Large Audio-Language Models and Applications

Wenwu Wang

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

University of Surrey

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

Invited Talk on the CCF Speech and Audio Workshop @ Suzhou

29/07/2026

Many thanks to…

- **Xinhao Mei**
- **Haohe Liu**
- **Xubo Liu**
- **Jinhua Liang**
- **Yi Yuan**
- **Yun Chen**
- **Xinran Liu**
- **Yiming Zhang**
- **Liting Gao**
- **Qiuqiang Kong**
- **Jisheng Bai**
- **Zehua Chen**
- **Yuanbo Hou**
- **Yuelan Cheng**
- **Junqi Zhao**
- **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**
  - Motivation
  - Example models & datasets
  - Open challenges
  - Integration of audio and language models
- **Applications of LALMs to various cross-modal generation tasks**
  - Audio captioning (e.g. audio to text generation)
  - Audio question answering and reasoning
  - Text to audio generation & composition & storytelling
  - Language guided audio source separation
  - LLMs for controllable audio and speech editing
  - Neural audio coding
- **Conclusions and future works**

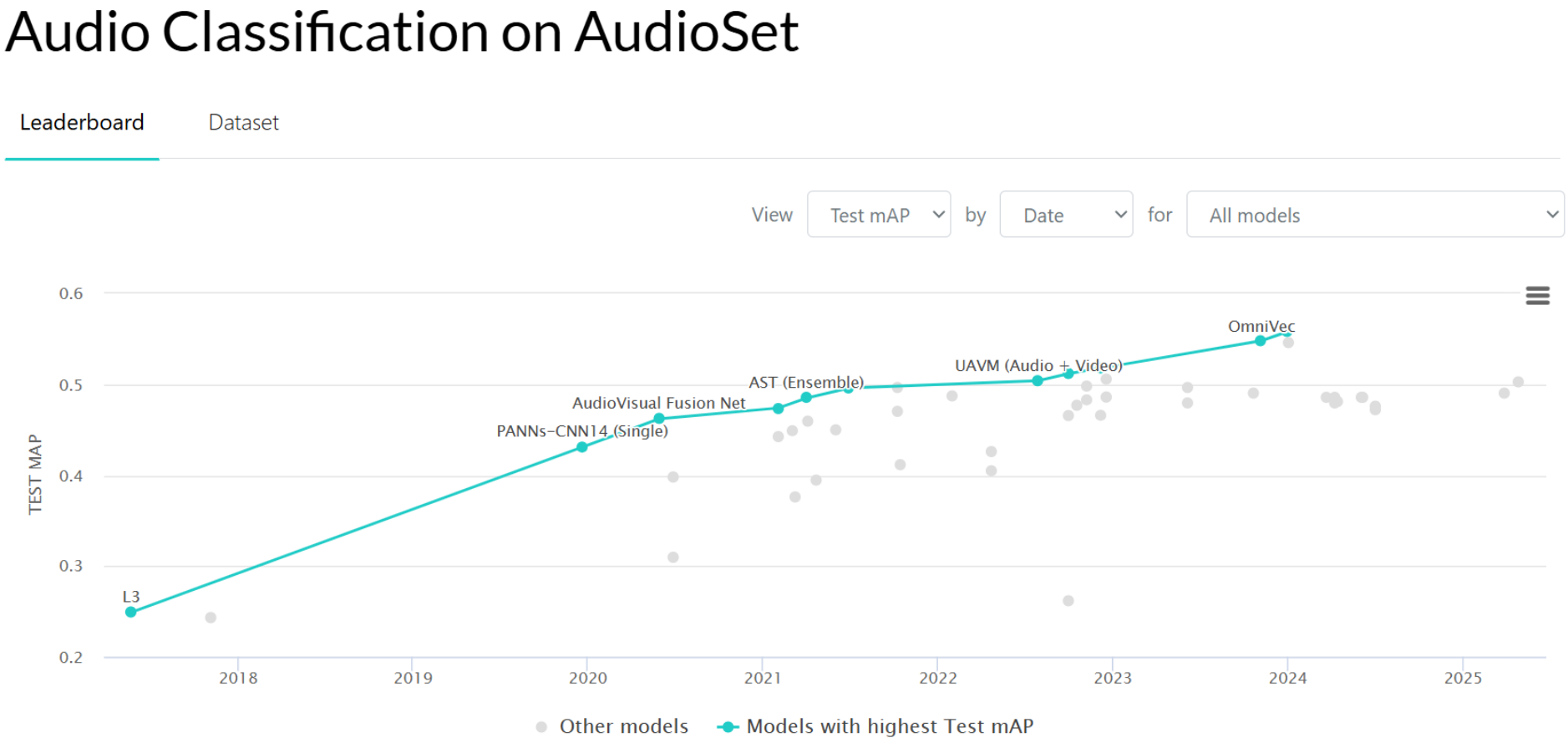
(Large) Audio Models

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
- SSLAM (Alex et al, 2025): self-supervised learning from audio mixtures

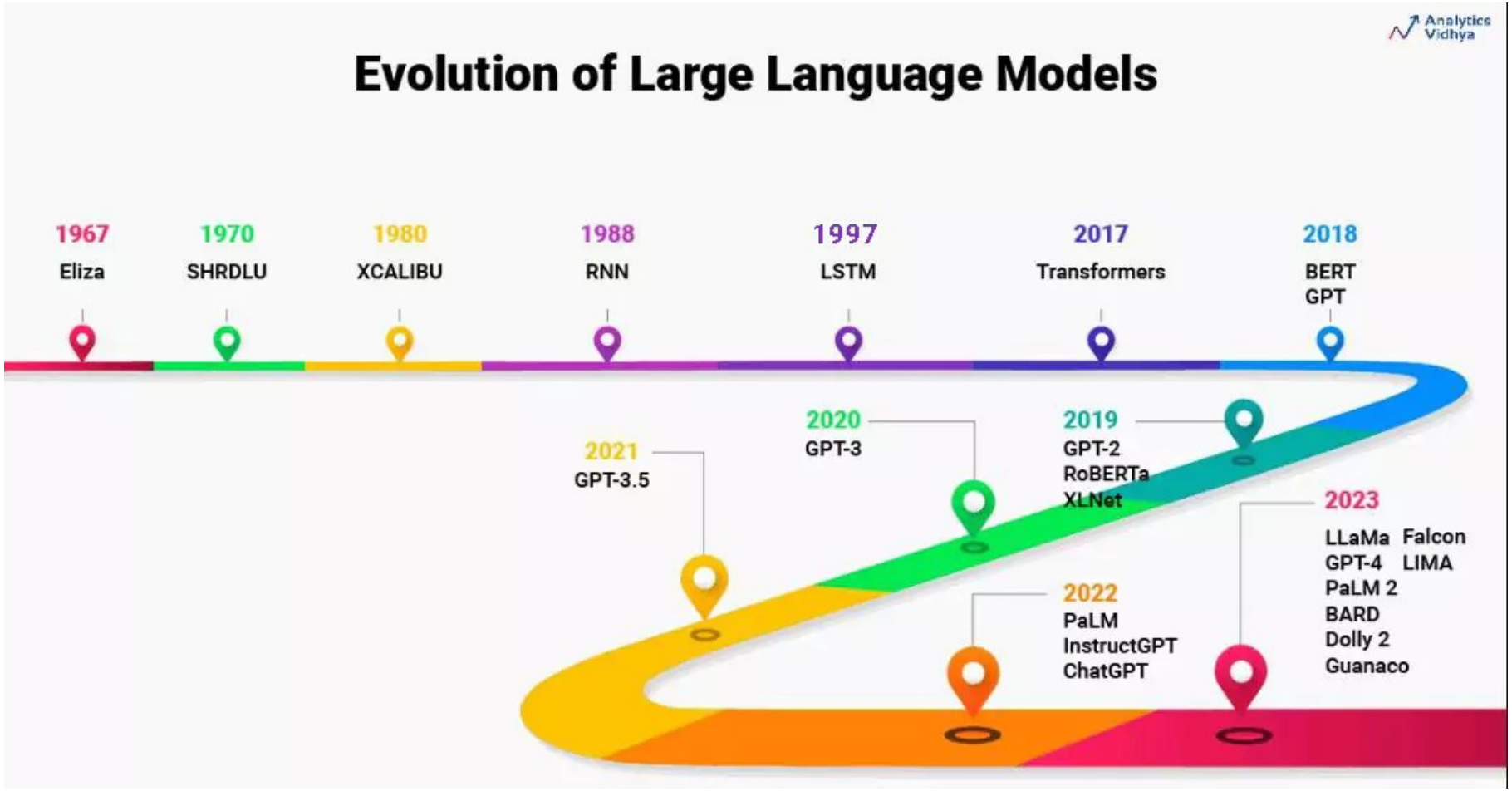
mean Average Precision (mAP) on AudioSet: **31.4** (2017) -> **55.8** (2025)

(Large) Audio Models



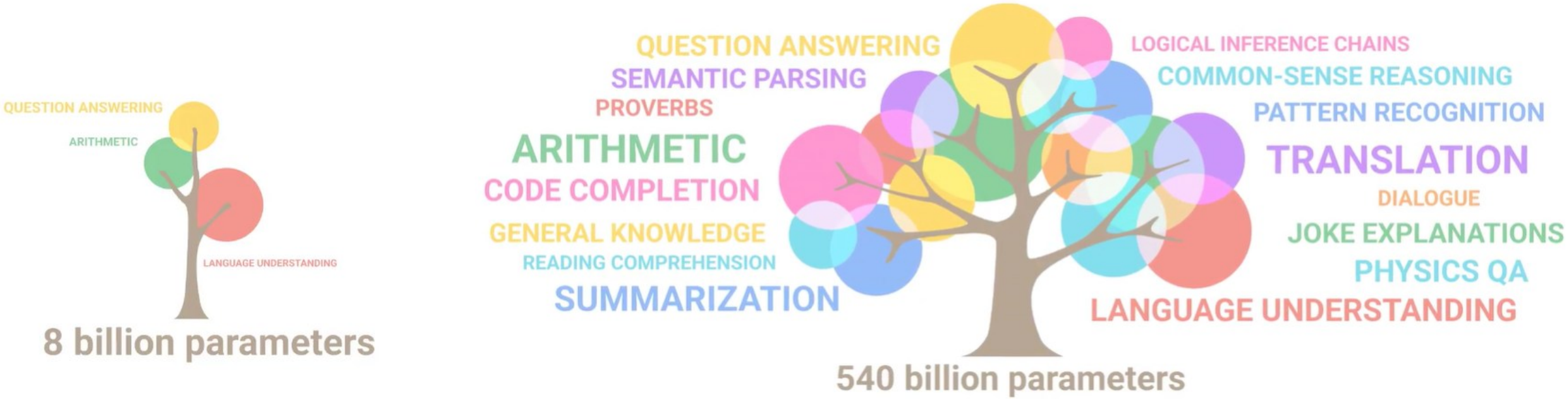
Leaderboard:
https://paperswithcode.com/sota/audio-classification-on-audioset
https://www.codesota.com/audio/classification

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?

