Audio Processing with Natural Language Intelligence

Dr. Xubo Liu

Invited Talk on IEEE SPS Early-Career Seminar, 27 May, 2026

Talk Outline

Enabling interactive and seamless human-machine interaction:

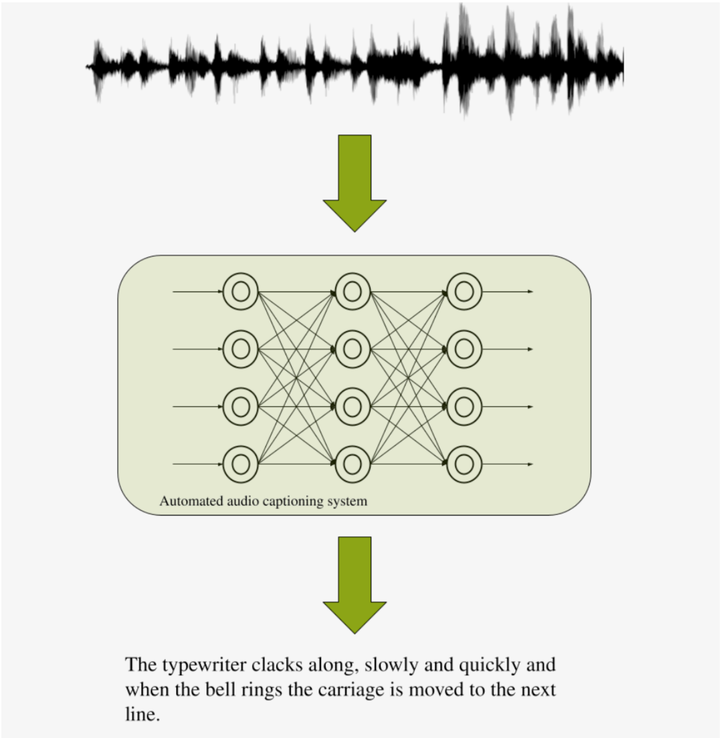
- Natural Language as a Seamless Interface
- Natural Language Intelligence-Powered **Audio Technologies**

- **Understanding**: Audio Captioning, QA & Reasoning, Speech Recognition
- **Editing & Generation**: Language-Queried Audio Source Separation, Text-to-Audio Generation/Storytelling
- **Representation Learning**: Audio Coding / Tokenisation

Audio Understanding

Audio Captioning

Generating a natural language description given an audio clip



X Mei, X Liu, MD Plumbley, W Wang, “Automated Audio Captioning: An Overview of Recent Progress and New Challenges”, EURASIP Journal on Audio, Speech, and Music Processing, 2022

