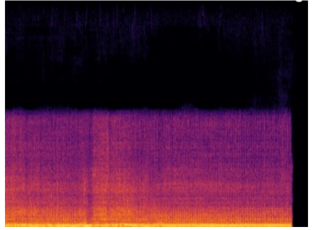
Mean absolute error for training

Human query: “The engine sound of a vehicle”

Mix

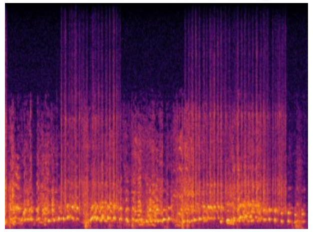


Separated

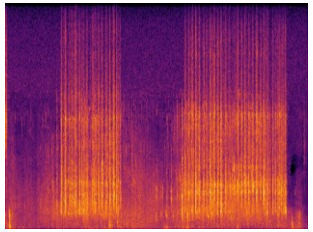


Human query: “The sound of hitting the keyboard”

Mix



Separated



X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M.D. Plumbley, and W. Wang,"Separate What You Describe: Language-Queried Audio Source Separation," in *Proc. 23rd Interspeech Conference* (INTERSPEECH 2022), 18-22 September, 2022, Incheon, Korea.

# Large Audio-Language Models - Examples

**Whisper**: automatic speech recognition (ASR)
**Wav2Vec 2.0**: speech to text (STT)
**DeepSpeech**: open-source ASR
**Coqui AI**: speech synthesis and TTS
**Jasper and QuartzNet**: ASR
**GPT-3 with Whisper**: ASR + LLMs
**Sonix.ai**: speech transcription and analysis
**SpeechBrain**: platform for speech models
**OpenSTT**: ASR
**Gemini**: text/image/speech
**WavLLM**: speech LLMs
**MuseNet**: music instrumental composition & style transfer
**MusicVAE**: melody generation, remixing, and style interpolation
**JukeBox**: raw music with vocals and lyrics
**Riffusion**: text to music generation
**MusicLM**: text to music generation
**REMI**: symbolic music generation (e.g. MIDI)
**DeepBach**: classic music composition
**AIVA**: music generation assistant

**Wav2CLIP**: audio and language mapping
**AudioCLIP**: audio-text-image alignment
**CLAP**: audio-text alignment
**CLAP-LAION**: audio-text alignment
**Pengi**: audio classification & AQA
**Qwen-Audio**: speech, music, general audio
**AudioLM**: Speech/music generation
**LTU**: audio QA and reasoning
**SALMONN**: speech-audio-music LLMs
**ImageBind**: image, text, audio, depth
**ONE-PEACE**: audio-text-image
**AudioGPT**: Speech, Music, Talking Head
**UniAudio**: speech/sound/music/singing

**AudioLDM**: text to audio generation
**AudioLDM 2**: text to audio generation
**Re-AudioLDM**: text to audio generation
**T-CLAP**: audio-text alignment
**WavCraft**: text prompted audio editing
**APT-LLM**: LLM based AQA and reasoning

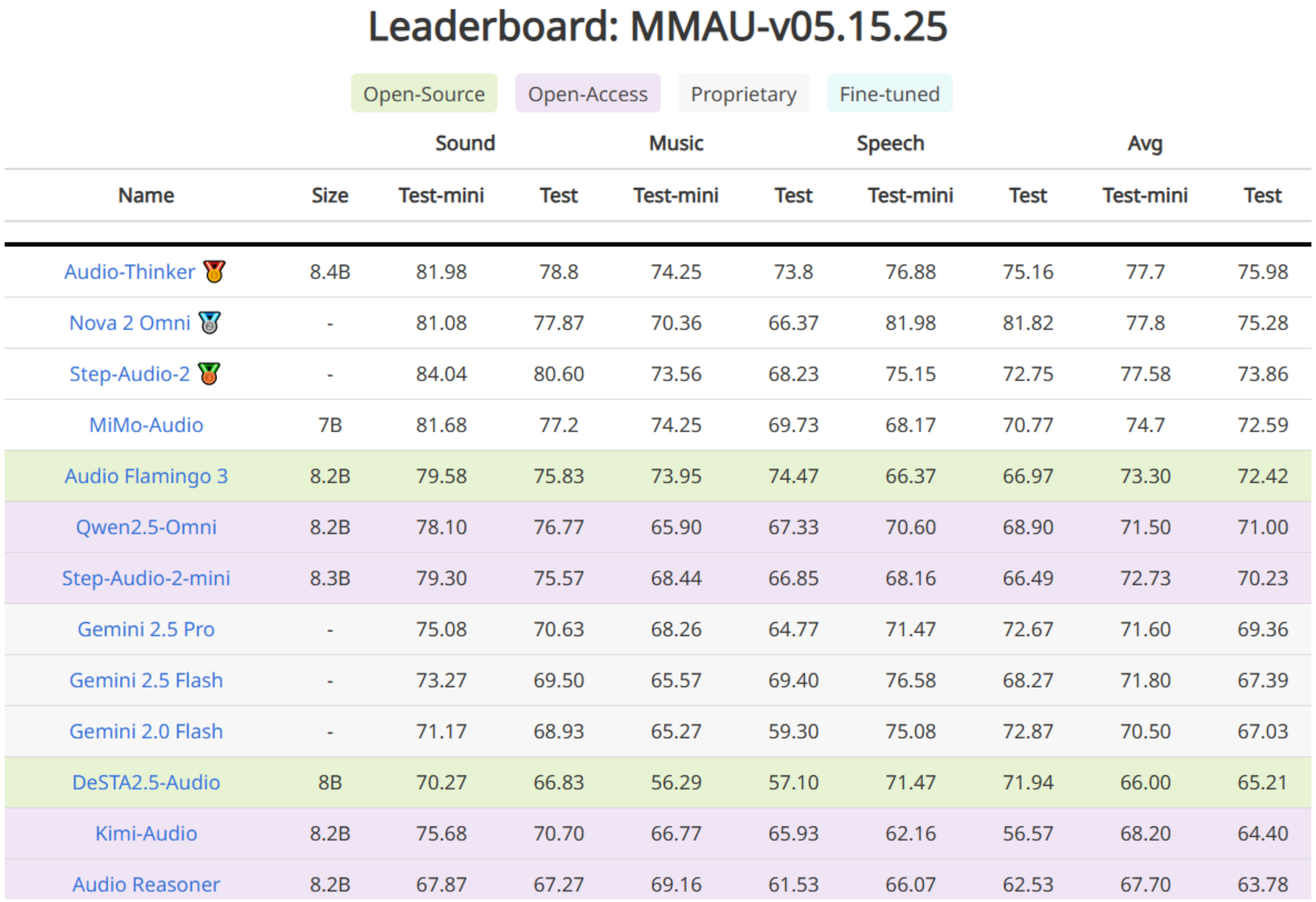
# Large Audio-Language Datasets

- AudioCaps (Kim et al, 2019)
- Clotho (Drossos et al, 2020)
- SoundDescs (Koepke et al, 2021)
- LAION-Audio-630K (Wu et al, 2023)
- Auto-ACD (Xu et al. 2024)

- **WavCaps (Mei et al, 2023)**
- **Sound-VECaps (Yuan et al, 2024)**
- **AudioSetCaps (Bai et al, 2025)**

- CLEAR (Abdelnour et al, 2018)
- DAQA (Fayek and Johnson, 2019)
- Clotho-AQA (Lipping et al, 2022)
- MUSIC-AVQA (Li et al, 2022)
- mClothoAQA (Behera et al, 2023)
- OpenAQA-5M (Gong et al, 2023)

# Large Audio-Language Models- Recent Progress



## Leaderboard: MMAU-v05.15.25

Open-Source | Open-Access | Proprietary | Fine-tuned

| | | Sound | | Music | | Speech | | Avg | |
|---|---|---|---|---|---|---|---|---|---|
| Name | Size | Test-mini | Test | Test-mini | Test | Test-mini | Test | Test-mini | Test |
| Audio-Thinker 🥇 | 8.4B | 81.98 | 78.8 | 74.25 | 73.8 | 76.88 | 75.16 | 77.7 | 75.98 |
| Nova 2 Omni 🥈 | - | 81.08 | 77.87 | 70.36 | 66.37 | 81.98 | 81.82 | 77.8 | 75.28 |
| Step-Audio-2 🥉 | - | 84.04 | 80.60 | 73.56 | 68.23 | 75.15 | 72.75 | 77.58 | 73.86 |
| MiMo-Audio | 7B | 81.68 | 77.2 | 74.25 | 69.73 | 68.17 | 70.77 | 74.7 | 72.59 |
| Audio Flamingo 3 | 8.2B | 79.58 | 75.83 | 73.95 | 74.47 | 66.37 | 66.97 | 73.30 | 72.42 |
| Qwen2.5-Omni | 8.2B | 78.10 | 76.77 | 65.90 | 67.33 | 70.60 | 68.90 | 71.50 | 71.00 |
| Step-Audio-2-mini | 8.3B | 79.30 | 75.57 | 68.44 | 66.85 | 68.16 | 66.49 | 72.73 | 70.23 |
| Gemini 2.5 Pro | - | 75.08 | 70.63 | 68.26 | 64.77 | 71.47 | 72.67 | 71.60 | 69.36 |
| Gemini 2.5 Flash | - | 73.27 | 69.50 | 65.57 | 69.40 | 76.58 | 68.27 | 71.80 | 67.39 |
| Gemini 2.0 Flash | - | 71.17 | 68.93 | 65.27 | 59.30 | 75.08 | 72.87 | 70.50 | 67.03 |
| DeSTA2.5-Audio | 8B | 70.27 | 66.83 | 56.29 | 57.10 | 71.47 | 71.94 | 66.00 | 65.21 |
| Kimi-Audio | 8.2B | 75.68 | 70.70 | 66.77 | 65.93 | 62.16 | 56.57 | 68.20 | 64.40 |
| Audio Reasoner | 8.2B | 67.87 | 67.27 | 69.16 | 61.53 | 66.07 | 62.53 | 67.70 | 63.78 |

https://sakshi113.github.io/mmau_homepage/#leaderboard-v15-parsed

# Trends and Open Questions in LALMs

**Open challenges:**

- **Fusion of audio and language models**: aligning/fusing audio-text data
- **Applications to audio tasks**: addressing existing challenges in audio tasks
- **Extending to multi-modality**: text, audio, visual, or more modalities
- **Data scarcity**: audio-language dataset shortage for building audio-language models
- **Extending to multi-tasks**: exploring in new tasks & solving multi-task problems
- **Multi-lingual models and datasets**: lacking multi-lingual models and datasets
- **Real-time streaming**: demands for real-time processing for applications in live captioning, gaming, and customer service
- **Low-resource language support**: growing interest in training models for underrepresented languages for inclusiveness
- **Safety issues**: growing concerns about toxicity and privacy issues
- **Explaining models**: explaining and interpreting different choices and decisions made by the model
- **Object hallucination**: struggling in answering discriminative questions related to the identification of specific object sounds within an audio clip
- **Performance evaluations**: lack of common evaluation protocols, tools and benchmarks

**Applications**:

- Accessibility, voice assistants, content creation, and human-computer interaction

# Typical Methods for Fusing Audio and Language

- Aligning audio-texts with contrastive pretraining
  - Examples: CLAP, CLAP-LAION, AudioCLIP, Wav2CLIP, T-CLAP, etc.

- Tokenizing audio and texts, then followed by LLMs
  - Examples: Moshi, VITA, LSLM, Voicebox, FunAudioLLM, LauraGPT, etc.