UNIVERSITY OF SURREY

- 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

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/AudioLDM2/Re-AudioLDM/WavJourney: TTA generation
AudioSep/FlowSep/FlowSep2: language guided audio source separation
WavCraft/RFM-Editing/SFM-Adapter: text prompted audio/speech editing
APT-LLM: LLM based AQA and reasoning
DreamAudio: Customised text to audio generation
GCDance/TeMuDance: Language guided music to dance generation

(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)
- Audio-FLAN (Xue et al, 2025)
- SeaBench-Audio (Liu et al, 2025)
- MusicSem (Salganik et al, 2026)

- **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)
- AudioMCQ (He et al, 2025)
- Audio Flamingo 3 (Goel et al, 2025)
- Jamendo-MT-QA (Koh et al, 2026)

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 LLAMs

UNIVERSITY OF SURREY

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:

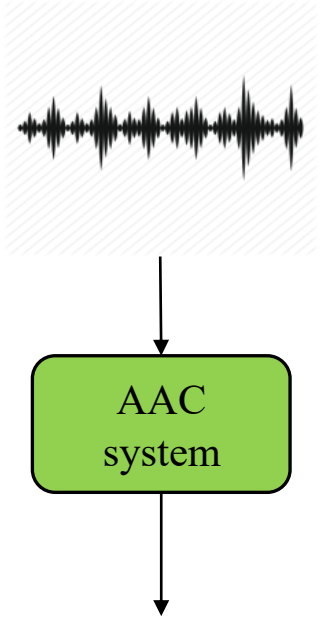
- Accessibility, voice assistants, content creation, and human-computer interaction

Typical Methods for Fusing Audio and Language

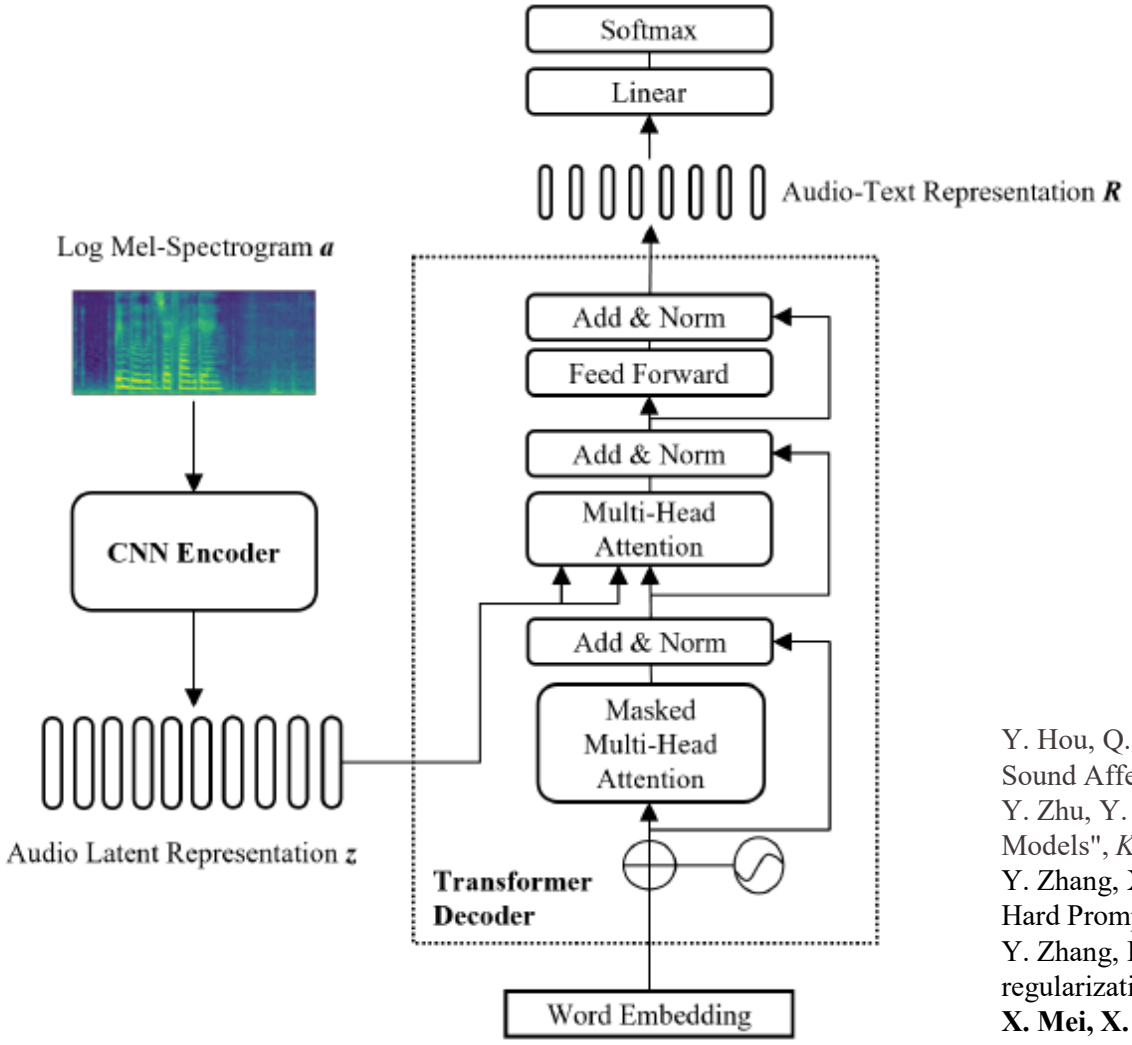
- Aligning audio-texts with contrastive pretraining
  - Examples: CLAP, CLAP-LAION, AudioCLIP, Wav2CLIP, T-CLAP, etc.
- Tokenizing audio and texts, then followed by LLMs
  - Examples: Moshi, VITA, LSLM, Voicebox, FunAudioLLM, LauraGPT, etc.
- Fusing embeddings with cross-attention
  - Examples: Q-former
- 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.



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.

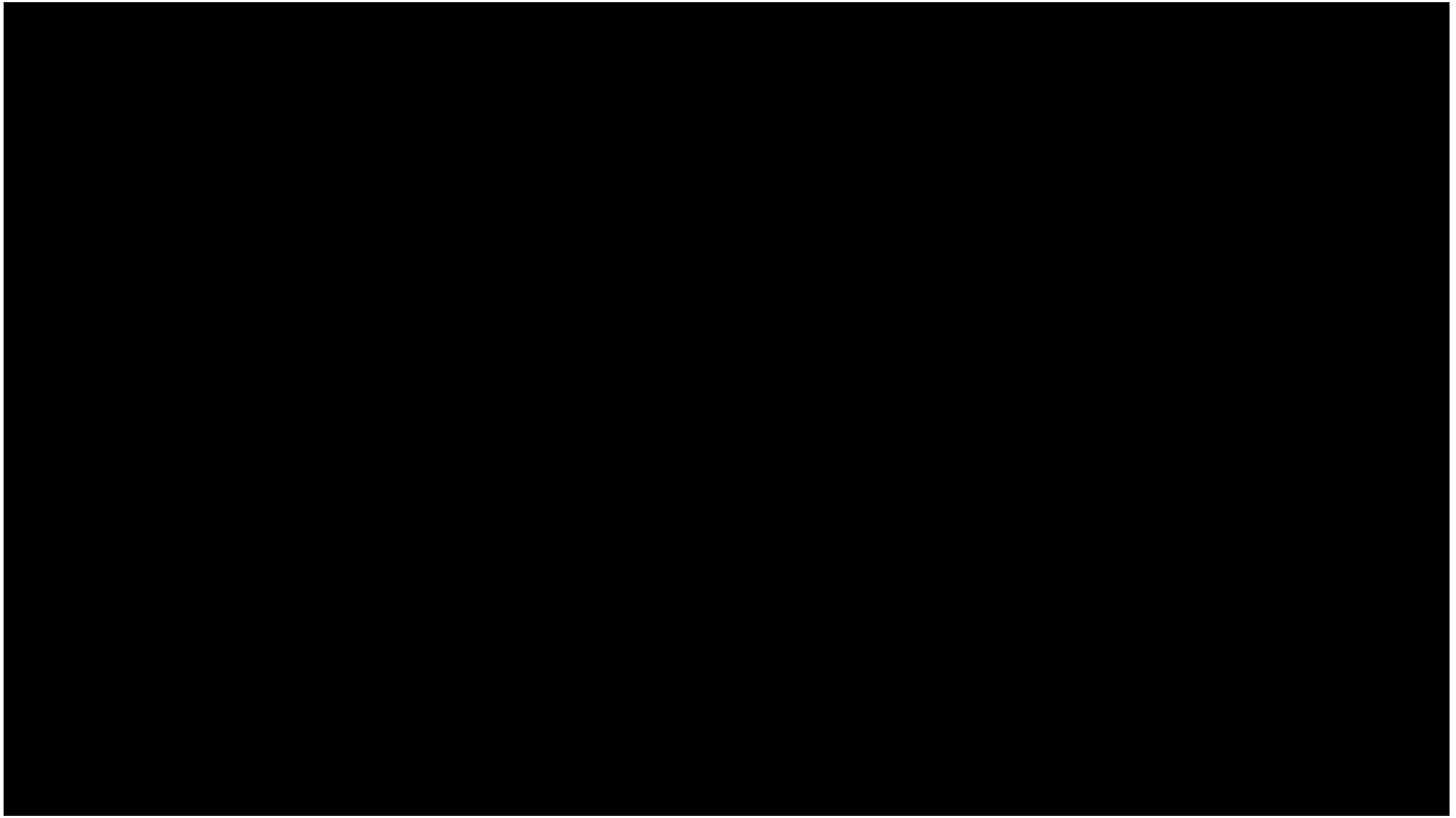
Y. Zhang, X. Xu, R. Du, H. Liu, Y. Dong, Z.-H. Tan, W. Wang, and Z. Ma, "Zero-Shot Audio Captioning Using Soft and Hard Prompts," *IEEE Transactions on Audio Speech and Language Processing,* vol. 33, pp. 2045 - 2058, May 2025.