Audio Captioning

An Example: CNN-Transformer Encoder-Decoder

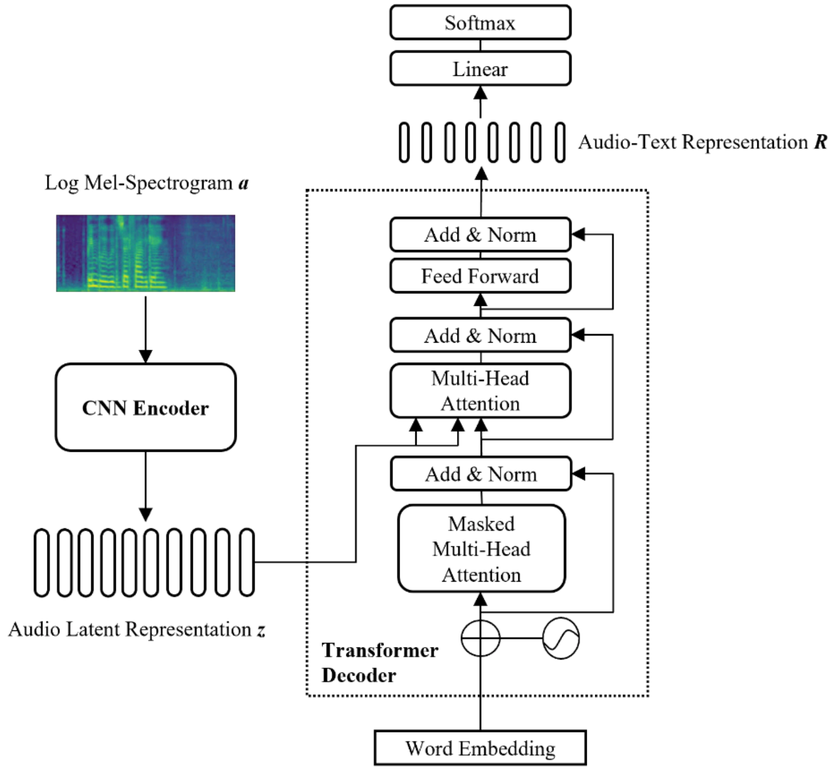


Figure 1: Sequence-to-sequence architecture with CNN encoder and Transformer decoder for audio captioning. The components in the dashed box indicate the Transformer decoder.

Research Focus:

- Transfer learning, representation learning, architecture [1-5]
- Interaction with vision modality [6-7]
- Diversity of audio captions [8-9]

[1] X Mei, Q Huang, X Liu, G Chen, J Wu, et al., “An Encoder-Decoder Based Audio Captioning System With Transfer and Reinforcement Learning”, Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE), 2021

[2] X Mei, X Liu, Q Huang, MD Plumbley, W Wang, “Audio Captioning Transformer”, Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE), 2021

[3] J Sun, X Liu, X Mei, V Kılıç, MD Plumbley, W Wang, “Dual Transformer Decoder based Features Fusion Network for Automated Audio Captioning”, INTERSPEECH, 2023

[4] X Liu, Q Huang, X Mei, T Ko, HL Tang, MD Plumbley, W Wang, “CL4AC: A Contrastive Loss for Audio Captioning”, Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE), 2021

[5] X Liu, X Mei, Q Huang, J Sun, J Zhao, H Liu, MD Plumbley, et al., “Leveraging Pre-trained BERT for Audio Captioning”, European Signal Processing Conference (EUSIPCO), 2022

[6] X Liu, Q Huang, X Mei, H Liu, Q Kong, J Sun, et al., “Visually-Aware Audio Captioning with Adaptive Audio-Visual Attention”, INTERSPEECH, 2023

[7] Ö Çaylı, X Liu, V Kılıç, W Wang, Knowledge Distillation for Efficient Audio-Visual Video Captioning”, European Signal Processing Conference (EUSIPCO), 2023

[8] X Mei, X Liu, J Sun, MD Plumbley, W Wang, “Diverse Audio Captioning via Adversarial Training”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022

[9] X Mei, X Liu, J Sun, MD Plumbley, W Wang, Towards Generating Diverse Audio Captions via Adversarial Training”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024

Audio Question Answering (QA) & Reasoning

Acoustic Prompt Tuning (APT):

An **Efficient** adapter extending LLMs/VLMs to the audio domain using an improved soft-prompting approach