- Fusing embeddings with cross-attention
  - Examples: Q-former

- Cascading acoustic models with LLMs
  - Examples: naive ASR+LLM+TTS

- Combination of the above schemes
  - Examples: SPIRIT-LM, Spectron, etc.

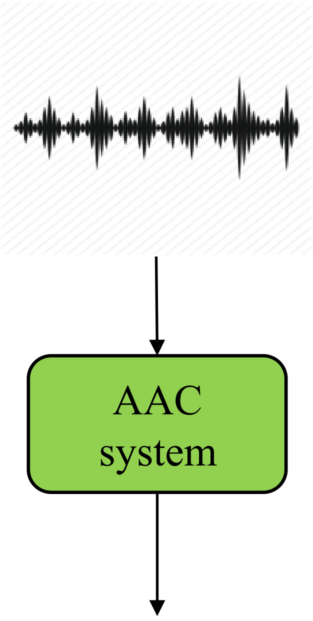
# Task 1: Audio to Text Generation

- **Automated audio captioning** (AAC) is a cross-modal translation task which aims at generating a natural language description given an audio clip.

- This task requires detecting the audio events and their spatial-temporal relationships and describing these information using natural language.

- Applications
  - Audio retrieval
  - Assist hearing-impared to understand environmental sounds
  - Subtitle for sounds in TV programs

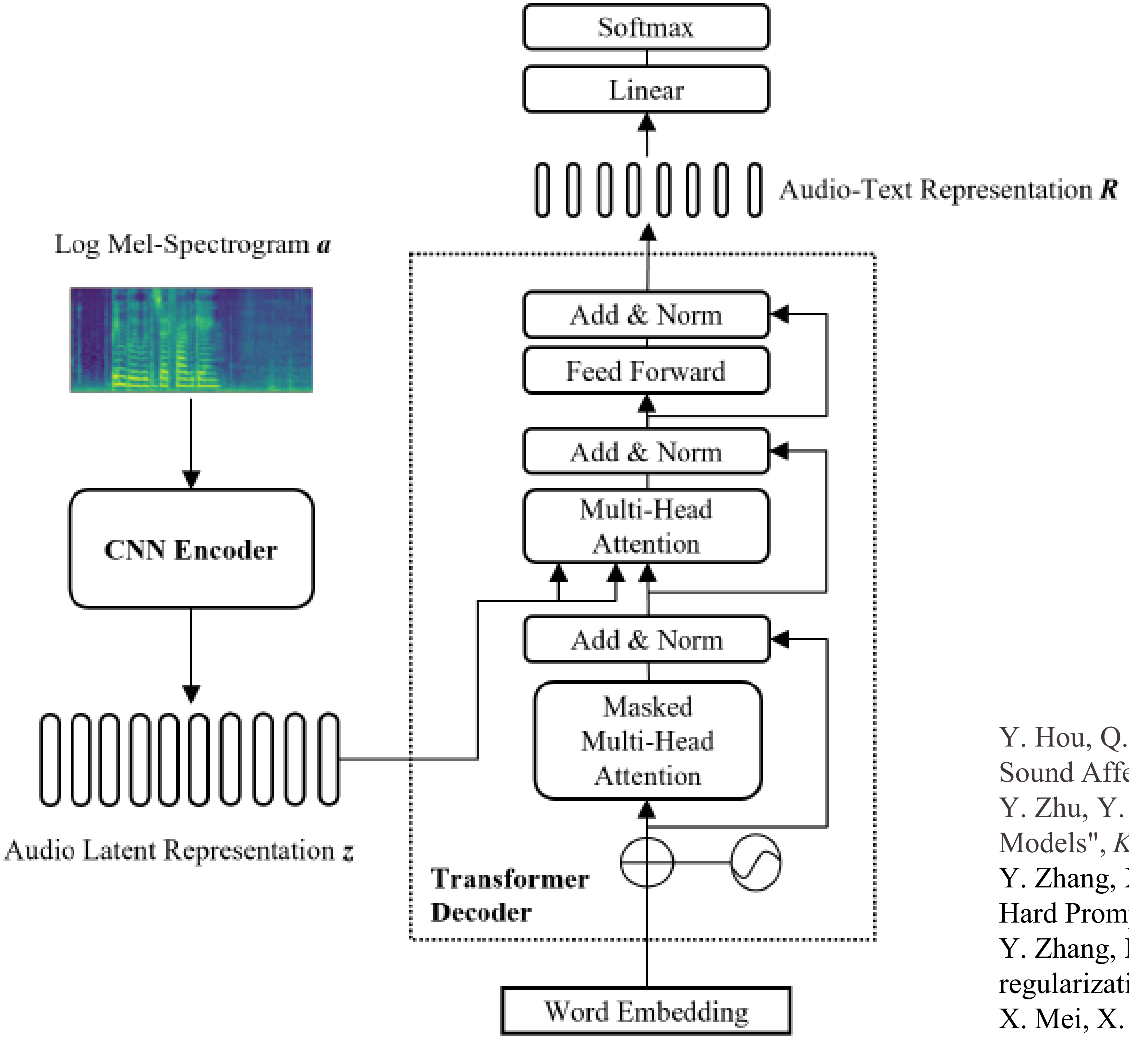
- AAC started in 2017, and has received increasing attention in recent three years with freely available datasets released and being held as a task in DCASE Challenges 2020-2022.



AAC system

“a woman talks nearby as water pours”

# An Example: CNN-Transformer Encoder-Decoder



**Common challenges in automated audio captioning**:

- Data scarcity
- Representations of audio, text and audio-text
- Diversity of captions
- Multi-lingual captioning
- Interactions with other modalities (e.g. vision)
- ....

Y. Hou, Q. Ren, A. Mitchell, W. Wang, J. Kang, T. Belpaeme, and D. Botteldooren, "Soundscape Captioning using Sound Affective Quality Network and Large Language Model," *IEEE Transactions on Multimedia*, 2026.
Y. Zhu, Y. Zhang, L. Xiao, W. Wang, and A. Men, "Zero-shot Diverse Audio Captioning with Diffusion Models", *Knowledge Based Systems*, 2026.
Y. Zhang, X. Xu, R. Du, H. Liu, Y. Dong, Z.-H. Tan, W. Wang, and Z. Ma, "Zero-Shot Audio Captioning Using Soft and Hard Prompts," *IEEE Transactions on Audio Speech and Language Processing,* vol. 33, pp. 2045 - 2058, May 2025.
Y. Zhang, H. Yu, R. Du, Z.-H. Tan, W. Wang, Z. Ma, Y. Dong, "ACTUAL: audio captioning with caption feature space regularization," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 2643 - 2657, 2023.
X. Mei, X. Liu, M. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges", *EURASIP Journal on Audio Speech and Music Processing*, 2022.
F. Xiao, J. Guan, H. Lan, Q. Zhu, and W. Wang, "Local information assisted attention-free decoder for audio captioning," *IEEE Signal Processing Letters*, vol. 29, pp. 1604-1608, 2022.

# Audio Captioning Demos

# Task 2: Audio Question Answering & Reasoning

**Acoustic Prompt Tuning (APT):** an adapter extending LLMs/VLMs to the audio domain using an improved soft-prompting approach

**Motivation:**

- **Existing works** on LLAMs used pretrained audio embeddings as soft prompt and adjust the LLM with Parameter-Efficient Fine-Tuning (PEFT).
- However, they **cannot generalize to multi-modal setting**, e.g., audio-visual language models.
- **Can we extend the off-the-shelf language models to the audio domain rather than training a dedicated LLM (~7B to 70B)?**

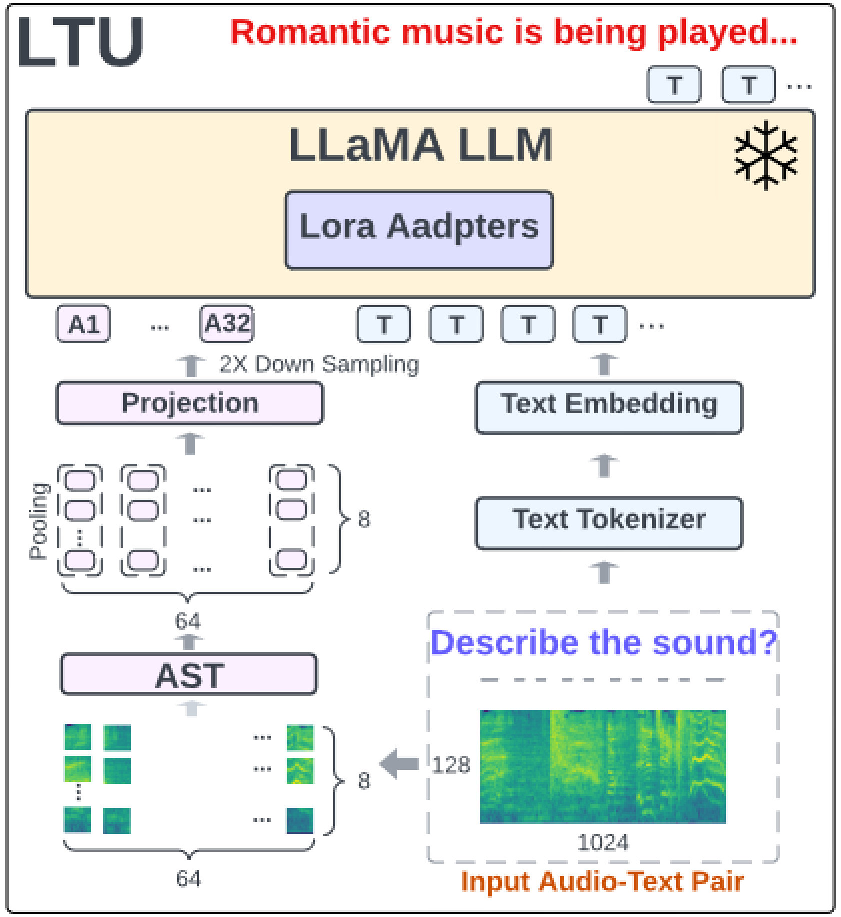
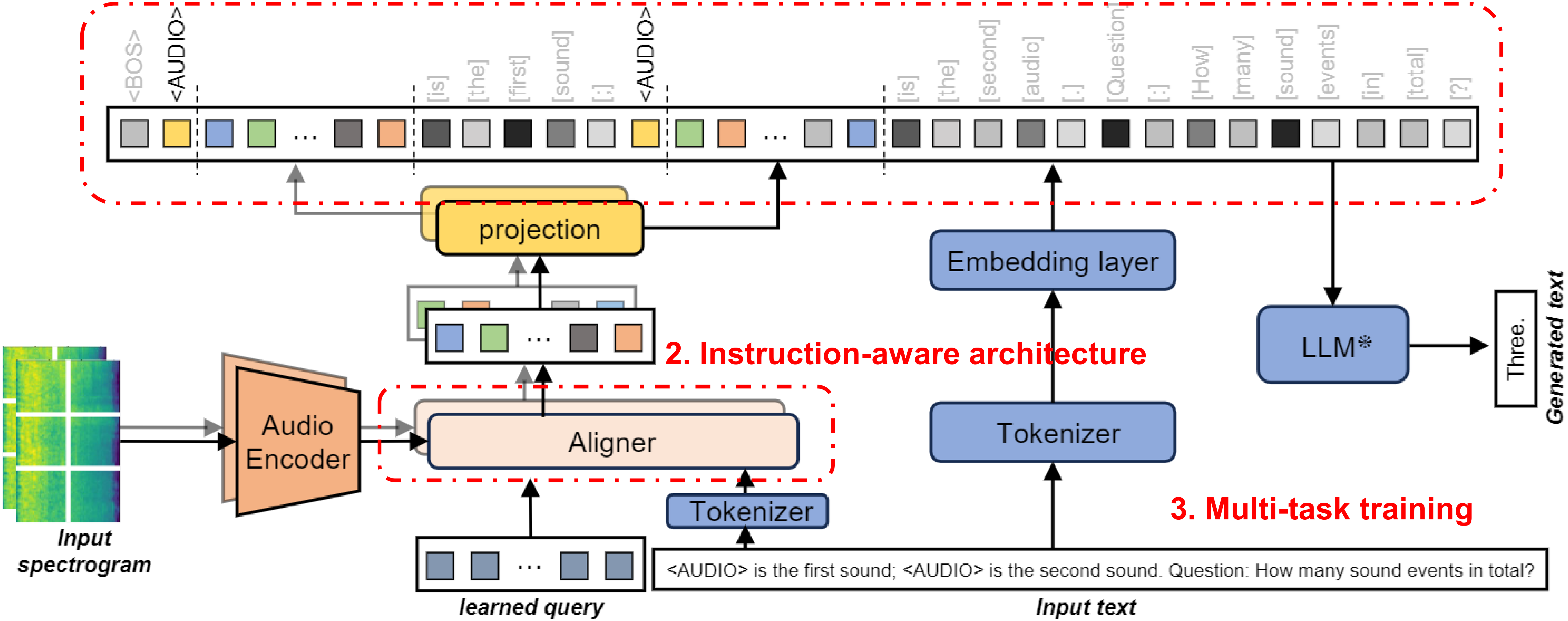


Fig. An example of an audio LLM structure (LTU (Gong et al, 2023)).