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 LALMs 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)?**

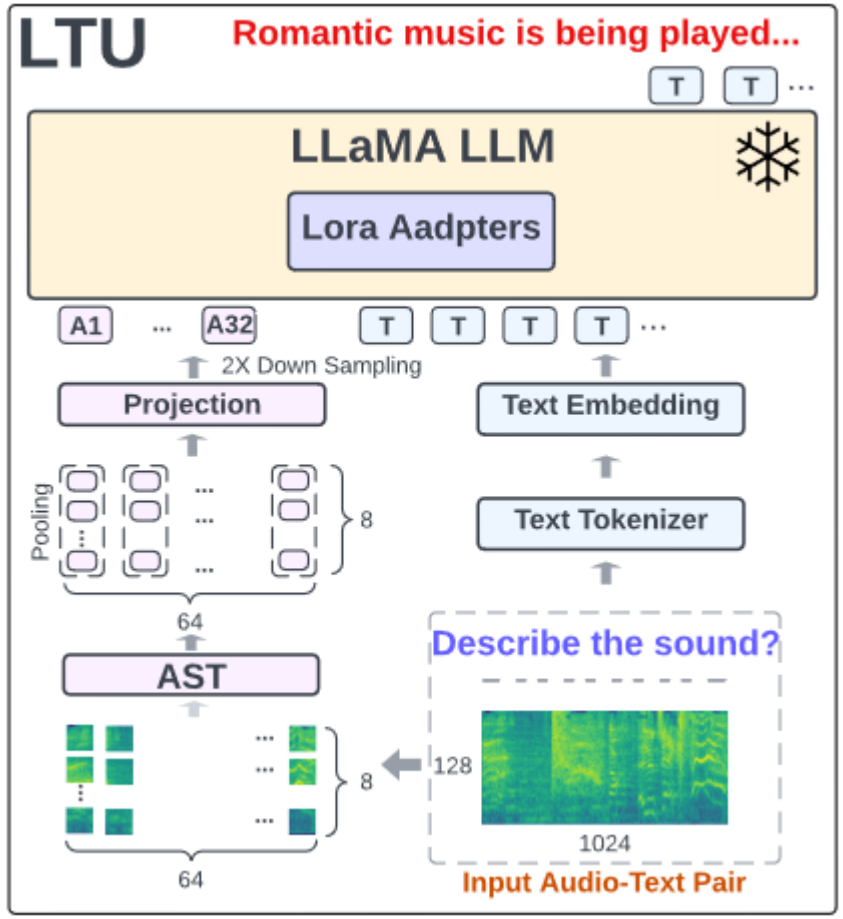
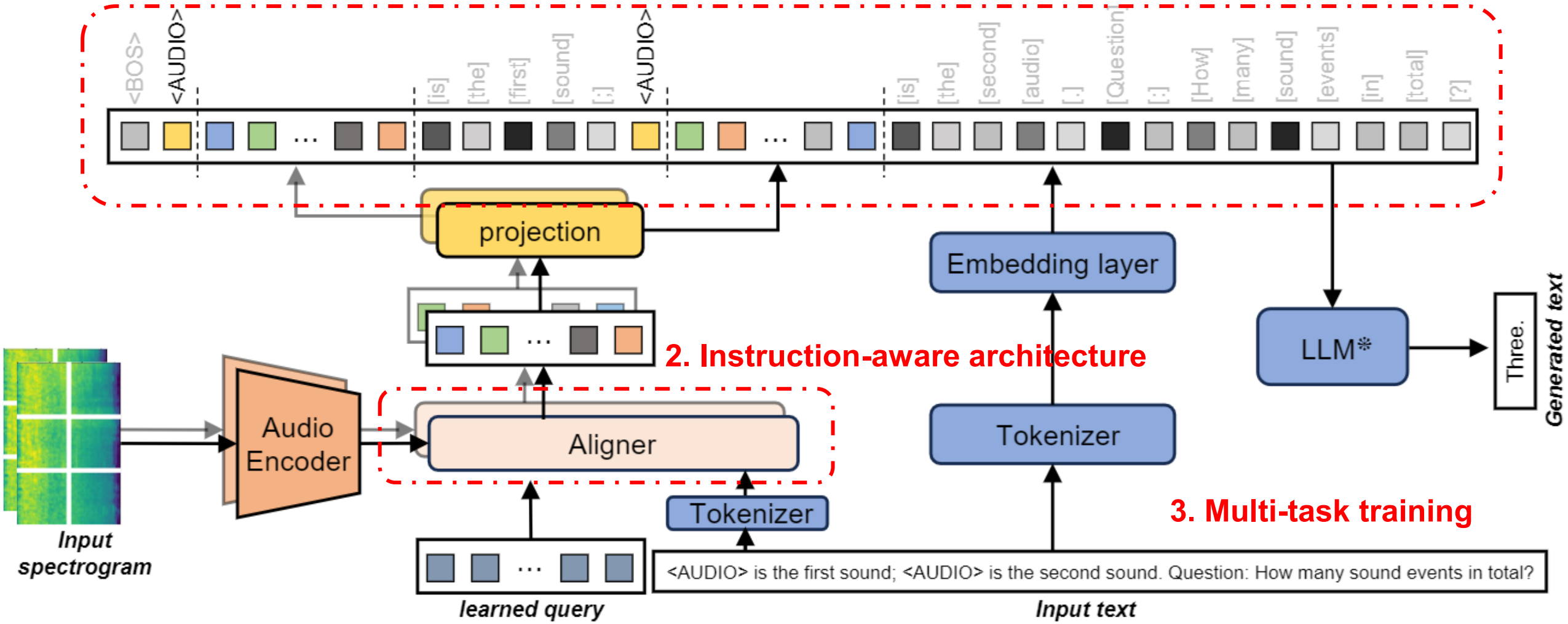


Fig. An example of an audio LLM structure (LTU (Gong et al, 2023)).

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

Audio Reasoning – **APT**

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



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)

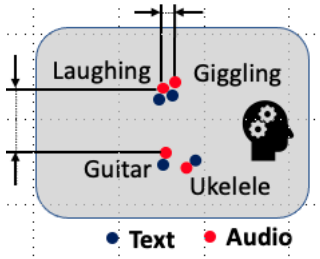
Text-to-Music Generation

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

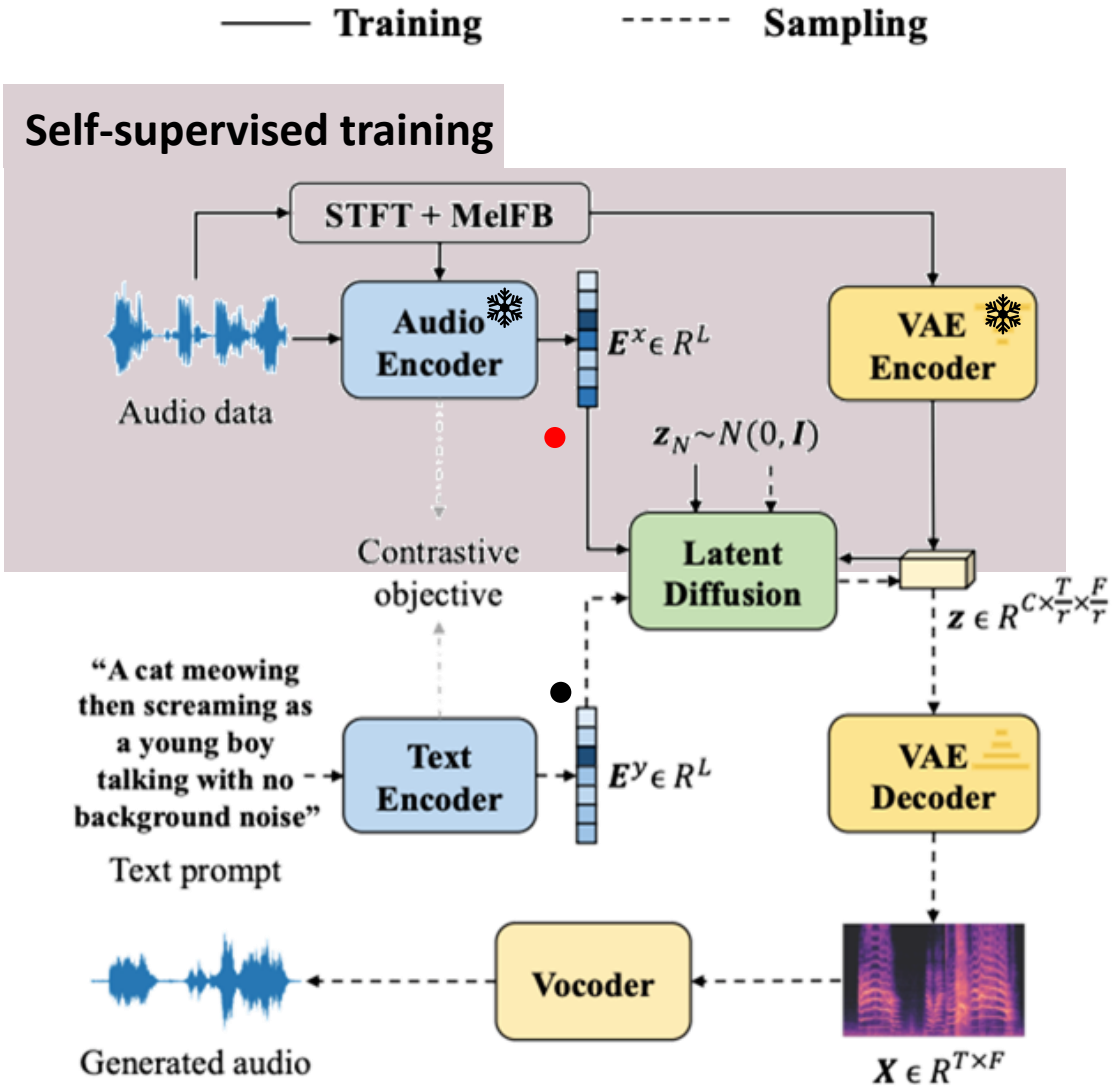
Others

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

AudioLDM

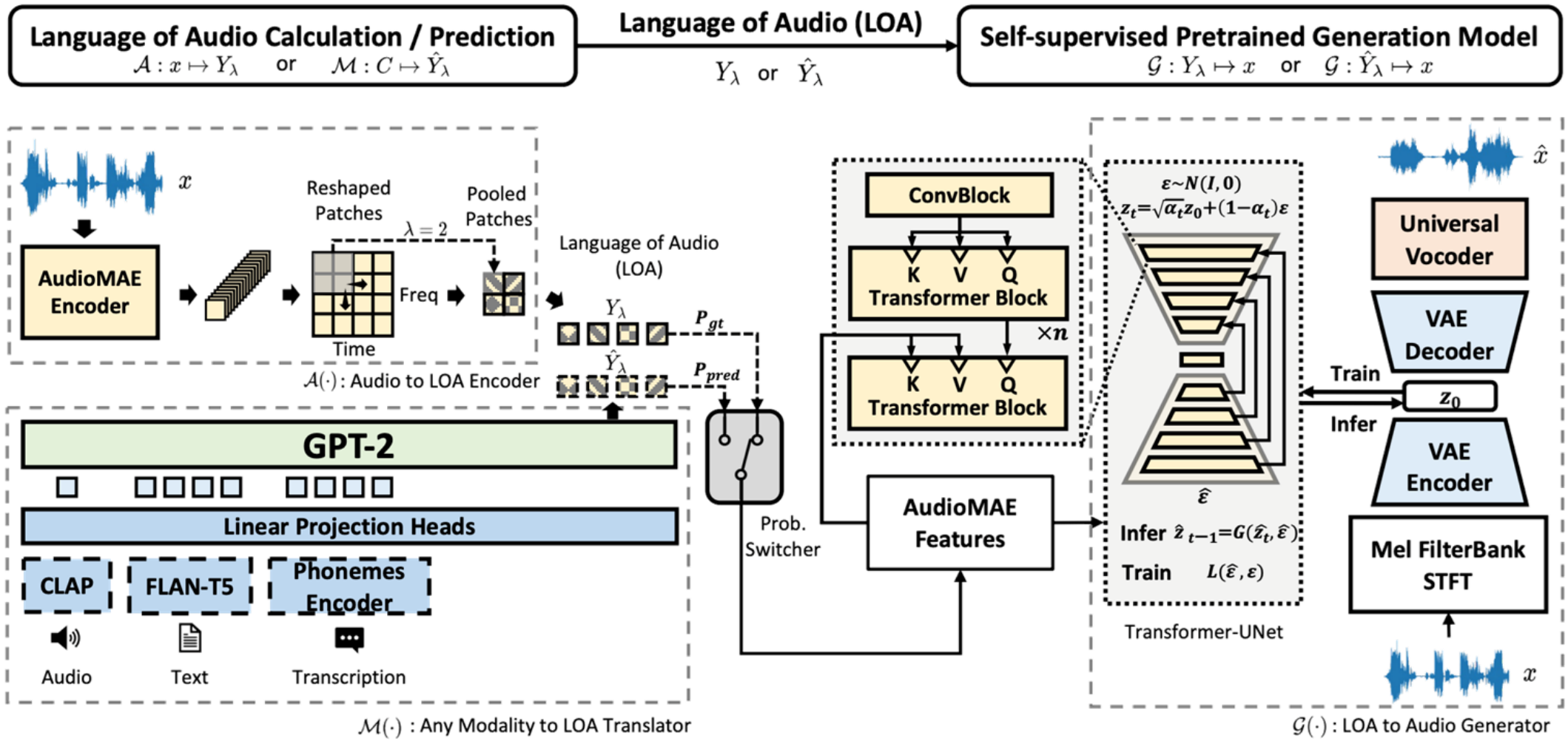


1. **Contrastive Language-Audio Learning (CLAP) Encoders**
   - Align audio and text in one space.
2. **Latent Diffusion Models**
   - Learn to generate VAE latent conditioned on CLAP embedding
3. **Mel-spectrogram Autoencoder**
   - Learn latent representations.
4. **Mel-to-Waveform Vocoder**
   - Reverse 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