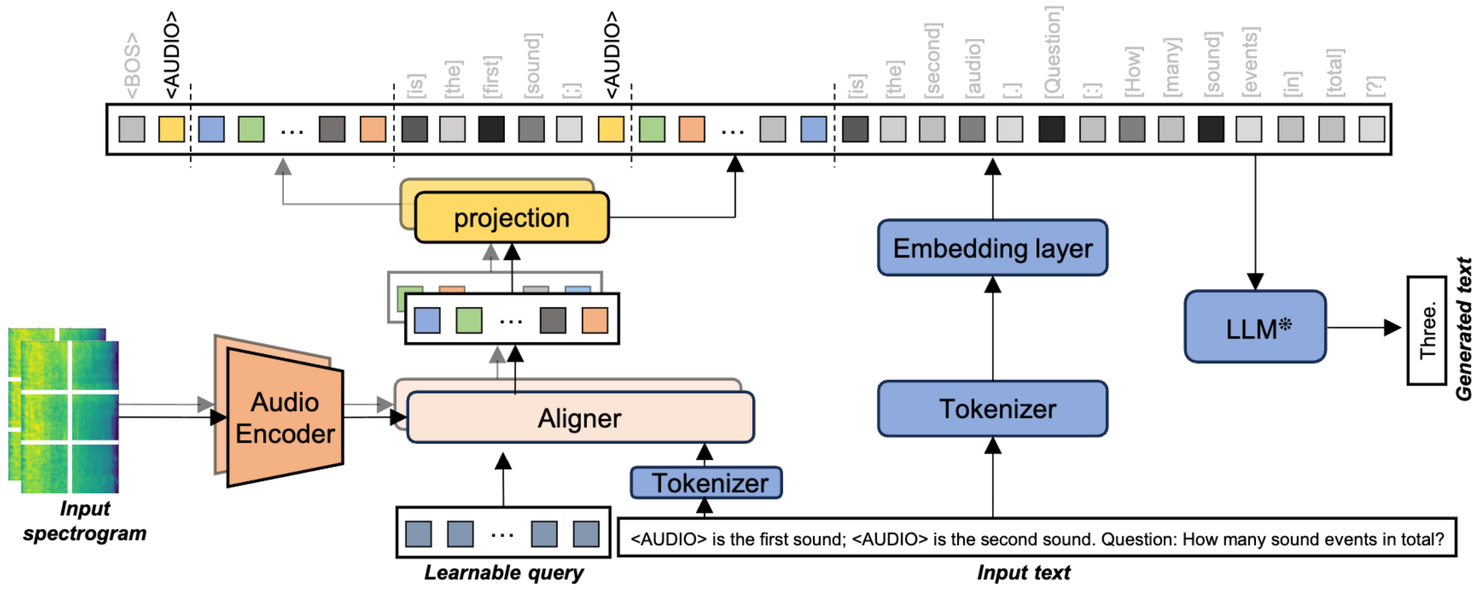
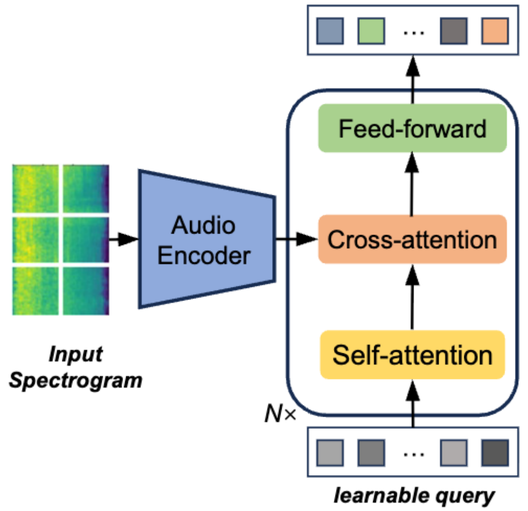


Fig. 1. Illustration of the proposed APT-LLM. APT-LLM includes three components: an audio encoder, an audio aligner, and an LLM. The audio encoder extracts audio feature maps from input spectrograms. The audio aligner then projects each audio feature map to 32 acoustic embeddings according to the input text. These acoustic embeddings, together with the added embeddings of the audio token "[AUDIO]", are juxtaposed with text embeddings. The interleaved audio-text embeddings are fed into the LLM to generate the output text. APT-LLM can ingest multiple audio clips in a sequence and thus benefit from diverse tasks during training.



Structure of audio aligner in APT.

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

Audio Question Answering (QA) & Reasoning

Multi-Task Training

TABLE XIV
TEMPLATE OF AUDIO TAGGING, AUTOMATED AUDIO CAPTIONING, LANGUAGE-QUERIED SOUND EVENT DETECTION, TEMPORAL EVENT RETRIEVAL, AND SOUND EVENT COUNTING.