J. Liang, X. Liu, W. Wang, M. D. Plumbley, H. Phan, E. Benetos, "Acoustic Prompt Tuning: Empowering Large Language Models with Audition Capabilities", *IEEE Transactions on Audio Speech and Language Processing*, vol. 33, 949 - 961, 2025. arXiv:2312.00249

# Task: Audio Reasoning – **APT**

**Acoustic Prompt Tuning (APT):** an adapter extending LLMs/VLMs to the audio domain using an improved soft-prompting approach


1. Interleaved acoustic and text embeddings

2. Instruction-aware architecture

3. Multi-task training

<AUDIO> is the first sound; <AUDIO> is the second sound. Question: How many sound events in total?

# **APT** – Demos

"first_recording": "Creaking pier.wav"

🔊

"second_recording": "Machetes sliding 2.wav"

🔊

"first_recording": "Creaking pier.wav",
"second_recording": "Machetes sliding 2.wav",
"**question**": "In which recording are the sound events more evenly distributed?",
"**answer**": "second“

"first_recording": "Rain and Storm.wav"

🔊

"second_recording": "Car vs. Freight Train.wav"

🔊

"first_recording": "Rain and Storm.wav",
"second_recording": "Car vs. Freight Train.wav",
"**question**": "Does the second recording have a calming effect like the first recording?",
"**answer**": "yes"

Code: https://github.com/JinhuaLiang/APT
Paper: https://arxiv.org/abs/2312.00249

# Task 3: Text to Audio Generation

## Potential Applications:

Computational "foley artist":

- Game developer: e.g., A ghost is haunting a house.
- Audio producer: e.g., high heels hitting metal ground.
- Movie producer: e.g., the laser sound from a laser gun.
- …

Automatic content creation

- Endless music
- Audiobook with ambient noises
- White noise for meditation
- …

Data Augmentations

## Related Works:

### Label-to-Audio Generation

- Acoustic Scene (Kong et al., 2019), Sound event (Liu et al., 2019), FootStep (Comunit et al. 2019), …

### Text-to-Audio Generation

- DiffSound (Yang et al., 2022), AudioGen (Kreuk et al., 2022), Make-an-Audio (Huang et al., 2023)

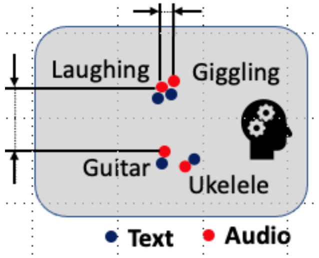
### Text-to-Music Generation

- MusicLM (Andrea et al., 2023)
- Moûsai (Flavio et al., 2023)
- Noise2Music (Huang et al., 2023)

### Others

- JukeBox (Dhariwal et al., 2020), AudioLM (Borsos et al., 2022), SingSong (Donahue et al., 2023),…

# AudioLDM

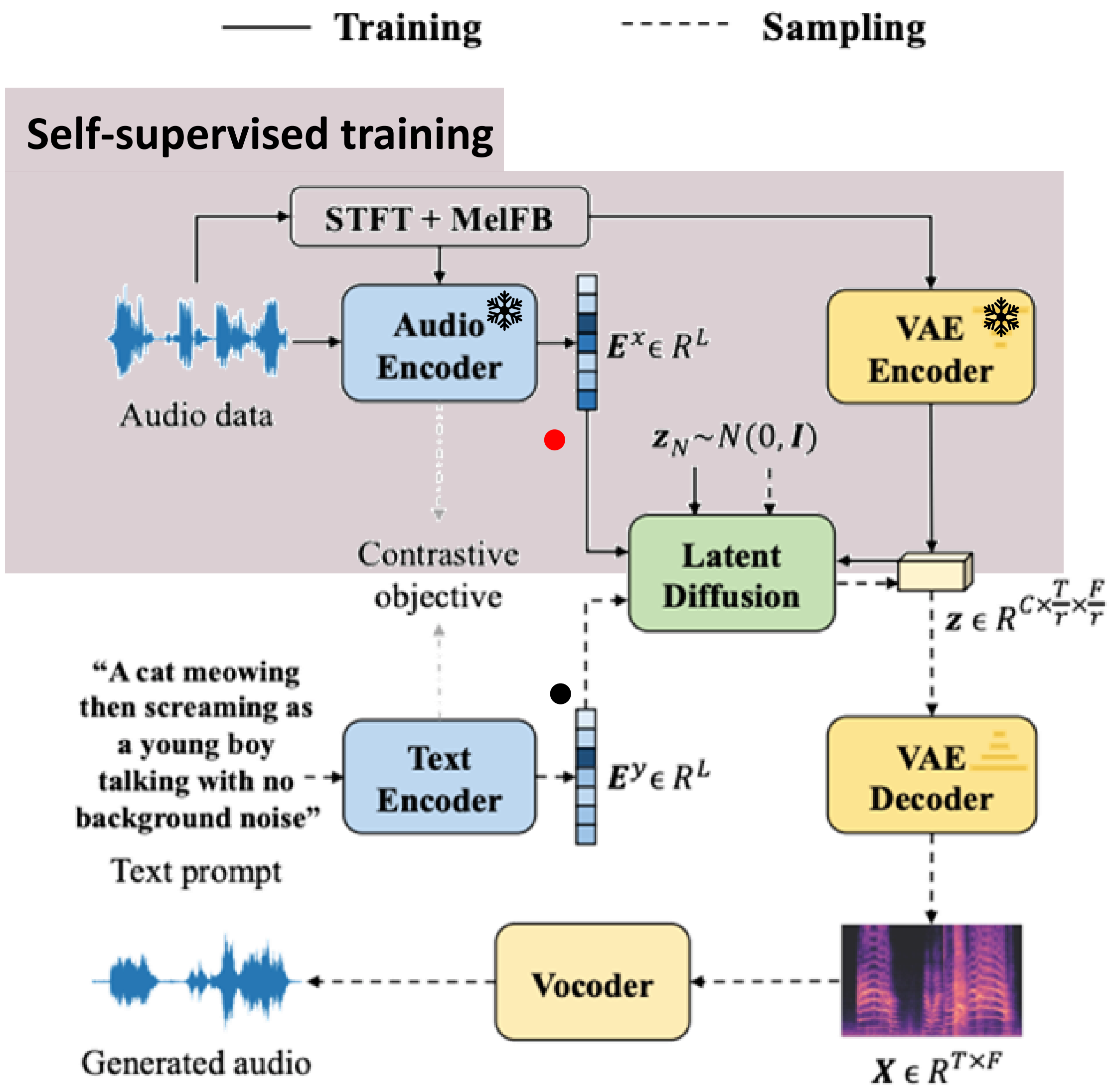


1. **Contrastive Language-Audio Learning (CLAP) Encoders**
   - Align audio and text in one space.

2. **Latent Diffusion Models**
   - Learn to generate VAE latent conditioned on CLAP embedding

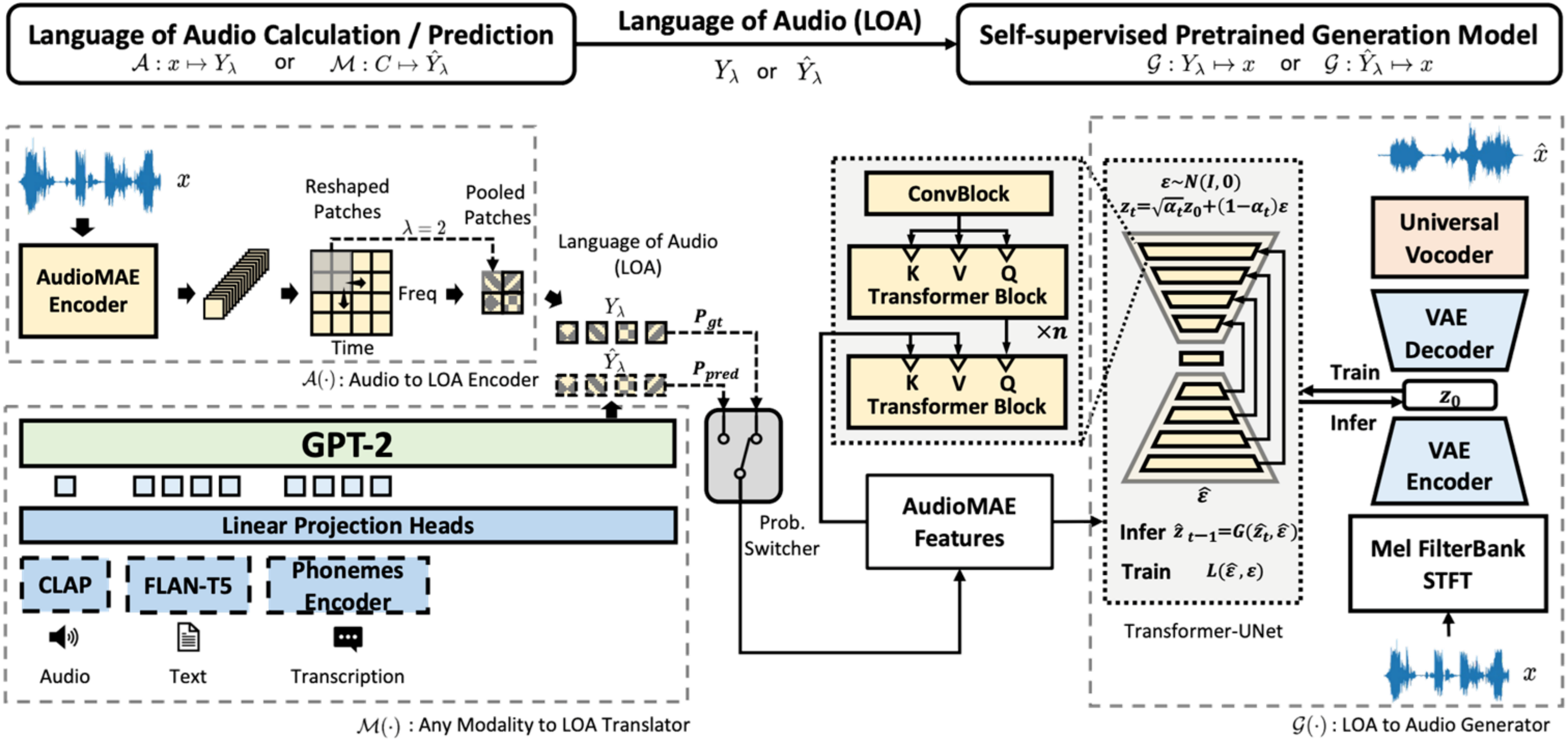
3. **Mel-spectrogram Autoencoder**
   - Learn latent representations.