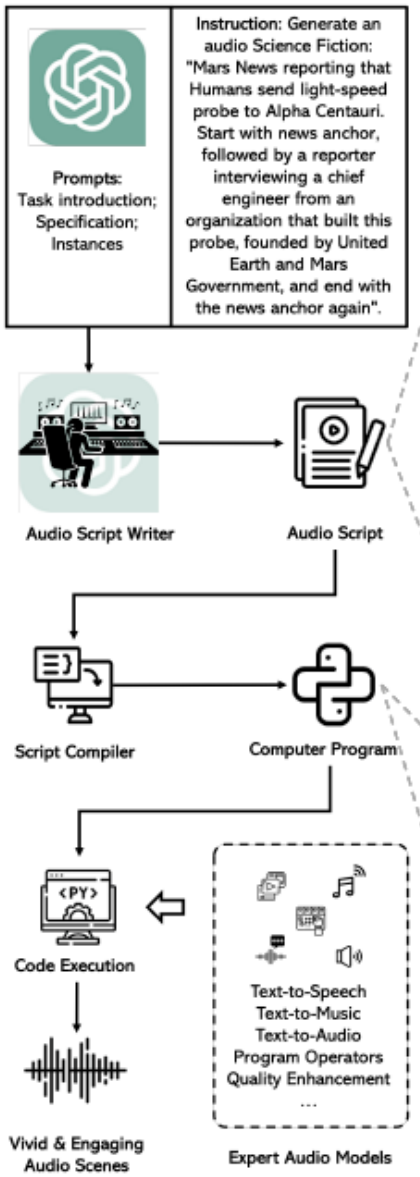
H. Liu , Y. Yuan , X. Liu , X. Mei , Q. Kong , Q. Tian , Y. Wang , W. Wang, Y. Wang , M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, pp. 2871-2883, 2024. [PDF] [code]

AudioLDM 2 - Demo

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/

WavJourney – Storytelling



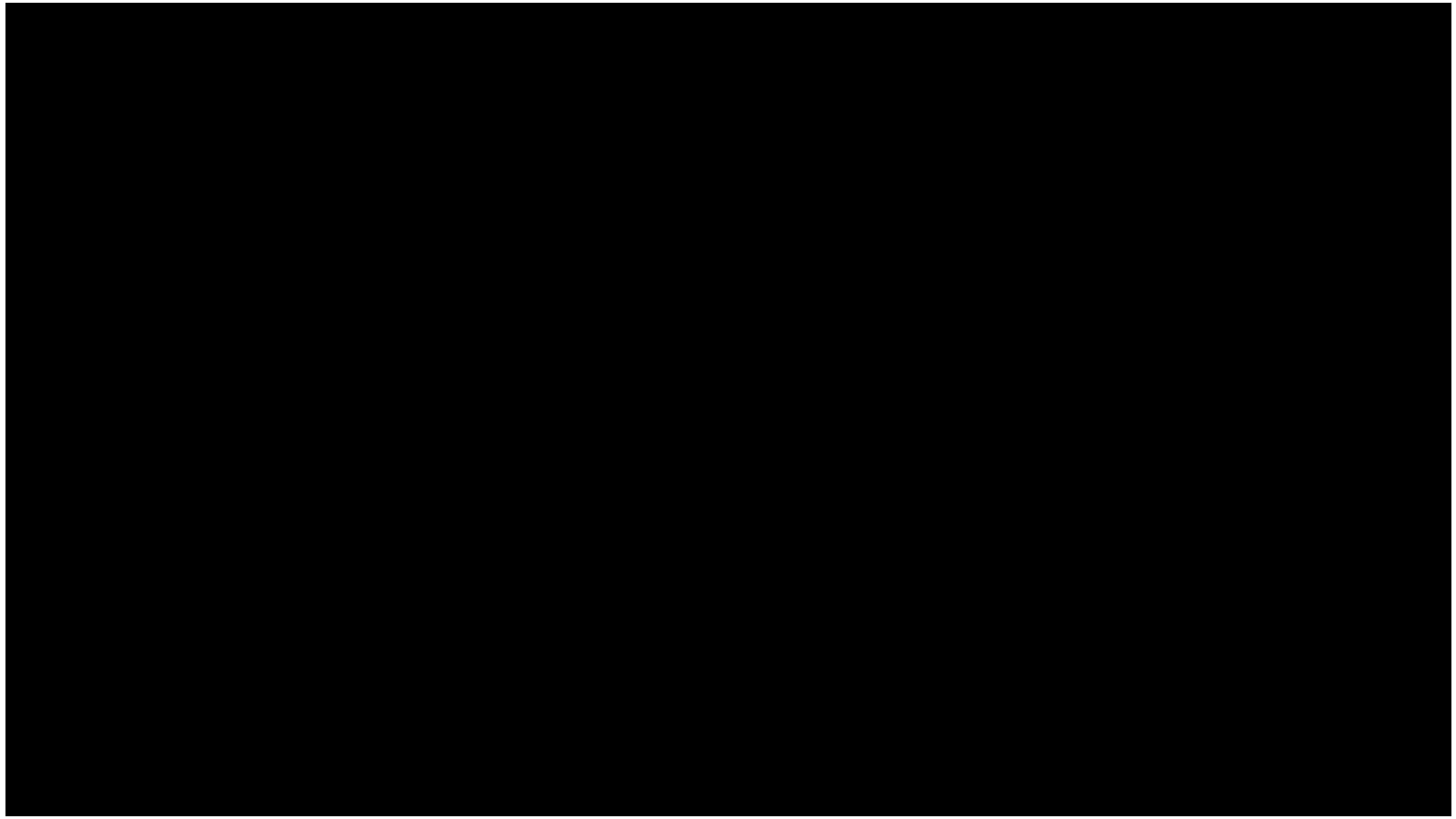
Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect 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

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

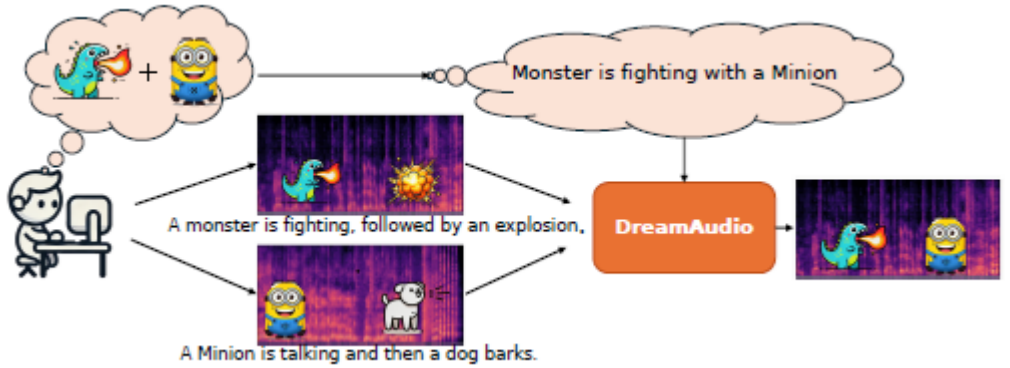
Script Compiler output (simplified pseudo code)

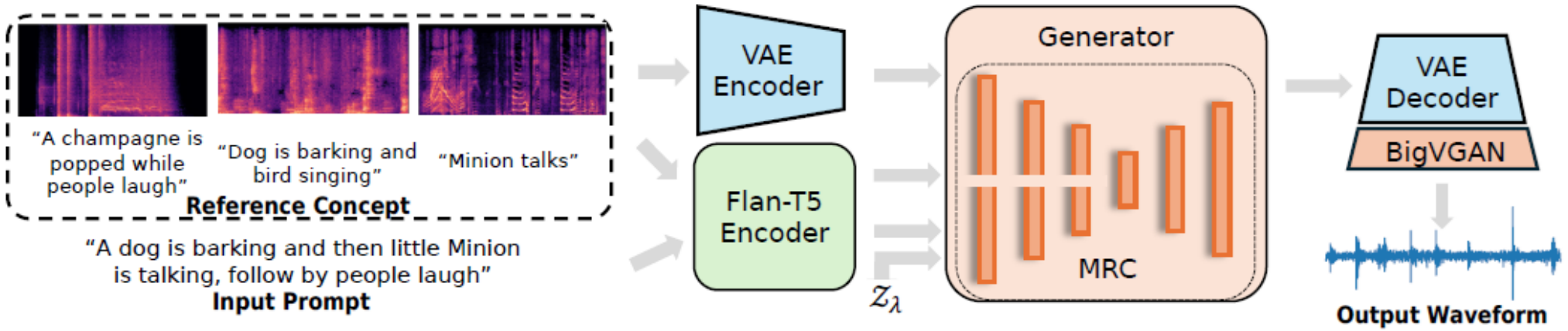
WavJourney – Sound Demo for Science Fiction Storytelling



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

DreamAudio – Customised Text to Audio Generation





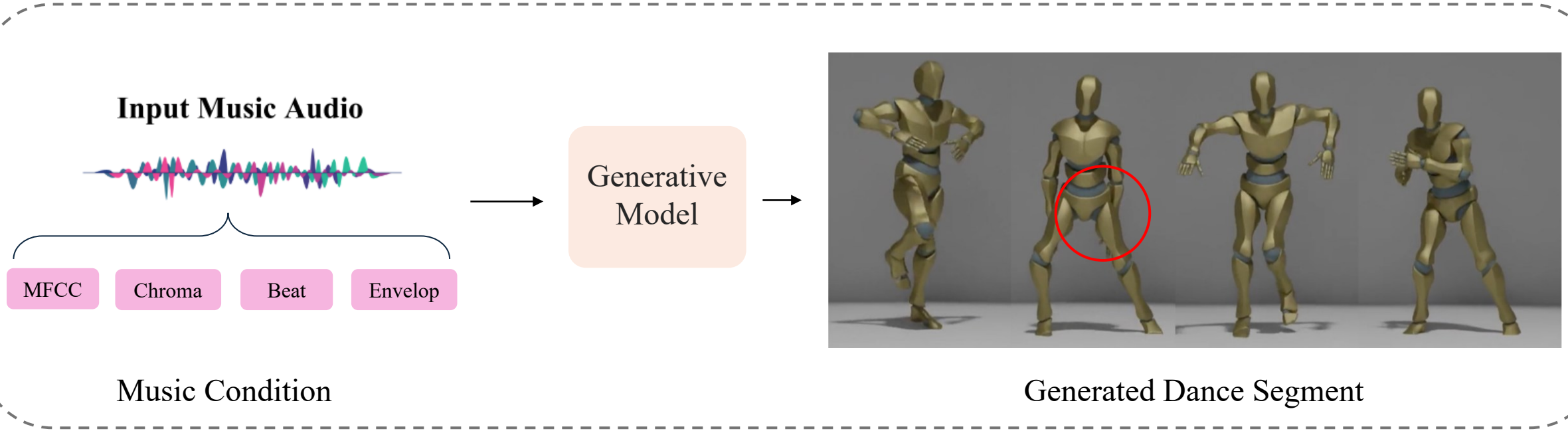
Y. Yuan, X. Liu, H. Liu, X. Kang, Z. Chen, Y. Wang, M. D. Plumbley, W. Wang, "DreamAudio: Customized Text-to-Audio Generation with Diffusion Models", *IEEE Transactions on Audio Speech and Language Processing*, 2026. [arXiv] [PDF] [code, data, & demo]

Controllable Music to Dance Generation

Focus only on hand-crafted musical features.