Audio Tagging	Audio Captioning	Language-Queried Sound Event Detection	Temporal Event Retrieval	Sound Event Counting
Summarize the audio with key words.	Summarize the audio succinctly.	Pinpoint the presence of {LABEL} with the time stamps.	Summarize the audio with key words in the interval of {STT} seconds to {EDT} seconds.	How many times can the sound {LABEL} be heard?
What sound events can be heard in the audio clip?	Present a short overview of the provided audio samples.	Indicate the start and end time of the audio event {LABEL}.	What sound events can be heard from {STT} seconds to {EDT} seconds?	How many instances of the sound {LABEL} can be perceived?
What auditory incidents can be recognized in the recording?	Provide a compact summary of the auditory content.	Document the exact times the sound {LABEL} taking place.	What auditory incidents can be recognized in the recording from {STT} seconds to {EDT} seconds?	What is the number of times the sound {LABEL} is detectable?
Which auditory occurrences can be detected?	Offer a brief outline of the audio clips that have been given.	Specify the time stamps for {LABEL} occurrence.	Which auditory occurrences can be detected during {STT} seconds to {EDT} seconds?	How frequently can one hear the sound {LABEL}?
Which sound occurrences can be perceived?	Render a compressed version of the audio's main points.	When the sound {LABEL} happens?	Which sound occurrences can be perceived between {STT} seconds and {EDT} seconds?	How often can the sound {LABEL} be perceived?
Present a concise breakdown of the given audio clips.	Describe the audio clip concisely.	Capture the exact times when {LABEL} is happening.	Present a concise breakdown of the recording from {STT} seconds to {EDT} seconds.	
List the sound events in the audio clip.	Explain the audio clip in a brief and straightforward manner.	Describe the time intervals during which {LABEL} takes place.	List the sound events in the interval of {STT} seconds to {EDT} seconds.	
Describe the recording with names of sound events.	Write a terse but informative summary of the sound.	State the precise moment at which {LABEL} occurs.	Name the auditory incidents within the {STT} to {EDT} seconds timeframe.	
Enumerate the audio events present in the audio.	Give a quick overview of the provided audio excerpts.	What time does the sound event {LABEL} take place?	Enumerate the audio events present between {STT} seconds and {EDT} seconds.	
Name the auditory incidents in the audio sample.	Outline the given audio samples briefly.	Capture the beginning and end time of the sound {LABEL}.	Describe the recording with names of sound events within the {STT} to {EDT} seconds timeframe.	
#Output: {LABEL}	#Output: {CAPTION}	#Output: {STT}s-{EDT}s	#Output: {LABEL}	#Output: {NUMBER}

TABLE V
PERFORMANCE (%) COMPARISON IN AUTOMATED AUDIO CAPTIONING TASKS. ↑ INDICATES THE HIGHER NUMBER, THE BETTER PERFORMANCE.

Model	AudioCaps		Clotho		Weighted average	
	SPICE↑	SPIDEr↑	SPICE↑	SPIDEr↑	SPICE↑	SPIDEr↑
Specialised systems trained with task-specific examples						
AT+CNN10 [41]	16.8	-	11.5	-	15.1	-
CNN-GPT2 [42]	16.7	43.8	11.1	21.5	14.9	36.7
WSAC+PD [43]	17.3	40.3	12.3	24.7	15.7	35.3
WavCaps [32]	18.2	48.5	13.3	29.7	16.6	**42.5**
One-for-all models for various audio tasks						
APT-LLM	19.1	40.2	13.2	24.8	**17.2**	35.3

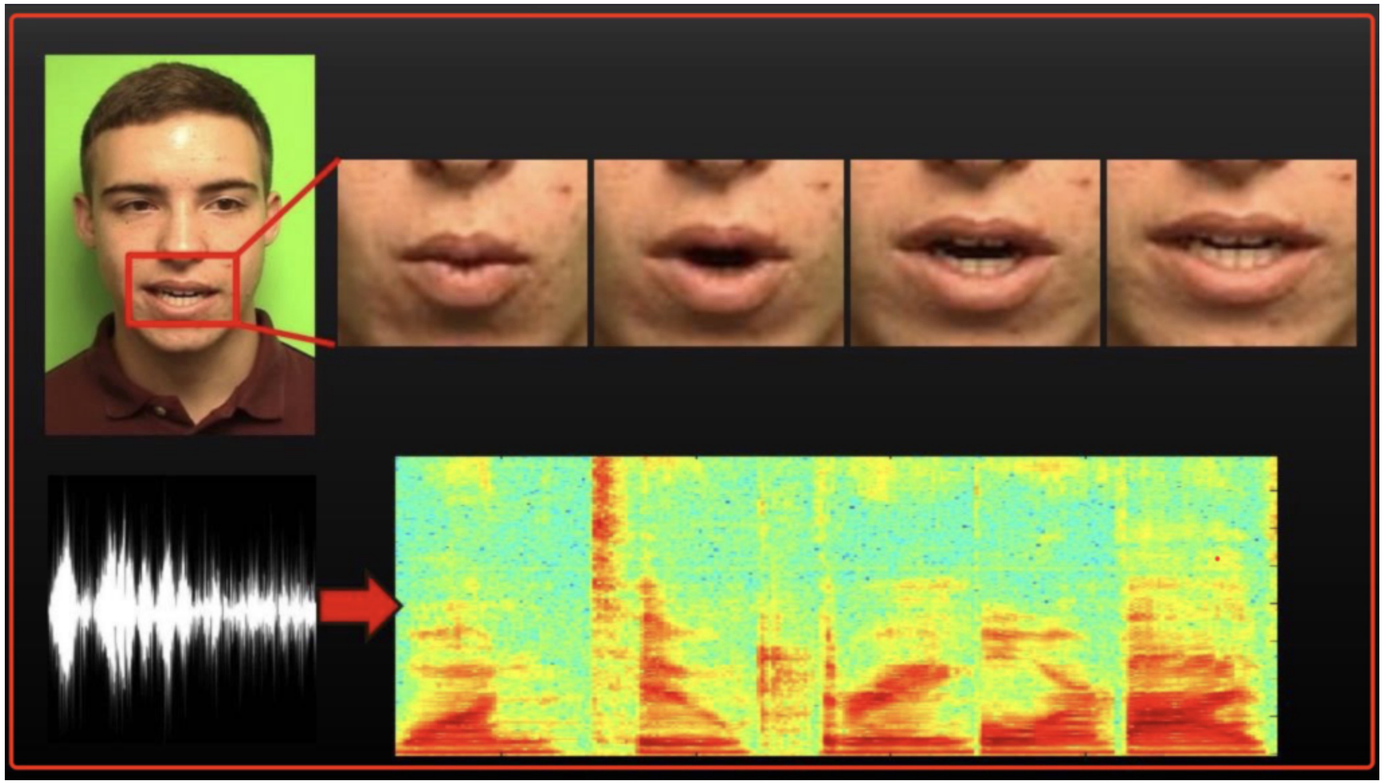
TABLE VI
ACCURACY (%) OF VARIOUS METHODS ON ESC-50 IN THE FEW-SHOT SETTINGS.