4. **Mel-to-Waveform Vocoder**
   - Reverse Mel back to waveform



H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M. D. Plumbley, "AudioLDM: text-to-audio generation with latent diffusion models," in *Proc. IEEE International Conference on Machine Learning* (ICML 2023), Hawaii, USA, 23-29 July, 2023.

# **AudioLDM 2** - Architecture



Language of Audio Calculation / Prediction
$\mathcal{A}: x \mapsto Y_\lambda$ or $\mathcal{M}: C \mapsto \hat{Y}_\lambda$
Language of Audio (LOA)
$Y_\lambda$ or $\hat{Y}_\lambda$
Self-supervised Pretrained Generation Model
$\mathcal{G}: Y_\lambda \mapsto x$ or $\mathcal{G}: \hat{Y}_\lambda \mapsto x$
$x$
AudioMAE Encoder
Reshaped Patches
Pooled Patches
$\lambda = 2$
Freq
Time
Language of Audio (LOA)
$Y_\lambda$
$P_{gt}$
$\hat{Y}_\lambda$
$P_{pred}$
$\mathcal{A}(\cdot)$: Audio to LOA Encoder
GPT-2
Linear Projection Heads
CLAP
FLAN-T5
Phonemes Encoder
Audio
Text
Transcription
$\mathcal{M}(\cdot)$ : Any Modality to LOA Translator
Prob. Switcher
ConvBlock
K V Q
Transformer Block
$\times n$
K V Q
Transformer Block
AudioMAE Features
$\varepsilon \sim N(I, 0)$
$z_t = \sqrt{\alpha_t} z_0 + (1 - \alpha_t)\varepsilon$
$\hat{\varepsilon}$
Infer $\hat{z}_{t-1} = G(\hat{z}_t, \hat{\varepsilon})$
Train $L(\hat{\varepsilon}, \varepsilon)$
Transformer-UNet
$\hat{x}$
Universal Vocoder
VAE Decoder
Train
$z_0$
Infer
VAE Encoder
Mel FilterBank STFT
$x$
$\mathcal{G}(\cdot)$ : LOA to Audio Generator

H. Liu, et al. “AudioLDM 2: Learning holistic audio generation with self-supervised pretraining.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871 - 2883, 2024.

# AudioLDM 2 - Demo

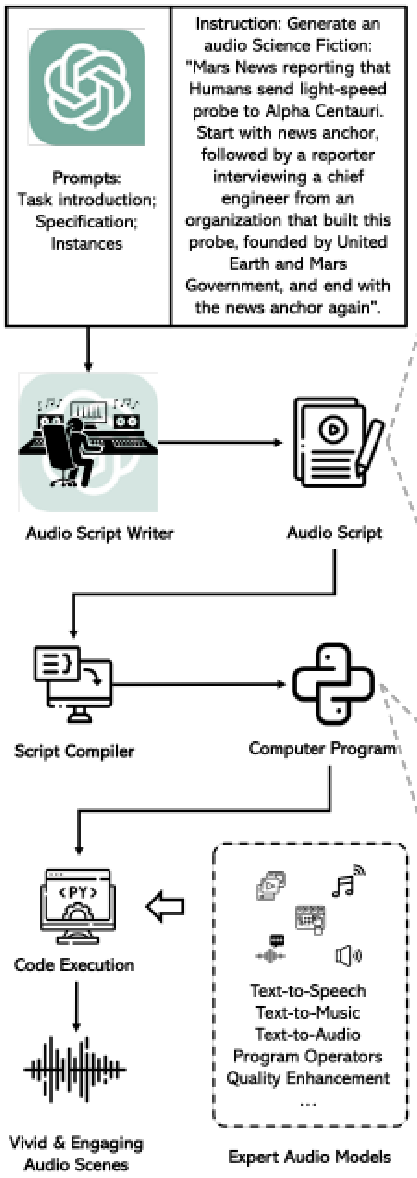
Text input: A traditional Irish fiddle playing a lively reel.

Up Next: The sound of a light saber

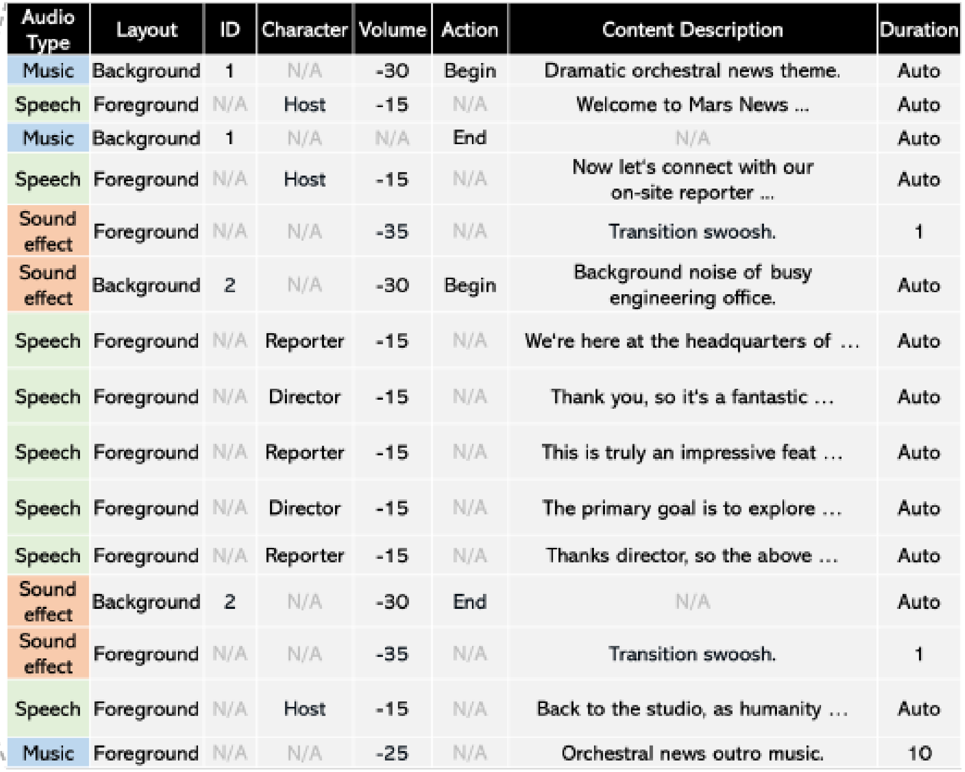
**We generated a total of 350 audio files with prompts (generated by ChatGPT) without cherry-picking.**

Codes and more demos: https://audioldm.github.io/audioldm2/

# Audio Storytelling - **WavJourney**

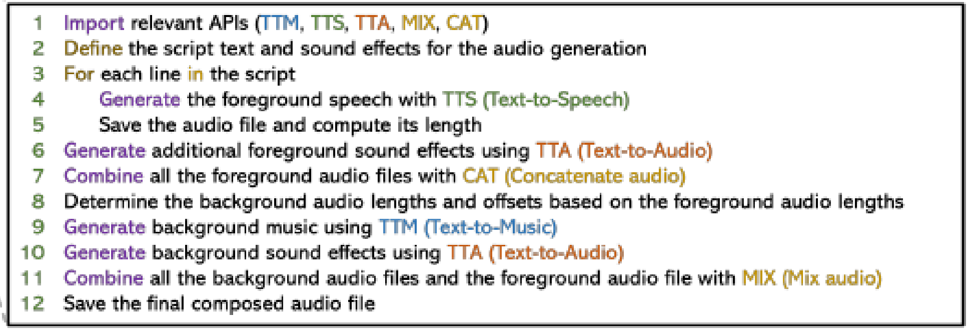


Instruction: Generate an audio Science Fiction: "Mars News reporting that Humans send light-speed probe to Alpha Centauri. Start with news anchor, followed by a reporter interviewing a chief engineer from an organization that built this probe, founded by United Earth and Mars Government, and end with the news anchor again".



| Audio Type | Layout | ID | Character | Volume | Action | Content Description | Duration |
|---|---|---|---|---|---|---|---|
| Music | Background | 1 | N/A | -30 | Begin | Dramatic orchestral news theme. | Auto |
| Speech | Foreground | N/A | Host | -15 | N/A | Welcome to Mars News ... | Auto |
| Music | Background | 1 | N/A | N/A | End | N/A | Auto |
| Speech | Foreground | N/A | Host | -15 | N/A | Now let's connect with our on-site reporter ... | Auto |
| Sound effect | Foreground | N/A | N/A | -35 | N/A | Transition swoosh. | 1 |
| Sound effect | Background | 2 | N/A | -30 | Begin | Background noise of busy engineering office. | Auto |
| Speech | Foreground | N/A | Reporter | -15 | N/A | We're here at the headquarters of ... | Auto |
| Speech | Foreground | N/A | Director | -15 | N/A | Thank you, so it's a fantastic ... | Auto |
| Speech | Foreground | N/A | Reporter | -15 | N/A | This is truly an impressive feat ... | Auto |
| Speech | Foreground | N/A | Director | -15 | N/A | The primary goal is to explore ... | Auto |
| Speech | Foreground | N/A | Reporter | -15 | N/A | Thanks director, so the above ... | Auto |
| Sound effect | Background | 2 | N/A | -30 | End | N/A | Auto |
| Sound effect | Foreground | N/A | N/A | -35 | N/A | Transition swoosh. | 1 |
| Speech | Foreground | N/A | Host | -15 | N/A | Back to the studio, as humanity ... | Auto |
| Music | Foreground | N/A | N/A | -25 | N/A | Orchestral news outro music. | 10 |

Audio script as a list of acoustic elements with attributes



```
Import relevant APIs (TTM, TTS, TTA, MIX, CAT)
Define the script text and sound effects for the audio generation
For each line in the script
    Generate the foreground speech with TTS (Text-to-Speech)
    Save the audio file and compute its length
Generate additional foreground sound effects using TTA (Text-to-Audio)
Combine all the foreground audio files with CAT (Concatenate audio)
Determine the background audio lengths and offsets based on the foreground audio lengths
Generate background music using TTM (Text-to-Music)
Generate background sound effects using TTA (Text-to-Audio)
Combine all the background audio files and the foreground audio file with MIX (Mix audio)
Save the final composed audio file
```

Script Compiler output (simplified pseudo code)