No genre information.

Based on 24-Joint dataset. Lack of hand motion.

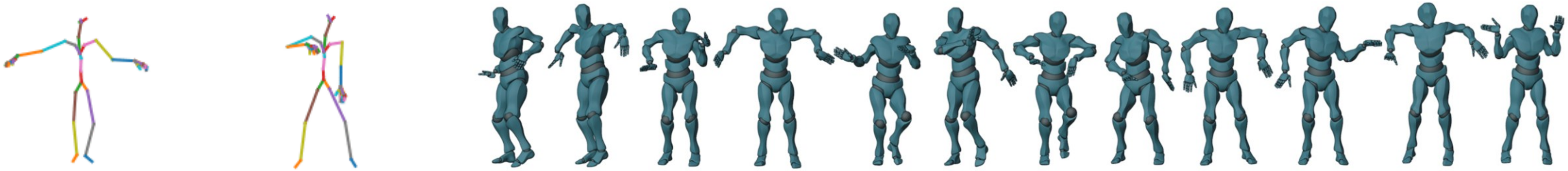


X. Liu, Z. Feng, D. Kanojia, W. Wang, "DGFM: Full Body Dance Generation Driven by Music Foundation Models", *NeurIPS 2024 Workshop on Audio Imagination*, 2024.

GCDance - Demo

This is a "HanTang" type of music.

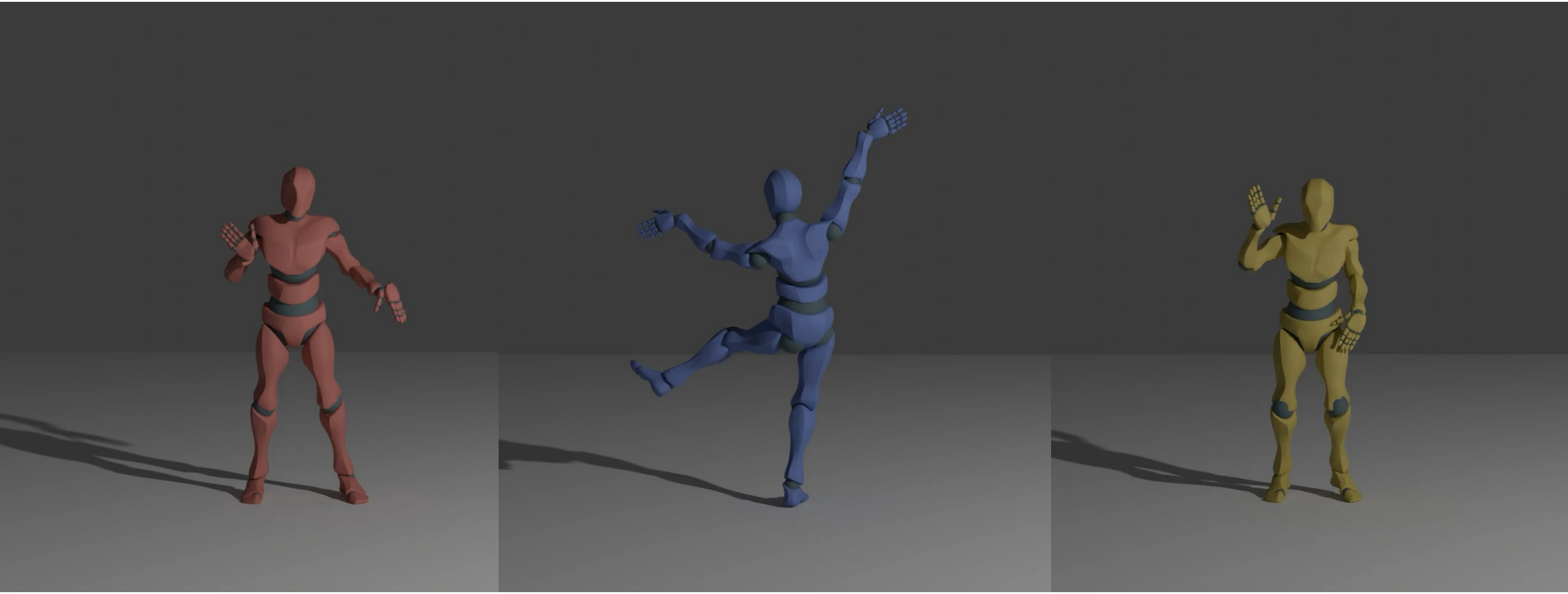
This is a "Dai" type of music.



This is a "Hiphop" type of music.

X. Liu, X. Dong, S. Qian, D. Kanojia, W. Wang, and Z. Feng, "GCDance: Genre-Controlled Music-Driven 3D Full Body Dance Generation," *IEEE Transactions on Multimedia*, 2026. [arXiv] [PDF] [code, data, & demo]

GCDance – More Demo

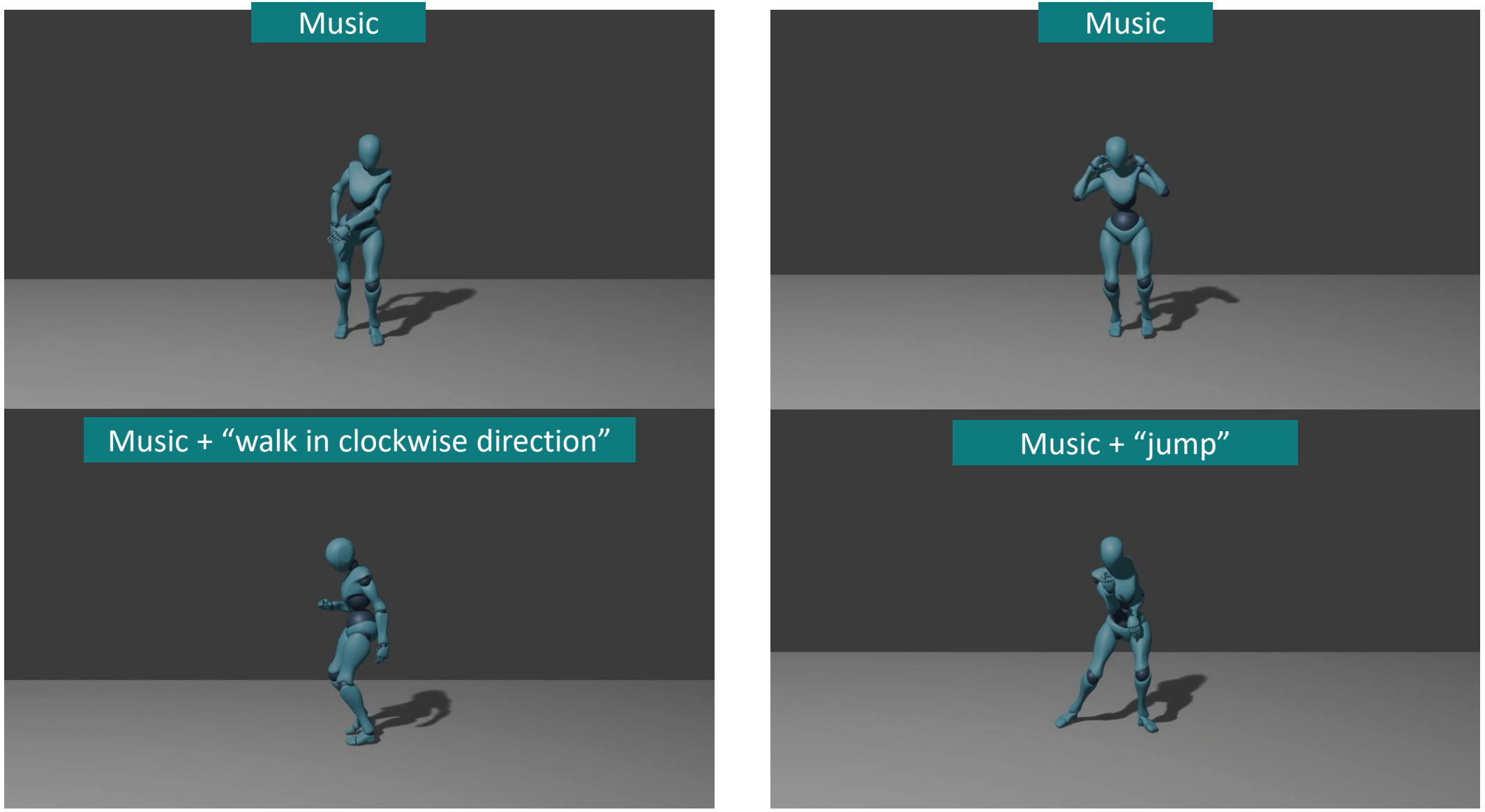


Breaking

HanTang

Korean

TeMuDance-Demo

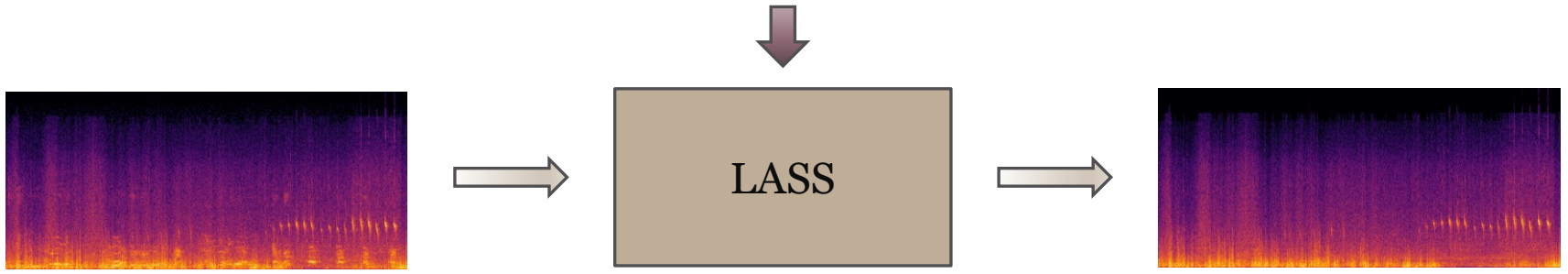


X. Liu, D. Kanojia, W. Wang, Z. Feng, "TeMuDance: Contrastive Alignment-Based Textual Control for Music-Driven Dance Generation," *ACM MM*, 2026.