	Accuracy↑	
	5-way	12-way
Specialised systems trained with task-specific examples		
ProtoNet [49]	88.2	77.7
MatchNet [50]	86.8	71.8
HPN [51]	88.7	78.7
Audio language models trained with contrastive learning		
TIP-adapter [52]	97.5	95.6
Treff adapter [14]	98.5	96.3
One-for-all models for various audio tasks		
APT-LLM	91.0	54.2

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

Multimodal Speech Recognition

Visual Speech Recognition (VSR) - Recognise speech using lip movements.



Applications:

- Improve Speech Recognition in noisy environments
- Help the hearing impaired

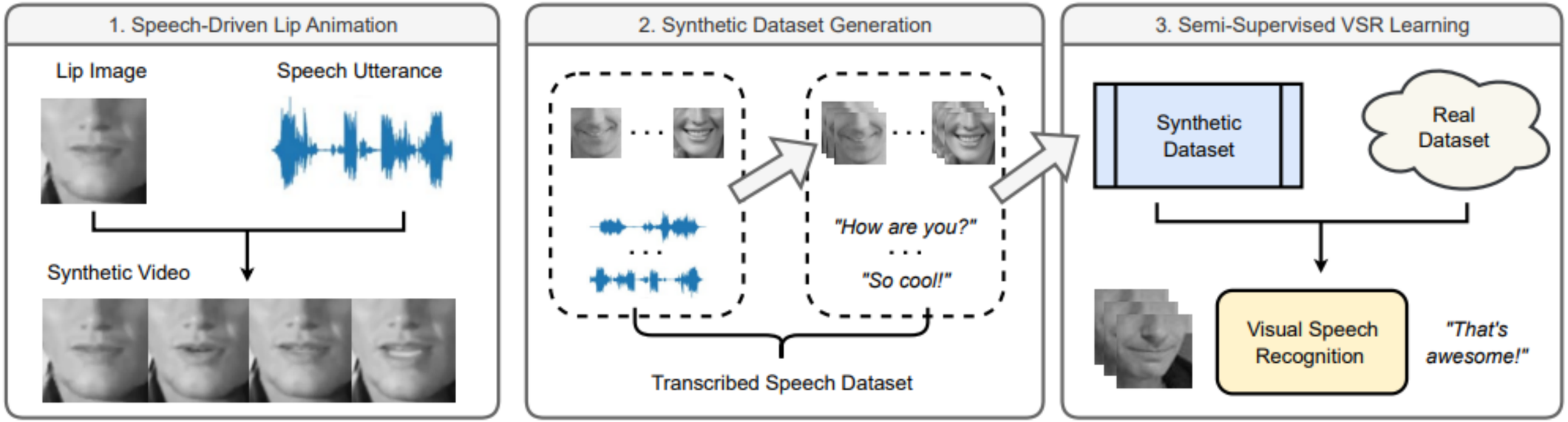
Key Challenges:

- Lack of large-scale labelled audio-visual data (e.g., LRS3 438 hours)
- Bias, license and privacy concerns using real human data

Multimodal Speech Recognition

SynthVSR: Scaling VSR with Synthetic Supervision

- A semi-supervised framework with large-scale synthetic supervision (3600+ hours)



X Liu, E Lakomkin, K Vougioukas, P Ma, H Chen, R Xie, M Doulaty, Ni Moritz, J Kolar, S Petridis, M Pantic, C Fuegen, “SynthVSR: Scaling Up Visual Speech Recognition with Synthetic Supervision”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023

Multimodal Speech Recognition

SynthVSR: Scaling VSR with Synthetic Supervision

- Generate synthetic lip movement videos from speech and face datasets; GAN-based approach
- SoTA VSR WER **16.9%** is achieved on LRS3, using **29x** less data than previous Google's system.

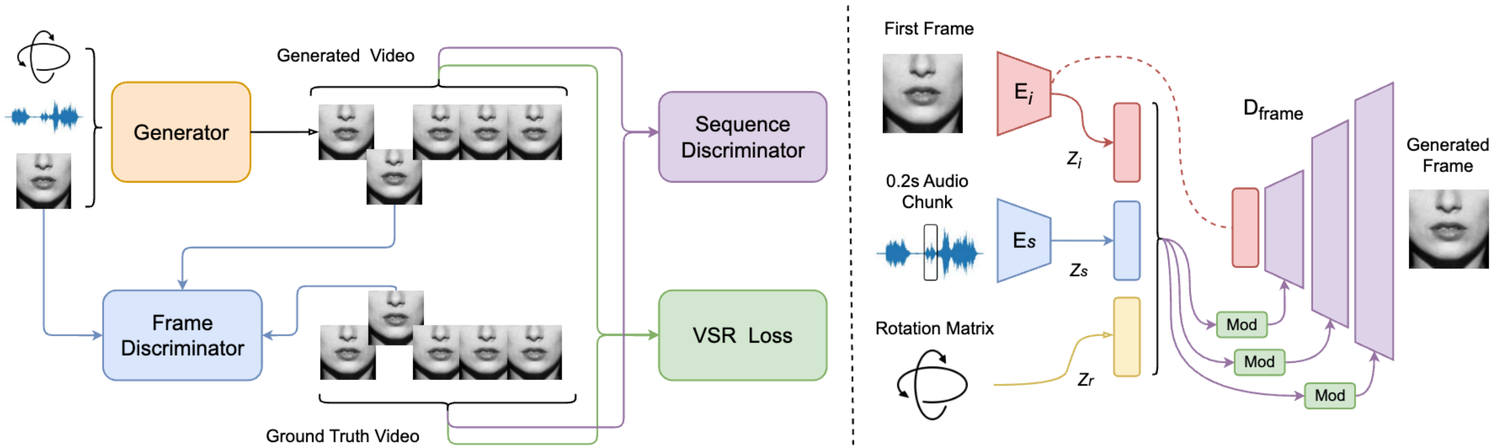


Figure 2. **Architecture of proposed speech-driven lip animation model.** Left: GAN-based speech-driven lip animation model generating lip movements given a lip image, a speech utterance, and a rotation sequence; Right: structure of the generator in the lip animation model.

X Liu, E Lakomkin, K Vougioukas, P Ma, H Chen, R Xie, M Doulaty, Ni Moritz, J Kolar, S Petridis, M Pantic, C Fuegen, "SynthVSR: Scaling Up Visual Speech Recognition with Synthetic Supervision", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023