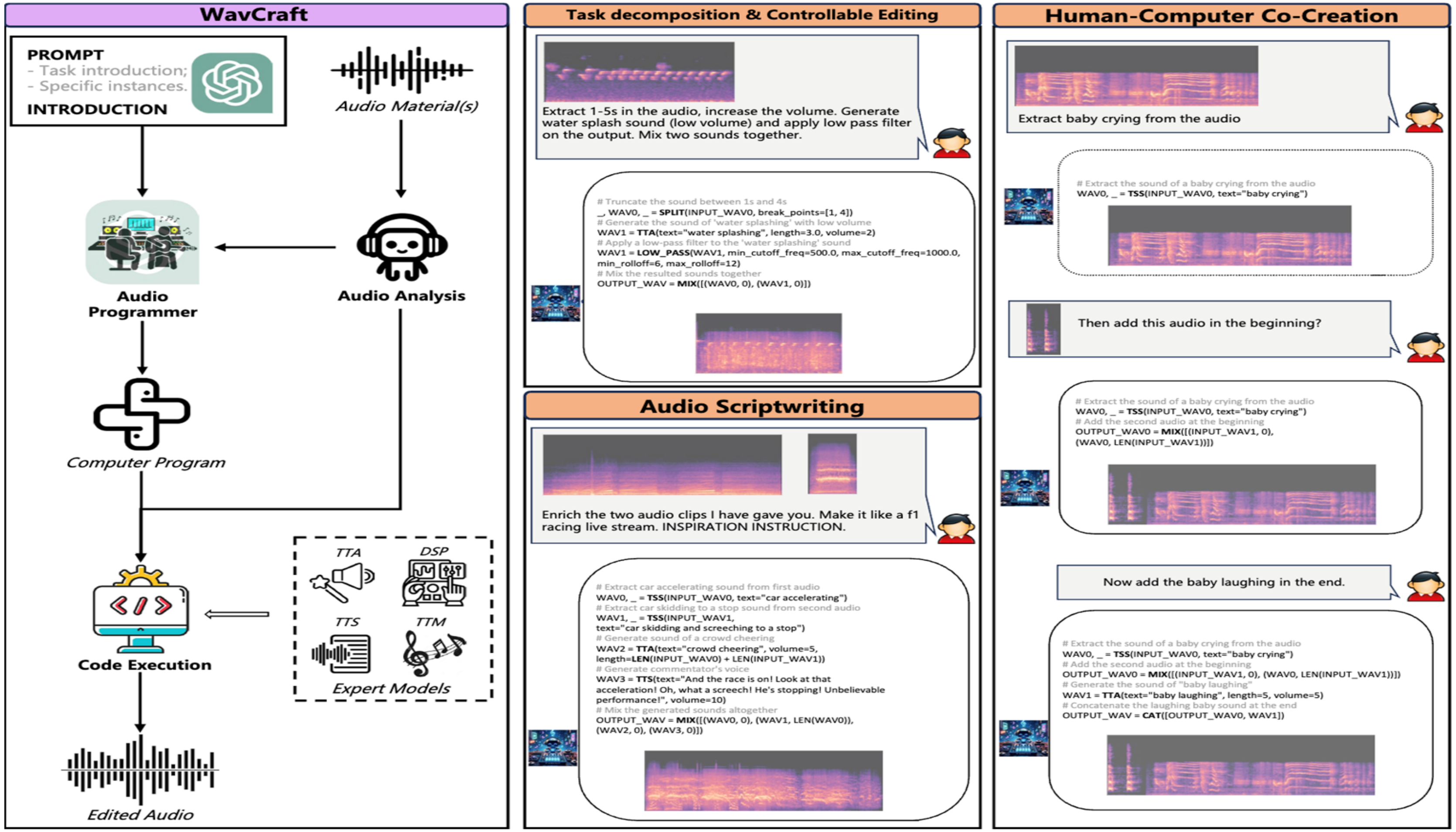
# **WavJourney** – Sound Demo for Science Fiction Storytelling

UNIVERSITY OF SURREY

| | |
|---|---|
| **Paper**: | https://arxiv.org/abs/2307.14335 |
| **Code**: | https://github.com/Audio-AGI/WavJourney |
| **Demo**: | https://huggingface.co/spaces/Audio-AGI/WavJourney |

# Task 4: LLMs for Controllable Audio Editing - **WavCraft**

**Audio editing** is to change the content of audio by following the instruction precisely.
This work introduces an audio agent that understands the user instruction, decomposes the instruction into several tasks, and allocates different tasks to the proper models.



Extract 1-5s in the audio, increase the volume. Generate water splash sound (low volume) and apply low pass filter on the output. Mix two sounds together.

```
# Truncate the sound between 1s and 4s
_, WAV0, _ = SPLIT(INPUT_WAV0, break_points=[1, 4])
# Generate the sound of 'water splashing' with low volume
WAV1 = TTA(text="water splashing", length=3.0, volume=2)
# Apply a low-pass filter to the 'water splashing' sound
WAV1 = LOW_PASS(WAV1, min_cutoff_freq=500.0, max_cutoff_freq=1000.0,
min_rolloff=6, max_rolloff=12)
# Mix the resulted sounds together
OUTPUT_WAV = MIX([(WAV0, 0), (WAV1, 0)])
```

Enrich the two audio clips I have gave you. Make it like a f1 racing live stream. INSPIRATION INSTRUCTION.

```
# Extract car accelerating sound from first audio
WAV0, _ = TSS(INPUT_WAV0, text="car accelerating")
# Extract car skidding to a stop sound from second audio
WAV1, _ = TSS(INPUT_WAV1,
text="car skidding and screeching to a stop")
# Generate sound of a crowd cheering
WAV2 = TTA(text="crowd cheering", volume=5,
length=LEN(INPUT_WAV0) + LEN(INPUT_WAV1))
# Generate commentator's voice
WAV3 = TTS(text="And the race is on! Look at that
acceleration! Oh, what a screech! He's stopping! Unbelievable
performance!", volume=10)
# Mix the generated sounds altogether
OUTPUT_WAV = MIX([(WAV0, 0), (WAV1, LEN(WAV0)),
(WAV2, 0), (WAV3, 0)])
```

Extract baby crying from the audio

```
# Extract the sound of a baby crying from the audio
WAV0, _ = TSS(INPUT_WAV0, text="baby crying")
```

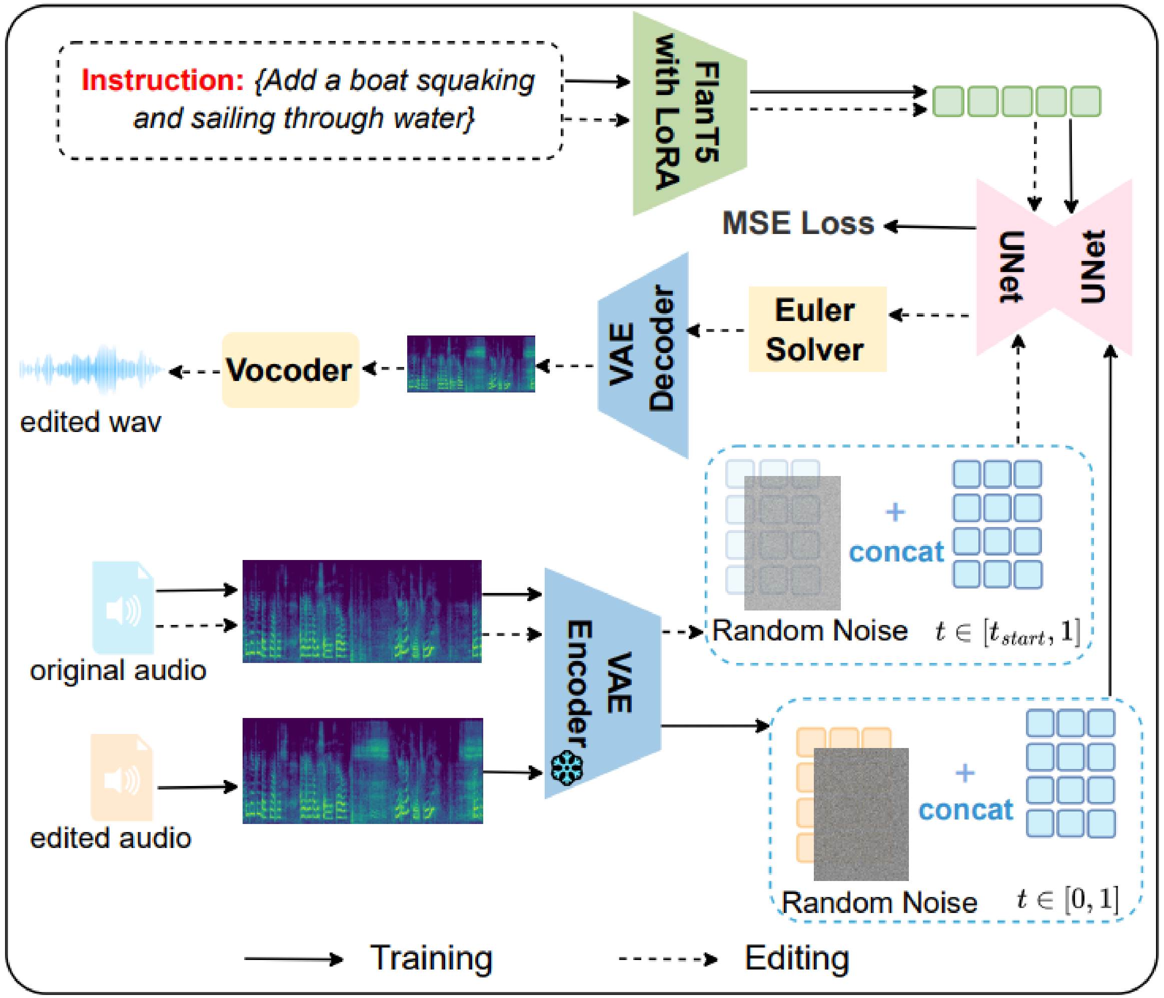
Then add this audio in the beginning?

```
# Extract the sound of a baby crying from the audio
WAV0, _ = TSS(INPUT_WAV0, text="baby crying")
# Add the second audio at the beginning
OUTPUT_WAV0 = MIX([(INPUT_WAV1, 0),
(WAV0, LEN(INPUT_WAV1))])
```

Now add the baby laughing in the end.

```
# Extract the sound of a baby crying from the audio
WAV0, _ = TSS(INPUT_WAV0, text="baby crying")
# Add the second audio at the beginning
OUTPUT_WAV0 = MIX([(INPUT_WAV1, 0), (WAV0, LEN(INPUT_WAV1))])
# Generate the sound of "baby laughing"
WAV1 = TTA(text="baby laughing", length=5, volume=5)
# Concatenate the laughing baby sound at the end
OUTPUT_WAV = CAT([OUTPUT_WAV0, WAV1])
```

# Task 4: LLMs for Controllable Audio Editing - **RFM**



Demos here:
https://katelin-glt.github.io/RFM-Editing-Demo/

**Remove continuous frying noises:**

Original:

Edited:

**Replace someone suddenly sneezes out loud with several pigeons cooing:**

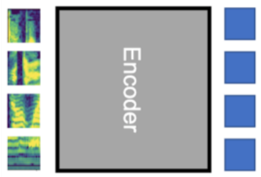
Original:

Edited:

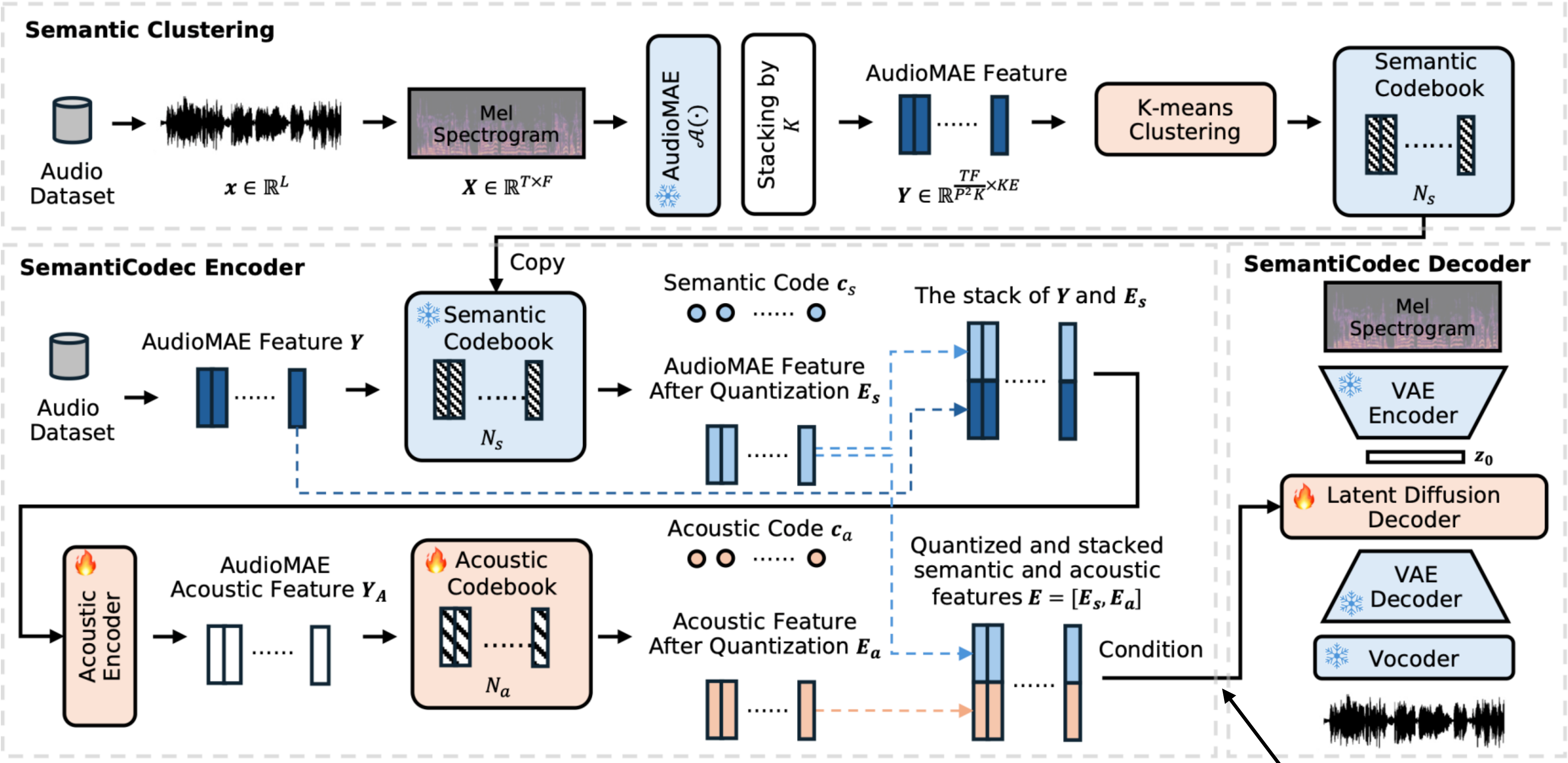
L. Gao, Y. Yuan, Y. Chen, Y. Cheng, Z. Li, J. Wen, S. Zhang, and W. Wang, "RFM-EDITING: Rectified Flow Matching for Text-Guided Audio Editing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2026), Barcelona, Spain, May 4-8, 2026.

# Task 5: Neural Audo Codec - **SemantiCodec**

- Ultra-low bit rate (0.31 kbps ~1.40 kbps, token rate 25, 50, or 100 per second) & Strong semantics in the token & Variable vocabulary sizes



AudioMAE Encoder



CrossAtten

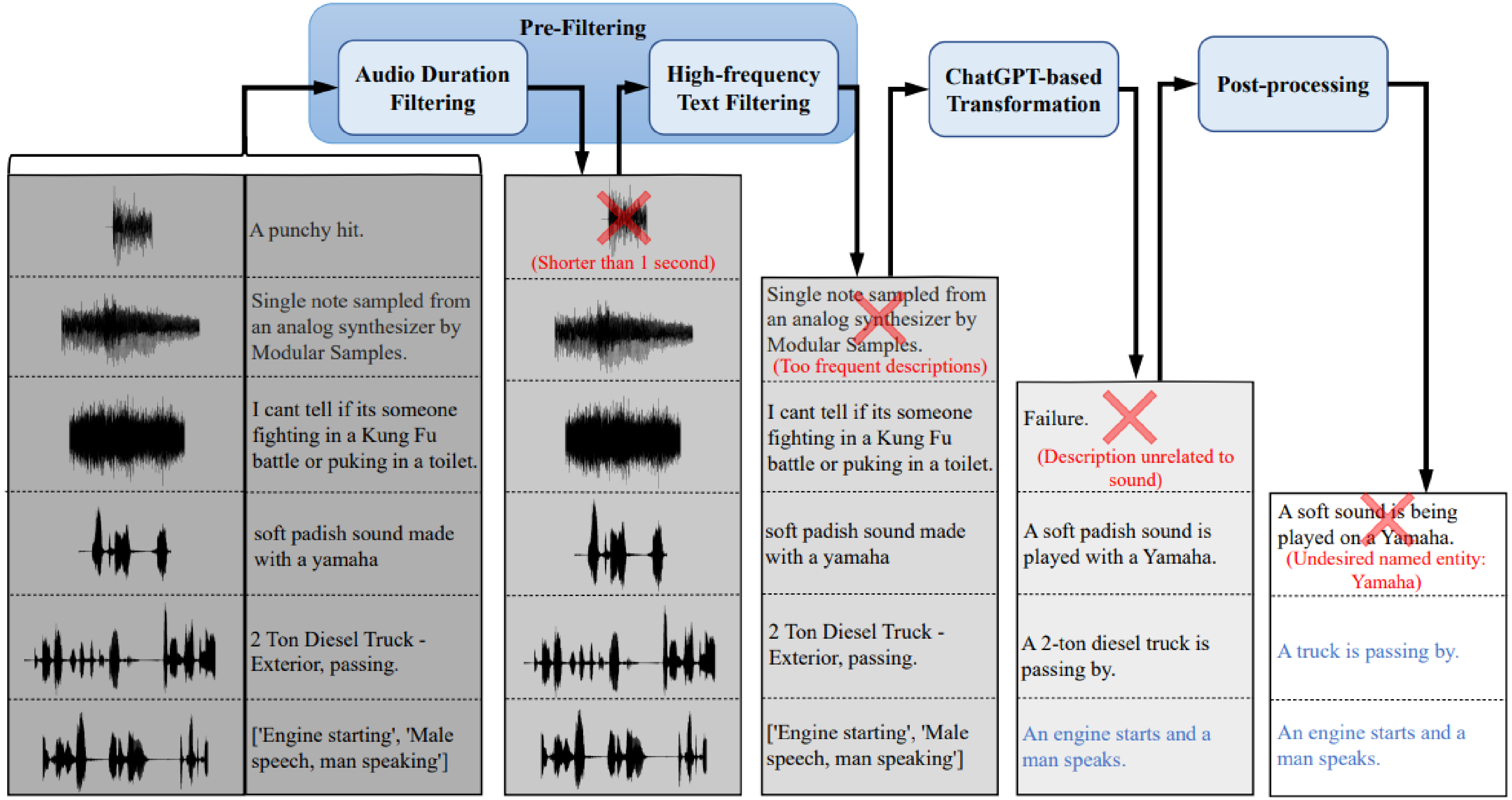
https://haoheliu.github.io/SemantiCodec/

# Task 5: Sound Demos

| | Original | HiFi-Codec (2.0 kbps) | Encodec (1.5 kbps) | DAC (1.41 kbps) | SemantiCodec (1.43 kbps) | DAC (0.47 kbps) | SemantiCodec (0.35 kbps) |
|---|---|---|---|---|---|---|---|
| Music (MUSDB18) | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 |
| General Audio (AudioSet) | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 |
| Speech (Libri) | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 | 🔊 |

**More sound demos:** https://haoheliu.github.io/SemantiCodec/

# Dataset: **WavCaps**



Pre-Filtering

Audio Duration Filtering

High-frequency Text Filtering

ChatGPT-based Transformation

Post-processing

A punchy hit.

Single note sampled from an analog synthesizer by Modular Samples.

I cant tell if its someone fighting in a Kung Fu battle or puking in a toilet.

soft padish sound made with a yamaha

2 Ton Diesel Truck - Exterior, passing.

['Engine starting', 'Male speech, man speaking']

(Shorter than 1 second)

Single note sampled from an analog synthesizer by Modular Samples.
(Too frequent descriptions)

I cant tell if its someone fighting in a Kung Fu battle or puking in a toilet.

soft padish sound made with a yamaha

2 Ton Diesel Truck - Exterior, passing.

['Engine starting', 'Male speech, man speaking']

Failure.
(Description unrelated to sound)

A soft padish sound is played with a Yamaha.

A 2-ton diesel truck is passing by.

An engine starts and a man speaks.

A soft sound is being played on a Yamaha.
(Undesired named entity: Yamaha)

A truck is passing by.

An engine starts and a man speaks.

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," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, pp. 3339 - 3354, June 2024.

Code: https://github.com/XinhaoMei/WavCaps

# WavCaps – Statistics

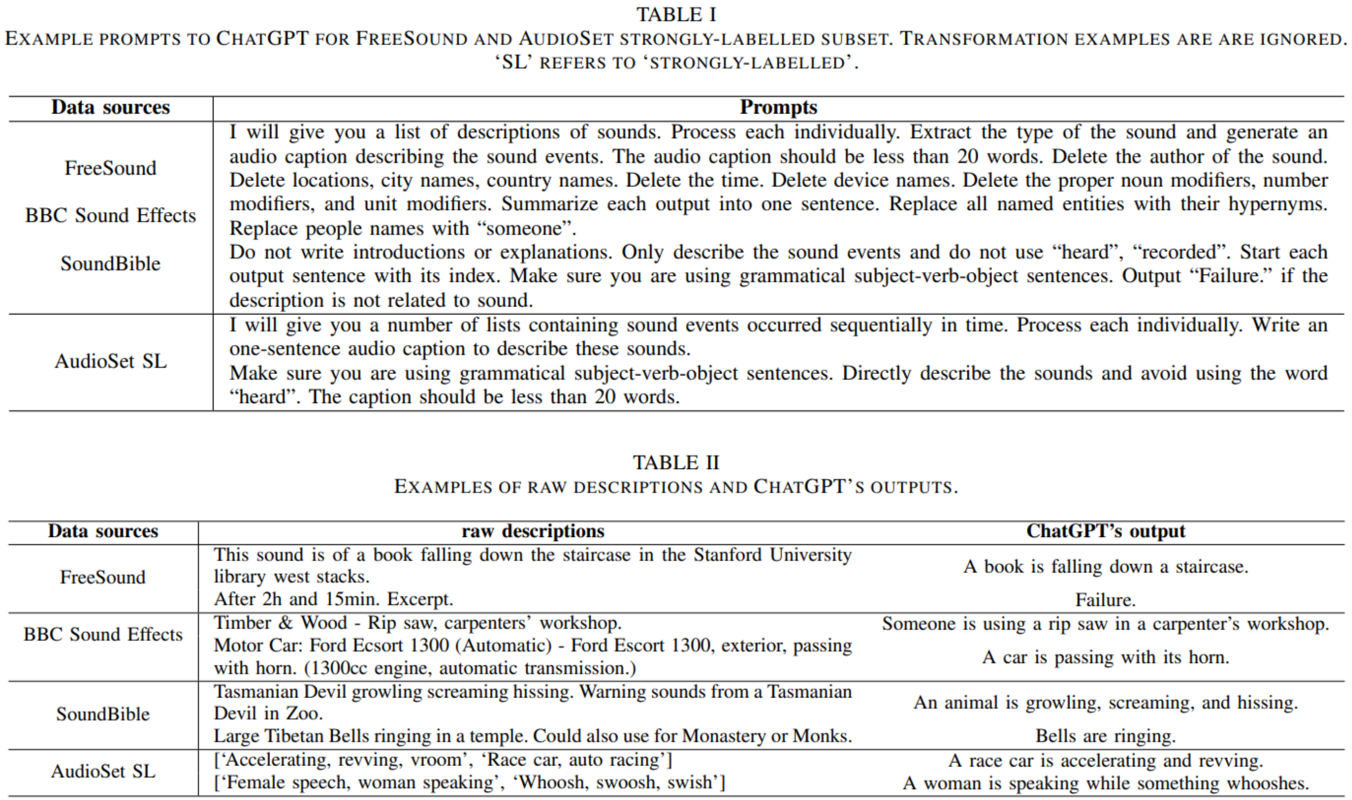


TABLE I

Example prompts to ChatGPT for FreeSound and AudioSet strongly-labelled subset. Transformation examples are are ignored. 'SL' refers to 'strongly-labelled'.