Task 4: Language-Queried Audio Source Separation

- LASS – Separate a **target source** from an audio mixture based on the **natural language descriptions** of the target source
- First attempt bridging audio source separation and natural language processing
- Support input arbitrary text to separate desired sound sources

Language Query: A bird is chirping under the thunder storm

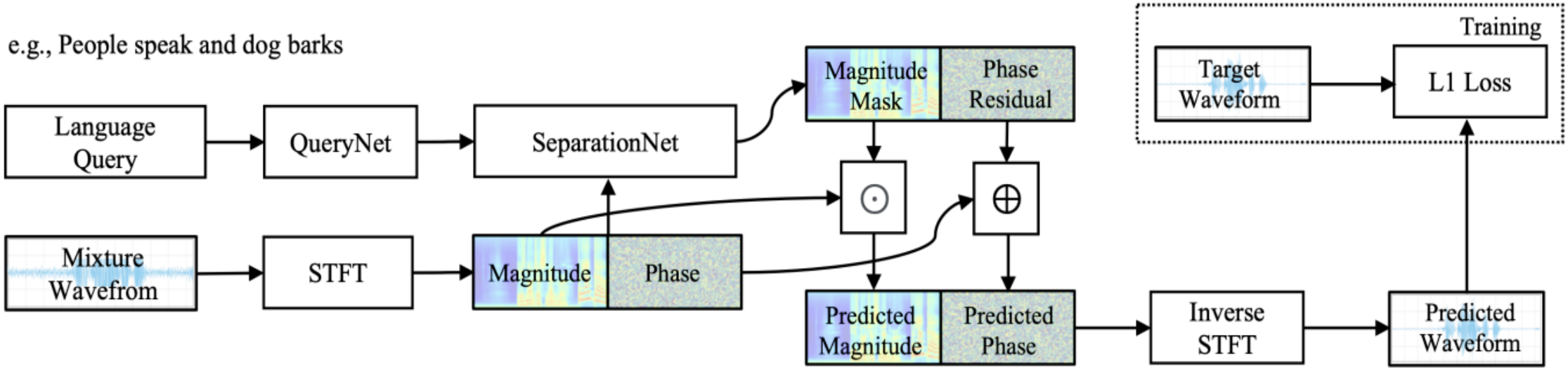


- **Existing methods**
  - LASS-Net (Liu et al 2022): first LASS model
  - SoundFilter (Kilgour 2022): exploiting audio supervision
  - CLIPSep (Dong et al. 2023): exploiting visual supervision
- **Challenges**
  - Small scale of data available for training
  - Open-domain source separation with texts
  - Overlapping sound events
  - Processing artefacts and incomplete separation

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.

Task 4: **AudioSep** - Architecture

- CLAP/CLIP + ResUNet, trained with **14,000** hours of multimodal data
- A foundation model for open-domain sound separation with texts
- Impressive zero-shot performance in separating speech, music, sounds



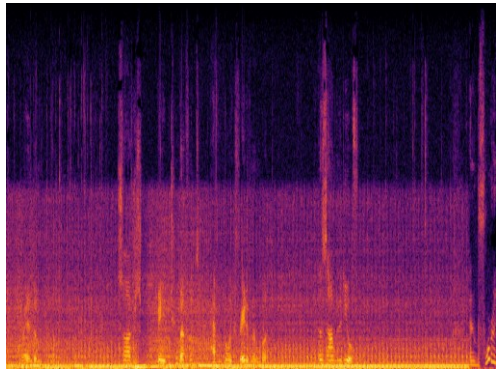
X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *IEEE Transactions on Audio Speech and Language Processing*, vol. 33, 458--471, 2025. [PDF] [code]

Demo: https://huggingface.co/spaces/Audio-AGI/AudioSep

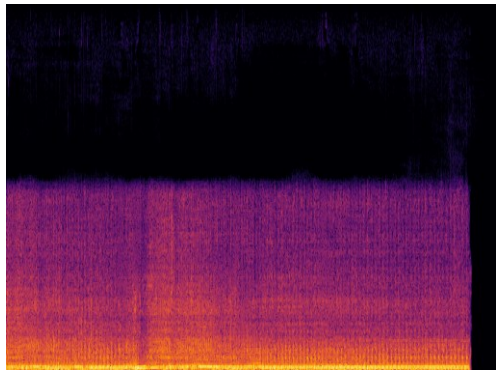
Task 4: AudioSep - Demo

Human query: “The engine sound of a vehicle”

Mix

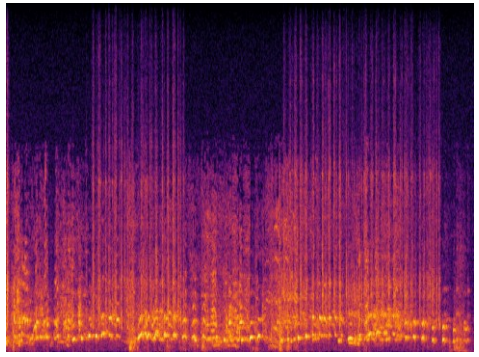


Separated

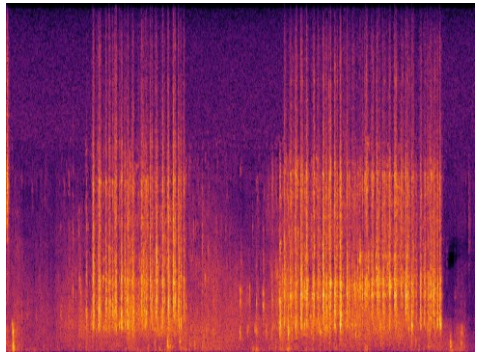


Human query: “The sound of hitting the keyboard”

Mix



Separated



Task 4: **FlowSep** - Motivation

Existing Ideas:

- Discriminative approaches.
- Time-frequency masking on spectrogram to remove the noise sound sources.

Challenges:

- Challenging with overlapping sound events.
- Excessive and insufficient masking leads to **artifacts**, including spectral holes and incomplete separation.

A “New” Idea:

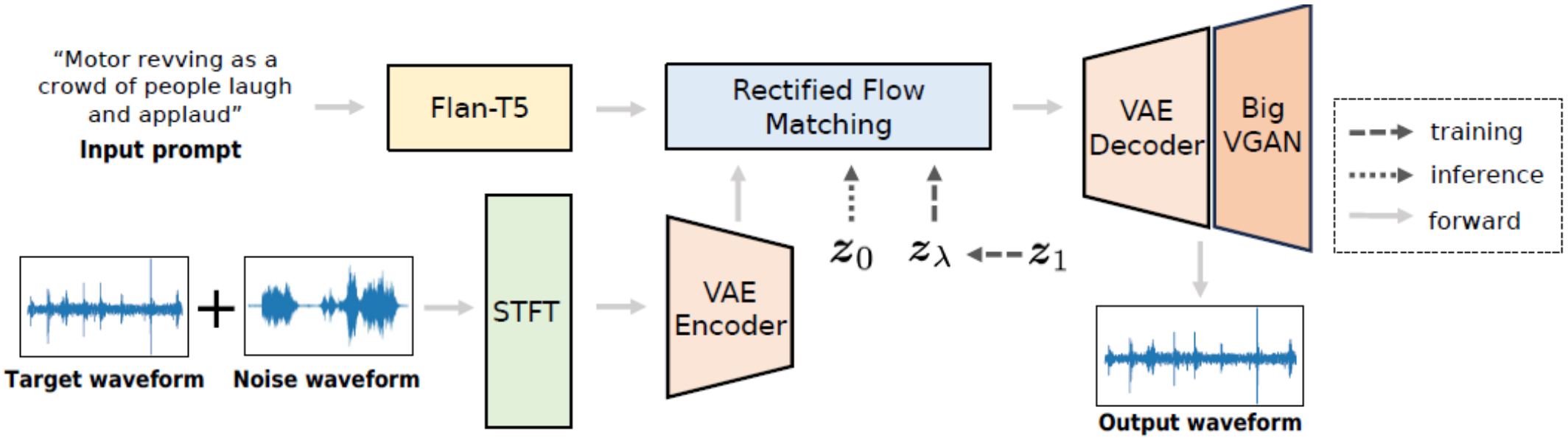
Using the generative approaches.

Diffusion-based generation framework with rectified flow matching.

Separation system by generating new audio samples with the noisy clips and text prompts as a condition.

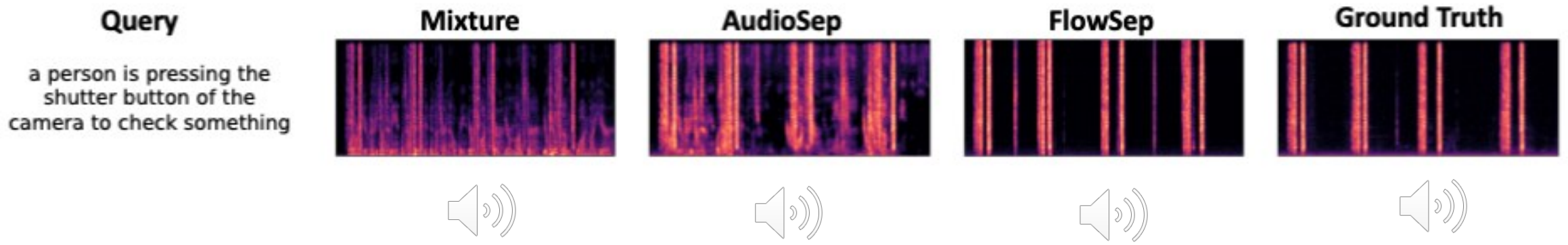
Task 4: **FlowSep** - Architecture

- Text-to-audio generation model as the backbone, Rectified Flow Matching for feature generation.
- Extended VAE latent space to integrate the noise audio feature.
- Flan-T5 text encoder, VAE latent decoder and BigVGAN vocoder.



Y. Yuan, X. Liu, H. Liu, M.D. Plumbley, and W. Wang,"FlowSep: Language-Quried Sound Separation with Rectified Flow Matching" in *ICASSP 2025*.