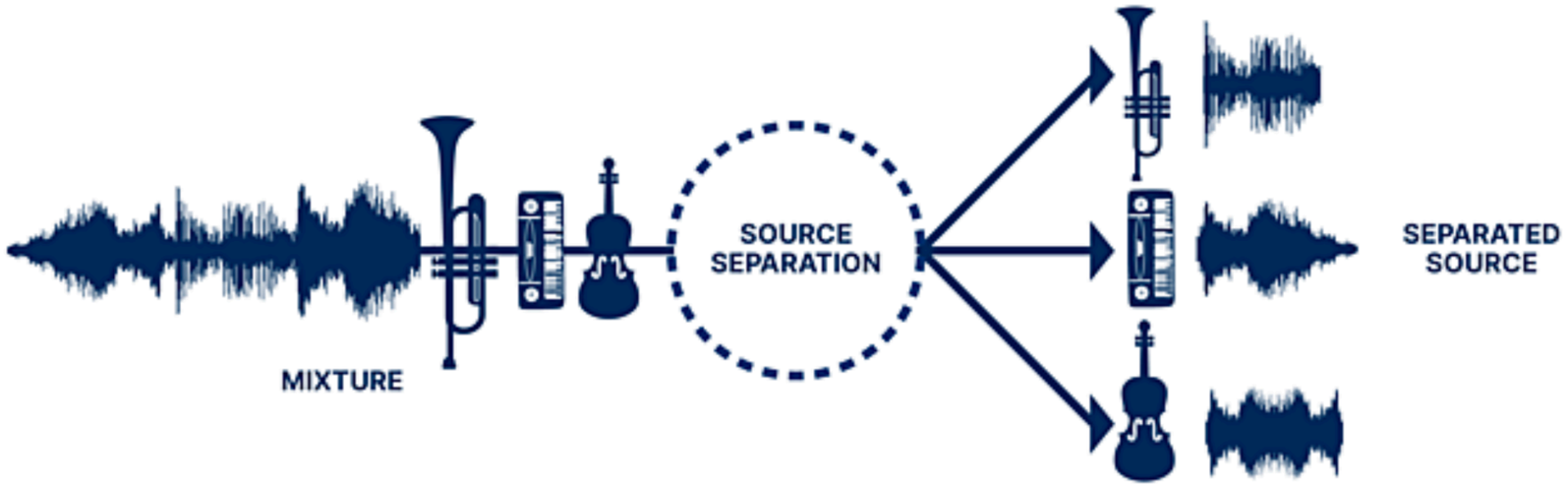
Audio Editing & Generation

Language-Queried Audio Source Separation

Audio Source Separation - a fundamental technology for audio editing

Query-based audio source separation (separate the sounds of interests)

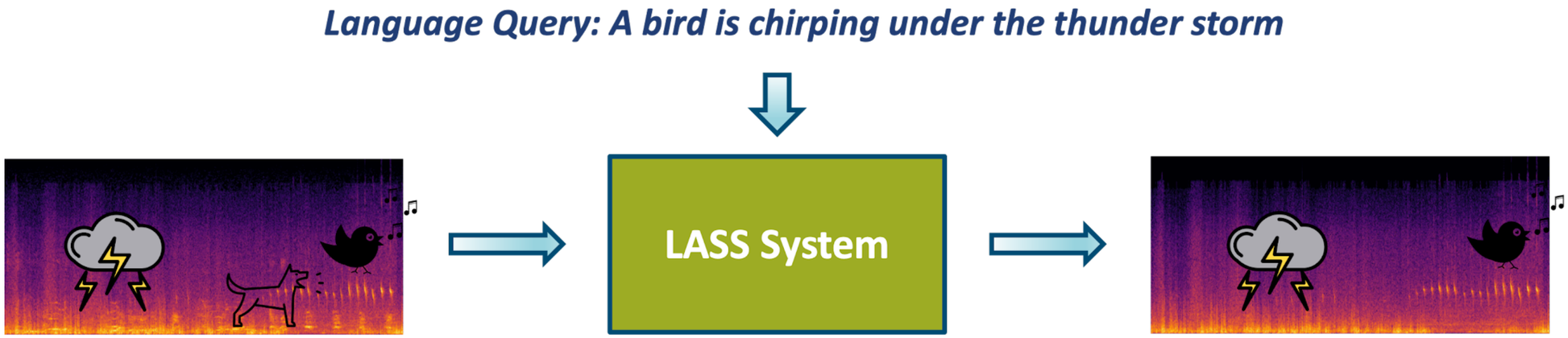
- Queries: vision, audio, labels
- Not flexible and straightforward to separate the desired sounds