| Data sources | Prompts |
|---|---|
| FreeSound<br>BBC Sound Effects<br>SoundBible | I will give you a list of descriptions of sounds. Process each individually. Extract the type of the sound and generate an audio caption describing the sound events. The audio caption should be less than 20 words. Delete the author of the sound. Delete locations, city names, country names. Delete the time. Delete device names. Delete the proper noun modifiers, number modifiers, and unit modifiers. Summarize each output into one sentence. Replace all named entities with their hypernyms. Replace people names with "someone".<br>Do not write introductions or explanations. Only describe the sound events and do not use "heard", "recorded". Start each output sentence with its index. Make sure you are using grammatical subject-verb-object sentences. Output "Failure." if the description is not related to sound. |
| AudioSet SL | I will give you a number of lists containing sound events occurred sequentially in time. Process each individually. Write an one-sentence audio caption to describe these sounds.<br>Make sure you are using grammatical subject-verb-object sentences. Directly describe the sounds and avoid using the word "heard". The caption should be less than 20 words. |

TABLE II

Examples of raw descriptions and ChatGPT's outputs.

| Data sources | raw descriptions | ChatGPT's output |
|---|---|---|
| FreeSound | This sound is of a book falling down the staircase in the Stanford University library west stacks.<br>After 2h and 15min. Excerpt. | A book is falling down a staircase.<br>Failure. |
| BBC Sound Effects | Timber & Wood - Rip saw, carpenters' workshop.<br>Motor Car: Ford Ecsort 1300 (Automatic) - Ford Escort 1300, exterior, passing with horn. (1300cc engine, automatic transmission.) | Someone is using a rip saw in a carpenter's workshop.<br>A car is passing with its horn. |
| SoundBible | Tasmanian Devil growling screaming hissing. Warning sounds from a Tasmanian Devil in Zoo.<br>Large Tibetan Bells ringing in a temple. Could also use for Monastery or Monks. | An animal is growling, screaming, and hissing.<br>Bells are ringing. |
| AudioSet SL | ['Accelerating, revving, vroom', 'Race car, auto racing']<br>['Female speech, woman speaking', 'Whoosh, swoosh, swish'] | A race car is accelerating and revving.<br>A woman is speaking while something whooshes. |

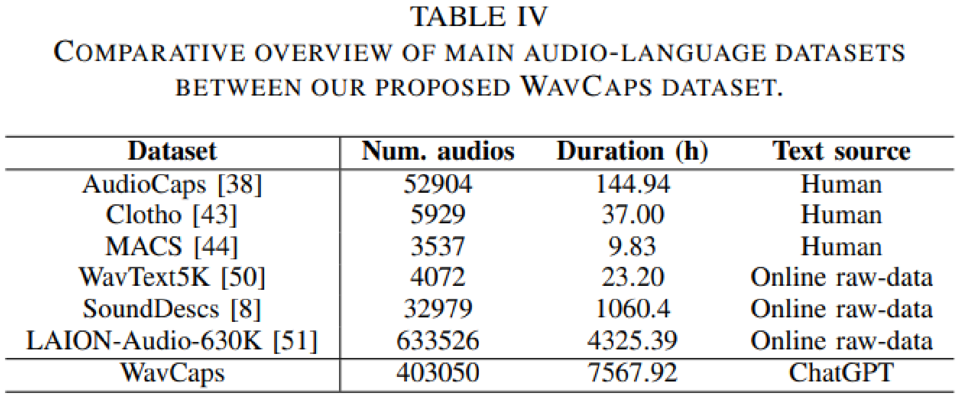


TABLE IV

Comparative overview of main audio-language datasets between our proposed WavCaps dataset.

| Dataset | Num. audios | Duration (h) | Text source |
|---|---|---|---|
| AudioCaps [38] | 52904 | 144.94 | Human |
| Clotho [43] | 5929 | 37.00 | Human |
| MACS [44] | 3537 | 9.83 | Human |
| WavText5K [50] | 4072 | 23.20 | Online raw-data |
| SoundDescs [8] | 32979 | 1060.4 | Online raw-data |
| LAION-Audio-630K [51] | 633526 | 4325.39 | Online raw-data |
| WavCaps | 403050 | 7567.92 | ChatGPT |

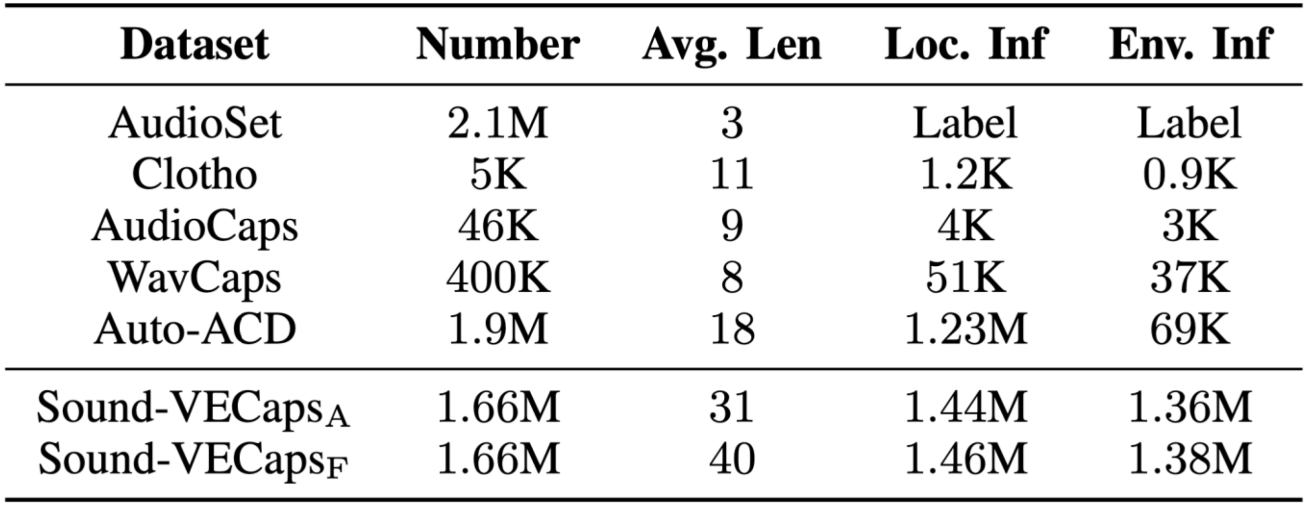
# Sound-VECaps

## Challenge

- Existing audio generation models struggle with complex and detailed prompts, leading to potential performance degradation.
- Captions of current audio datasets are too simple to provide detail information.

## Sound-VECaps

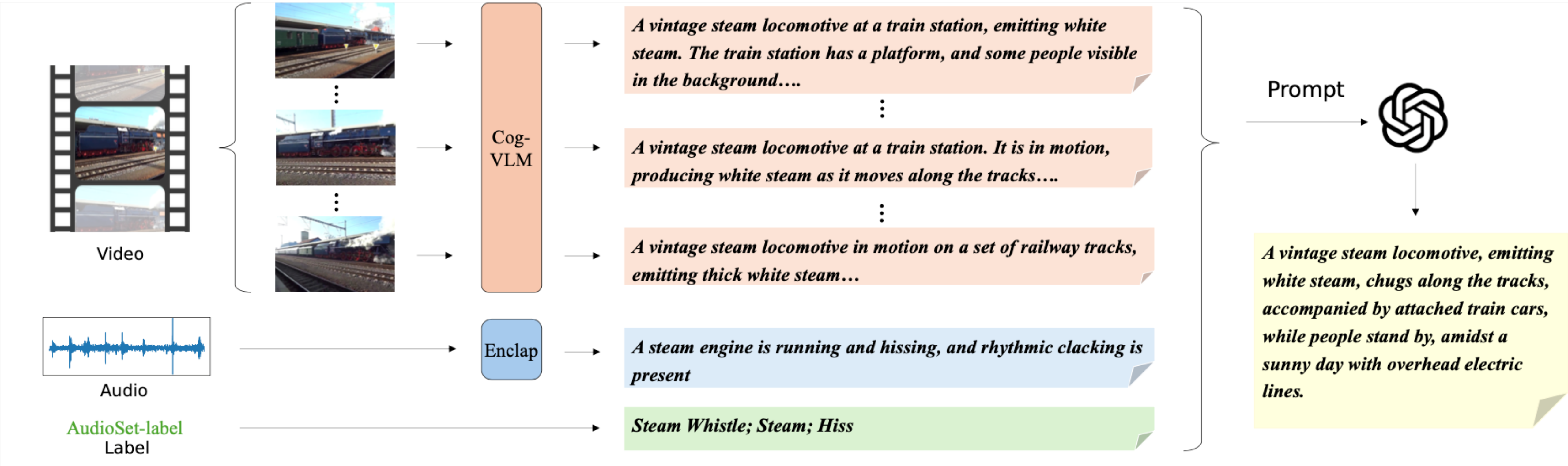
- 1.66M high-quality audio-caption pairs with enriched details including audio event orders, occurred places and environment information.



| Dataset | Number | Avg. Len | Loc. Inf | Env. Inf |
|---|---|---|---|---|
| AudioSet | 2.1M | 3 | Label | Label |
| Clotho | 5K | 11 | 1.2K | 0.9K |
| AudioCaps | 46K | 9 | 4K | 3K |
| WavCaps | 400K | 8 | 51K | 37K |
| Auto-ACD | 1.9M | 18 | 1.23M | 69K |
| Sound-VECaps$_A$ | 1.66M | 31 | 1.44M | 1.36M |
| Sound-VECaps$_F$ | 1.66M | 40 | 1.46M | 1.38M |

The analysis of audio-caption datasets, Loc and Env are the number of captions that include the location and environment information.