Task 4: FlowSep - Demo

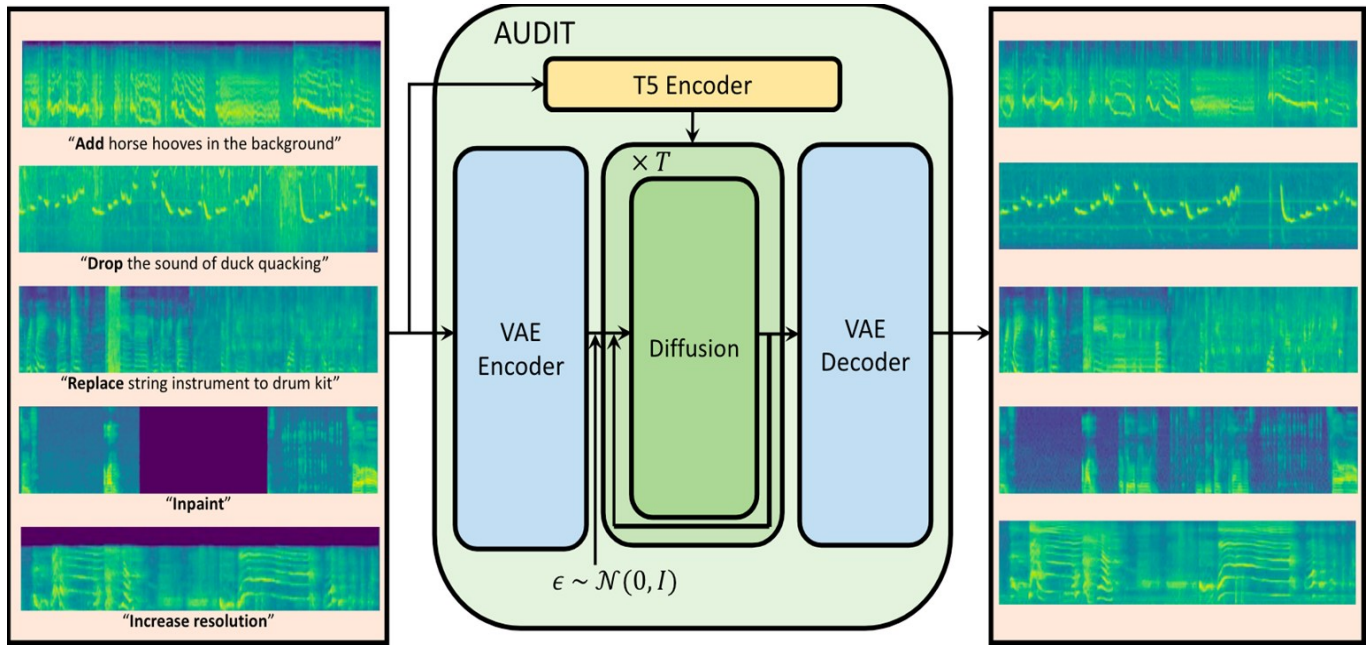


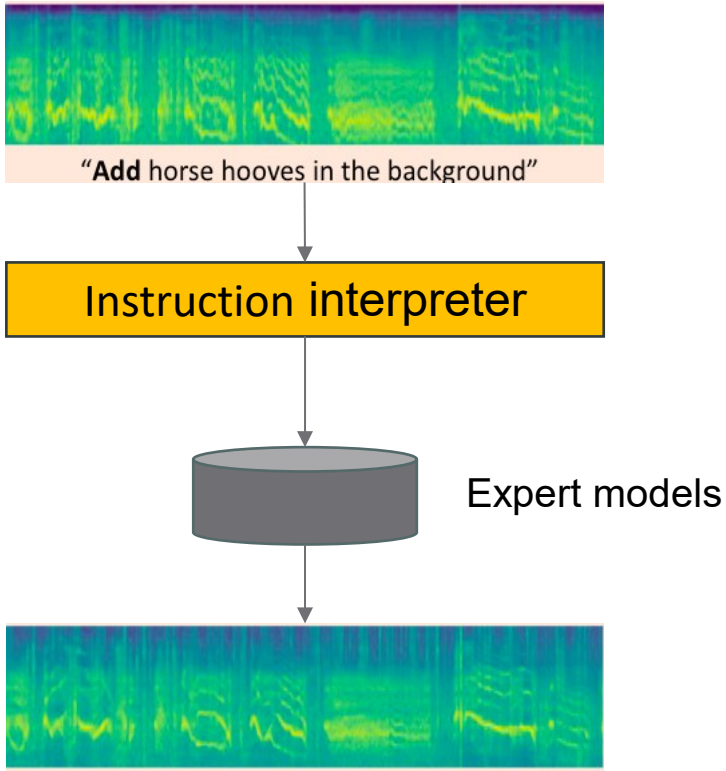
- Baseline models show incomplete separation with noticeable spectral gaps.
- FlowSep demonstrates promising capabilities in such situations.
- More demos please refer to https://audio-agi.github.io/FlowSep_demo/.

Y. Yuan, X. Liu, H. Liu, M.D. Plumbley, and W. Wang, “FlowSep: Language-Quried Sound Separation with Rectified Flow Matching" in *ICASSP 2025*.

Task 5: LLMs for Controllable Audio Editing

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.

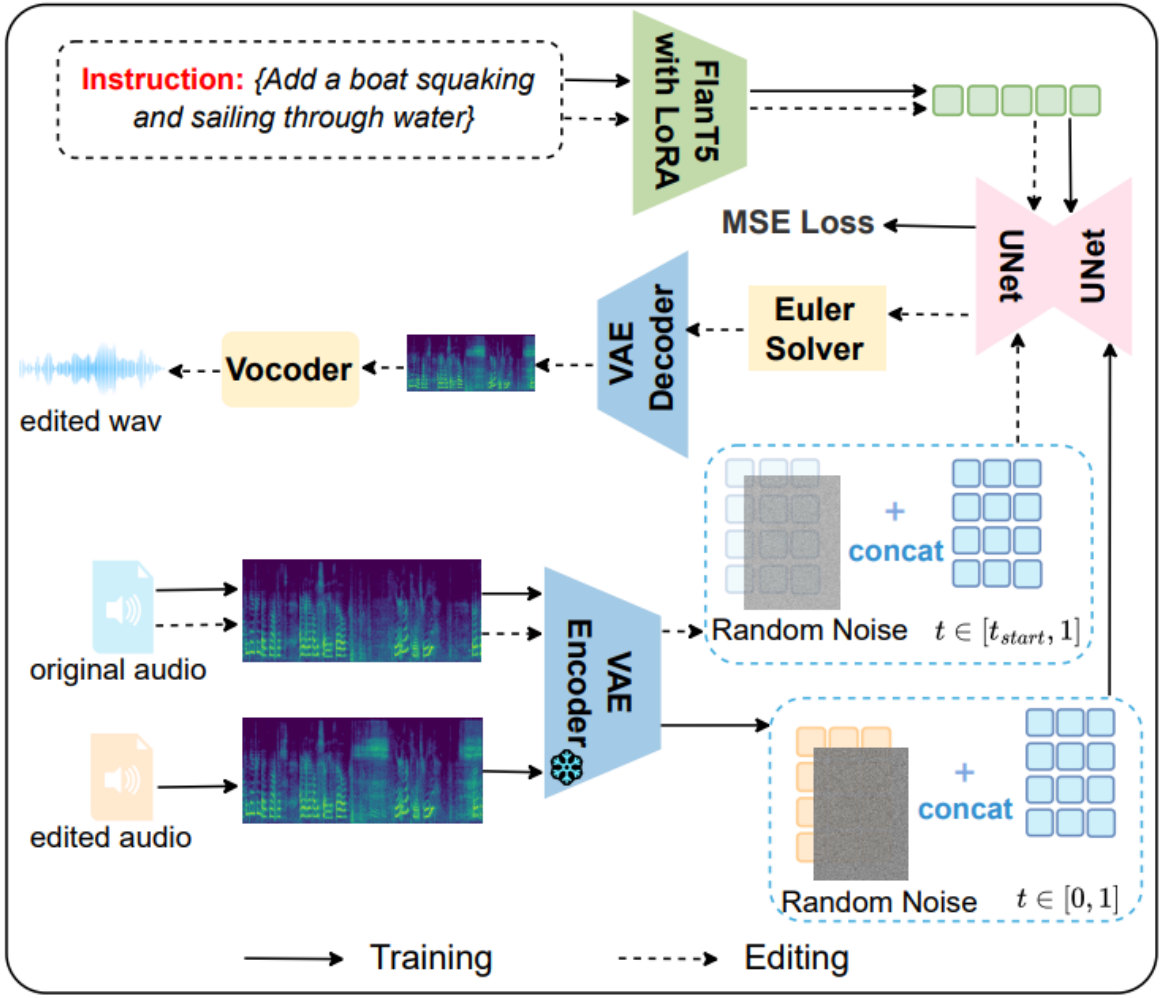




An example of end-to-end audio editing system: AUDIT

J. Liang, H. Zhang, H. Liu, Y. Cao, Q. Kong, X. Liu, W. Wang, M. Plumbley, and E. Benetos, "WavCraft: Audio Editing and Generation with Large Language Models", in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

RFM-EDITING – Controllable Audio Editing



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:

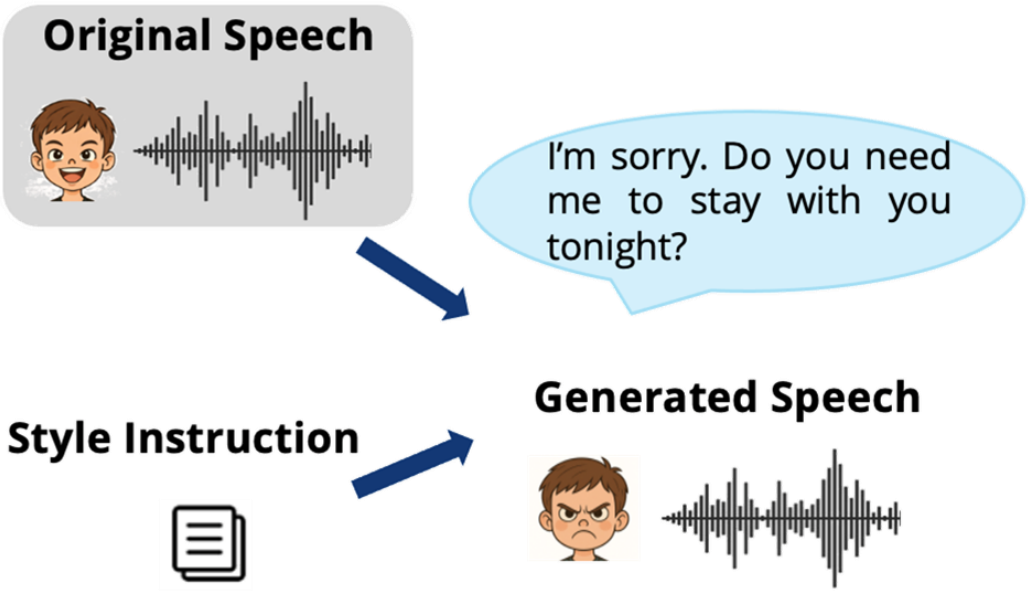
Original:

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. [PDF]

SFM-Adapter – Controllable Speech Editing

The goal of **Speech Style Edit:** ***Change How Speech Sounds, Not What It Says !***



Instruction-guided speech style editing: changes the speaking style (expressiveness and timbre) while preserving linguistic content.

Why It Matters

Human-Computer Interaction

More natural and engaging communication.

Voice Assistants

Hands-free help in daily life.

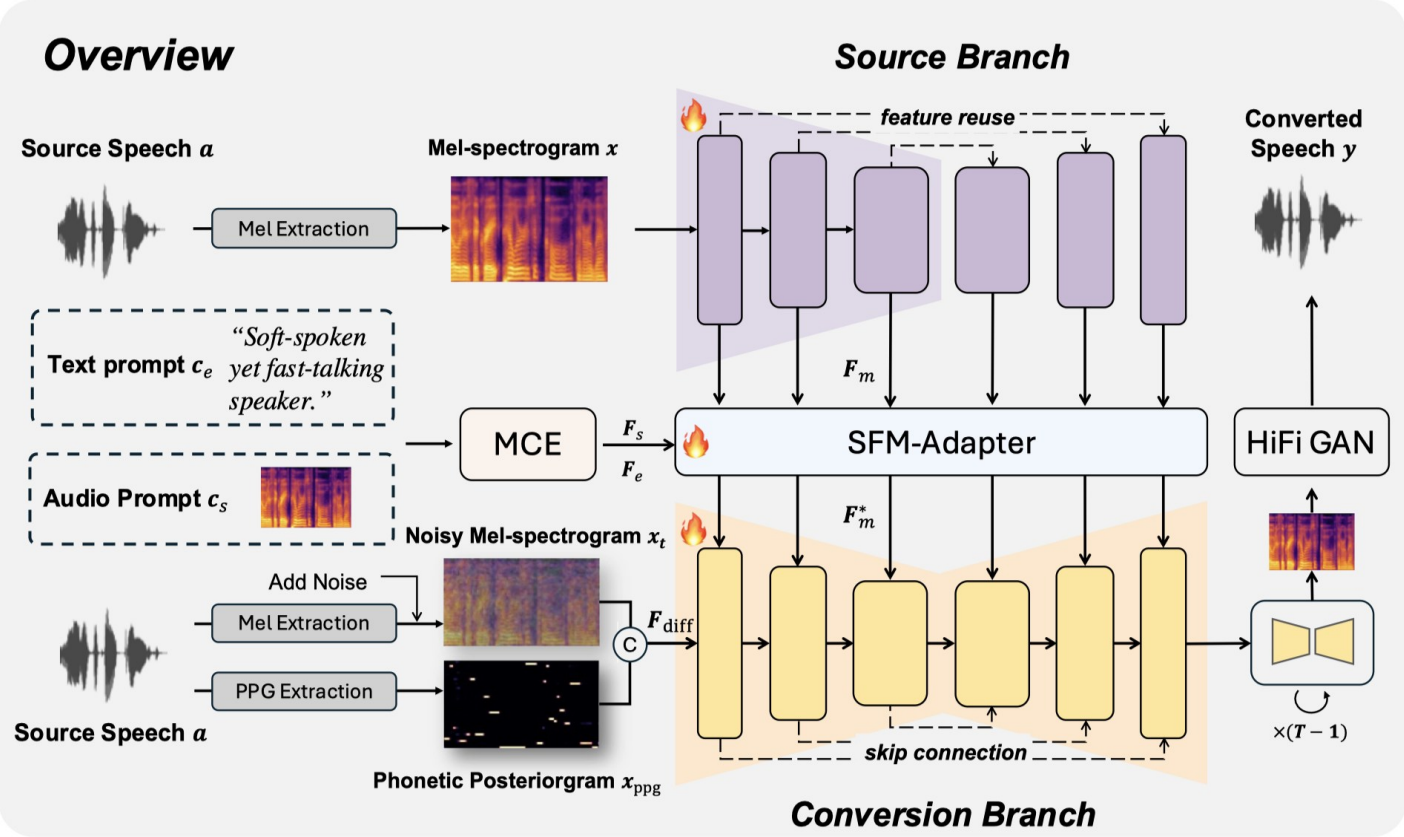
Digital Personas & Avatars

Bring virtual characters to life with expressive voices.

Y. Chen, H. Liu, Q. Chen, A. Singh, J. Zhao, W. Wang, P. J. B. Jackson, M. D. Plumbley, "SFM-Adapter: Style-aware Feature Manipulation Adapter for Speech Style Editing", *IEEE Transactions on Audio Speech and Language Processing*, 2026. [PDF] [code, data, & demo], vol. 34, pp. 3132-3145, 2026.

Method

Main framework: a reconstruction-based diffusion architecture that can incorporate external style conditions (*e.g., natural language and audio*).



Main components:

1. **Source Branch:** Extracts multi-scale features F_m from the input mel-spectrogram.
2. **Conversion Branch:** A diffusion-based U-Net model that synthesizes target speech by combining style features and PPG features.
3. **MCE:** Encodes the text and audio prompts into expressive features F_e and timbre features F_s
4. **SFM-Adapter:** Manipulates the source features using the requested target styles at multiple U-Net layers.

SFM-Adapter - Demos

Source Speech	Audio Prompt	Text Prompt	Target	SFM-Adapter
		Speaking with intention, her unhappy voice held a high pitch and low energy.		
		A fatigued disconsolate male voice, characterized by a slow speaking speed and a deep, low pitch, emits a subtle energy, resulting in a soothing audio experience that exudes tranquility.		

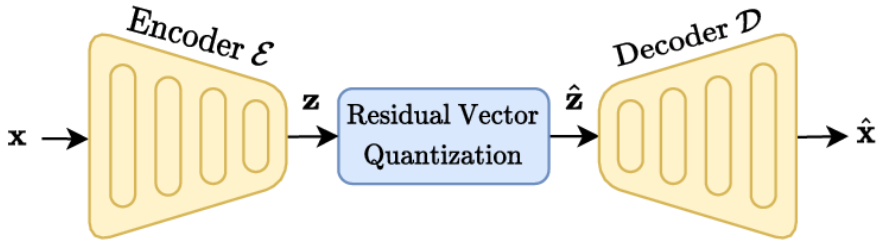
Y. Chen, H. Liu, Q. Chen, A. Singh, J. Zhao, W. Wang, P. J. B. Jackson, M. D. Plumbley, "SFM-Adapter: Style-aware Feature Manipulation Adapter for Speech Style Editing", *IEEE Transactions on Audio Speech and Language Processing*, 2026. [PDF] [code, data, & demo], vol. 34, pp. 3132-3145, 2026.

Task 6: Neural Audio Codec

Audio Codec:

- A device or software that encodes or decodes digital audio data for transmission, storage, or playback.
- **Co**mpressor-**dec**ompressor → “Codec”
- MP3, FLAC, AAC, Vorbis, etc.

Neural Audio Codec:



SoundStream (Zeghidour et al. 2021)

Encodec (Défossez et al. 2021)

Descript (Kumar et al. 2023)

HiFi-Codec (Yang et al. 2023)

SpeechTokenizer (Zhang et al, 2023)

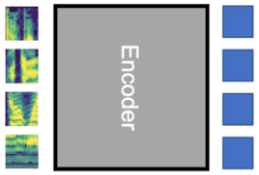
Codec Superb Benchmark (Wu et al. 2024)

- https://github.com/voidful/Codec-SUPERB

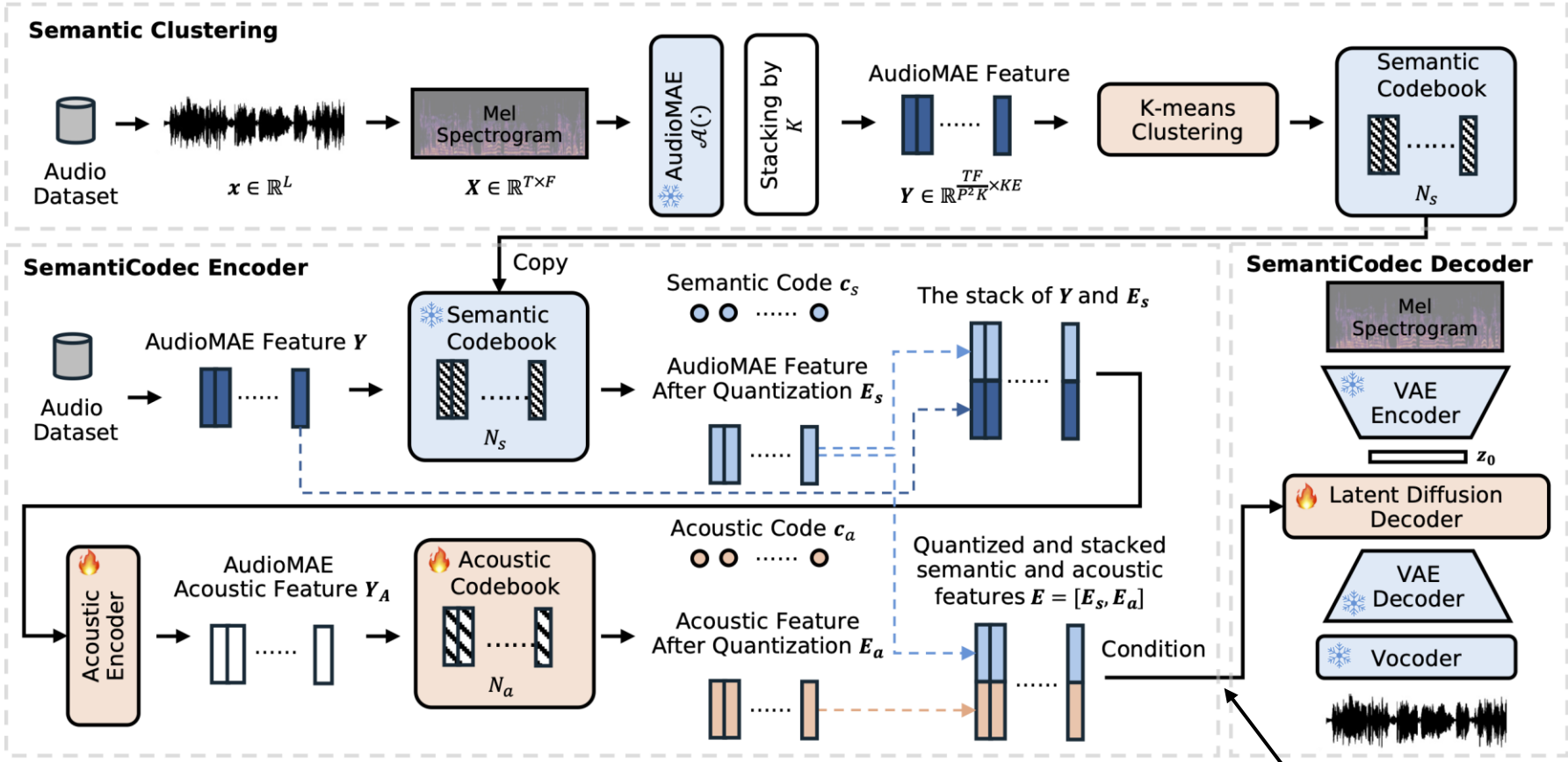
Task 6: **SemantiCodec**

Large scale k-means is challenging
AudioSet + Million Song Dataset + GigaSpeech
https://github.com/haoheliu/kmeans_pytorch

- 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



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

Task 6: SemantiCodec - 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/Sound-VECaps/AudioSetCaps**

Dataset	Quantity	Ave-Length	Vocabulary	Caption Source
Clotho	30K	11	4K	H
AutoCaps	57K	9	5K	H
LAION-Audio-630K	630K	7	311K	L
WavCaps	400K	8	24K	L
Auto-ACD	1.5M	18	20K	L+A+V
Sound-VECaps	1.6M	40	50K	L+A+V
AudioSetCaps	1.9M	28	21K	L+A

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, 2024. [PDF] [arXiv] [code] (https://github.com/XinhaoMei/WavCaps)

Y. Yuan, D. Jia, X. Zhuang, Y. Chen, Z. Chen, Y. Wang, Y. Wang, X. Liu, X. Kang, M.D. Plumbley, and W. Wang, "Sound-VECaps: Improving Audio Generation With Visual Enhanced Captions," in *ICASSP*, Hyderabad, India, April 6-11, 2025. [PDF]

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. [arXiv][code]

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)
  - Improving controllability/personalization/customisation of LALMs in various downstream tasks (e.g. generation and captioning)
  - Developing LALMs for long form audio activity understanding, reasoning
  - Developing LALMs for spatial audio generation and reasoning
  - Leveraging physics-based model + data driven models

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

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

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

SFM-Adapter: https://ychenn1.github.io/SFM-Adapter/

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

GCDance: paper, data & code:
https://xinranliu7715.github.io/gcdance/