# **Sound-VECaps** – Processing Pipeline



*A vintage steam locomotive at a train station, emitting white steam. The train station has a platform, and some people visible in the background….*

*A vintage steam locomotive at a train station. It is in motion, producing white steam as it moves along the tracks….*

*A vintage steam locomotive in motion on a set of railway tracks, emitting thick white steam…*

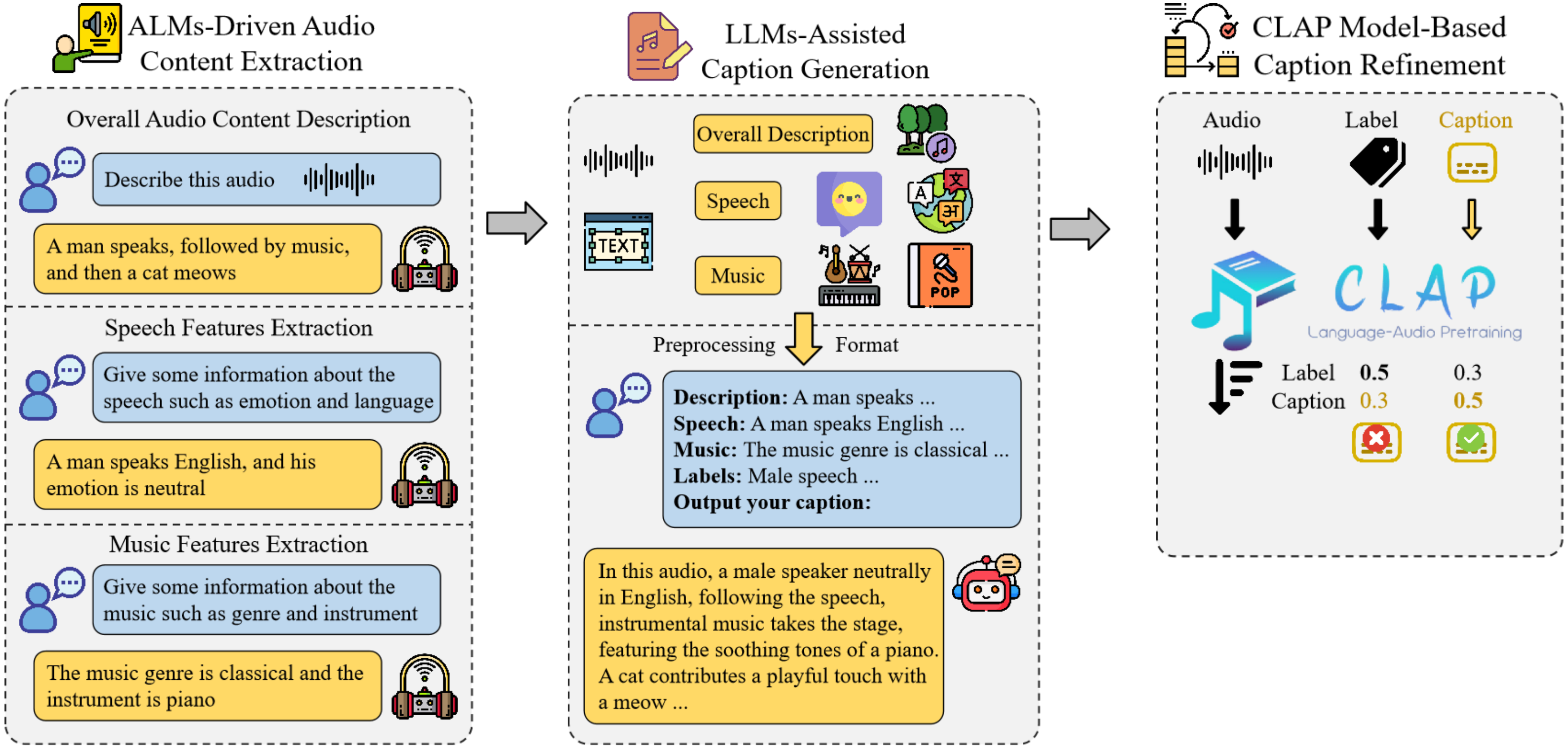
*A steam engine is running and hissing, and rhythmic clacking is present*

*Steam Whistle; Steam; Hiss*

*A vintage steam locomotive, emitting white steam, chugs along the tracks, accompanied by attached train cars, while people stand by, amidst a sunny day with overhead electric lines.*

Paper, data & code: https://yyua8222.github.io/Sound-VECaps-demo

# Dataset: **AudioSetCaps**


ALMs-Driven Audio Content Extraction
Overall Audio Content Description
Describe this audio
A man speaks, followed by music, and then a cat meows
Speech Features Extraction
Give some information about the speech such as emotion and language
A man speaks English, and his emotion is neutral
Music Features Extraction
Give some information about the music such as genre and instrument
The music genre is classical and the instrument is piano
LLMs-Assisted Caption Generation
TEXT
Overall Description
Speech
Music
POP
Preprocessing
Format
Description: A man speaks ...
Speech: A man speaks English ...
Music: The music genre is classical ...
Labels: Male speech ...
Output your caption:
In this audio, a male speaker neutrally in English, following the speech, instrumental music takes the stage, featuring the soothing tones of a piano. A cat contributes a playful touch with a meow ...
CLAP Model-Based Caption Refinement
Audio
Label
Caption
CLAP
Language-Audio Pretraining
Label 0.5 0.3
Caption 0.3 0.5

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," *IEEE Transactions on Audio Speech and Language Processing*, vol. 33, pp. 2817 - 2829, June 2025.

# **AudioSetCaps** – An Example

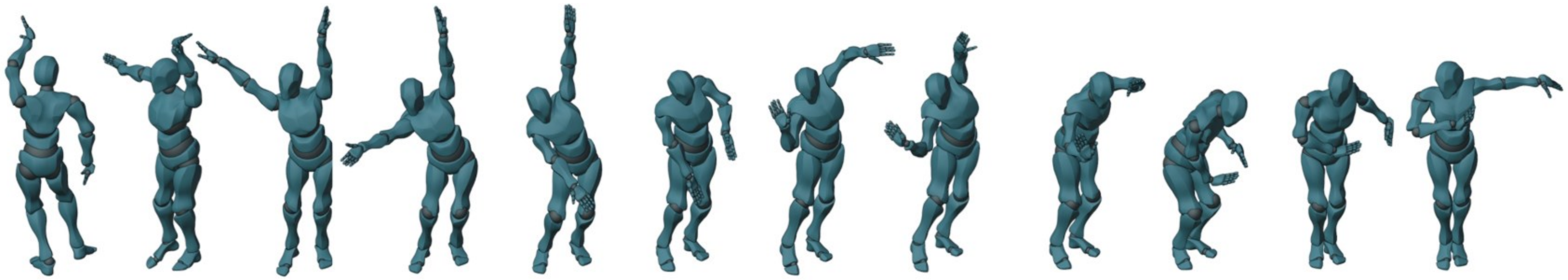
ID: Y0qH8FmqGI2U

| | Dataset-Caption | Mean Subjective Score |
|---|---|---|
| Label | Female speech, Woman speaking, Background noise, Generic impact sounds, Surface contact, Babbling, Tick, Human voice, Breathing, Baby laughter | 4 |
| AudioCaps | A human baby laughs and gurgles as a female sings gently | 4 |
| WavCaps | People are talking and babbling with a baby laughing and surface contact. | 4.2 |
| Auto-ACD | The sound of a laughing baby and women chatting and giggling can be heard at a busy spa. | 4 |
| **AudioSetCaps** | **A joyful interaction between a woman and a baby, as the infant giggles and the woman responds with a happy and upbeat tone.** | **4.4** |

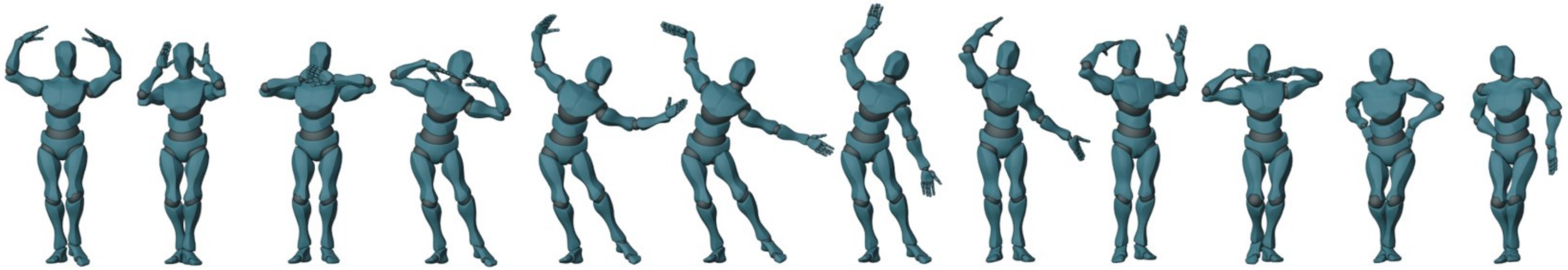
# Conclusion & Future Works

- **Summary**
  - Large language-audio models are promising - offering new opportunities to solve problems in conventional audio tasks and newly emerging audio tasks
  - These models often provide SOTA performance in many downstream tasks and may offer new capabilities that were not available in previous audio models.
- **Future Works**
  - Developing unified models for multi-tasks (e.g. understanding and generation) and multi-modal data (audio, visual, language)
  - Leveraging respective strengths of LLMs & audio models
  - Towards improved controllability in various downstream tasks (e.g. generation and captioning)
  - Leveraging physics-based model + data driven models

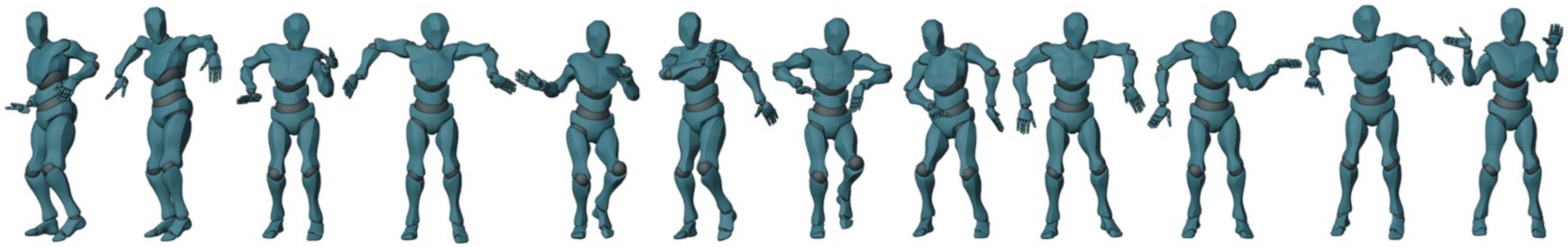
# Ongoing Work: Music to Dance Generation



*This is a "HanTang" type of music.*



*This is a "Dai" type of music.*



*This is a "Hiphop" type of music.*

# Paper, Codes, Demos, and More, …

**AudioLDM**:
Paper: https://arxiv.org/abs/2301.12503
Project Page: https://audioldm.github.io/
Github:

- Pretrained model: https://github.com/haoheliu/AudioLDM
- Evaluation tools: https://github.com/haoheliu/audioldm_eval

YouTube: https://www.youtube.com/watch?v=_0VTltNYhao

**SemantiCodec**:

Paper/code/demos at project page:
https://haoheliu.github.io/SemantiCodec/

**AudioSep**:
Code: https://github.com/Audio-AGI/AudioSep

**WavCaps**:
Paper: https://arxiv.org/abs/2303.17395
Code: https://github.com/xinhaomei/wavcaps

More code about other works avaliable at:
https://github.com/XinhaoMei/DCASE2021_task6_v2
https://github.com/XinhaoMei/ACT
https://github.com/liuxubo717/cl4ac

**AudioLDM2**:
Project Page: https://audioldm.github.io/audioldm2/

**APT**:
Code: https://github.com/JinhuaLiang/APT

**WavCraft**:
Code: https://github.com/JinhuaLiang/WavCraft

**WavJourney**:
Paper: https://arxiv.org/abs/2307.14335
Code: https://github.com/Audio-AGI/WavJourney
Demo: https://huggingface.co/spaces/Audio-AGI/WavJourney

**AudioSetCaps**:
Data and code: https://github.com/JishengBai/AudioSetCaps
https://huggingface.co/datasets/baijs/AudioSetCaps

**Sound-VECaps**: paper, data & code:
https://yyua8222.github.io/Sound-VECaps-demo

# Thank you for listening!

This work was sponsored by a Newton Institutional Links Award from the British Council, titled “Automated Captioning of Image and Audio for Visually and Hearing Impaired” & EPSRC projects titled “Making Sense of Sounds” and “AI for Sounds”.