Language-Queried Audio Source Separation

Audio Source Separation - a fundamental technology for audio editing

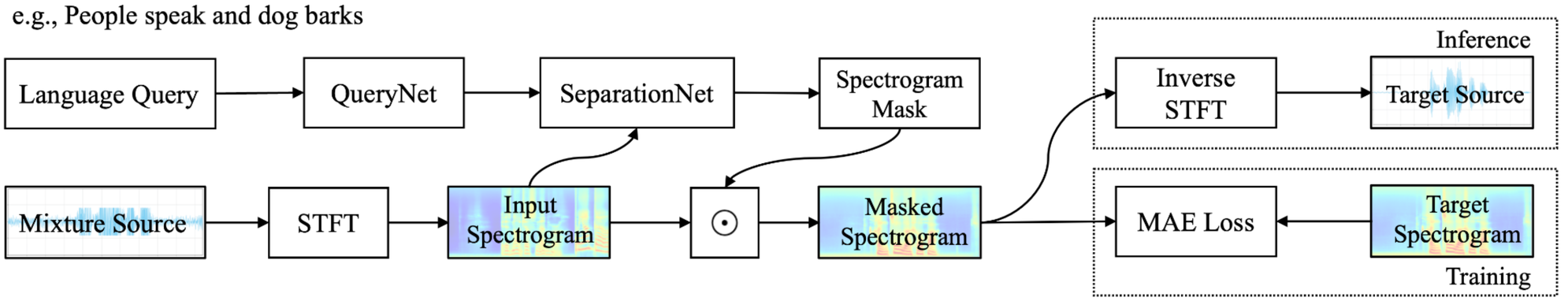
- LASS - Separate a **target source** from an audio mixture based on the **natural language descriptions** of the target source
- First attempt bridging source separation and natural language processing



X Liu, H Liu, Q Kong, X Mei, J Zhao, Q Huang, MD Plumbley, W Wang, “Separate What You Describe: Language-Queried Audio Source Separation”, INTERSPEECH 2022

Language-Queried Audio Source Separation

- LASS-Net:
  - QueryNet (BERT) + SeparationNet (ResUNet) + FiLM fusion
- How to construct training data?
  - Training with synthetic mixtures generating from audio-text datasets



X Liu, H Liu, Q Kong, X Mei, J Zhao, Q Huang, MD Plumbley, W Wang, “Separate What You Describe: Language-Queried Audio Source Separation”, INTERSPEECH 2022

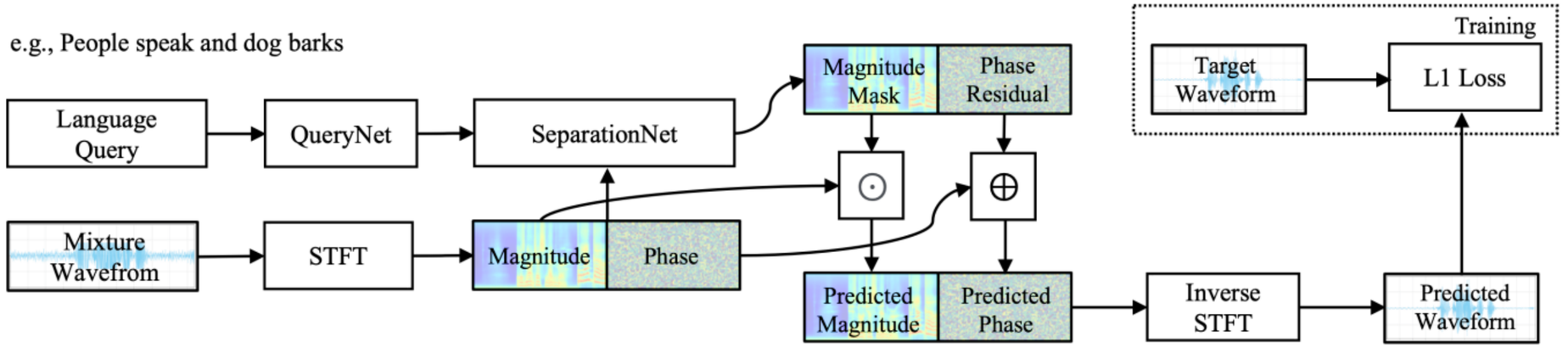
Language-Queried Audio Source Separation

AudioSep - Separate Anything You Describe

- CLAP/CLIP + ResUNet, scaling 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

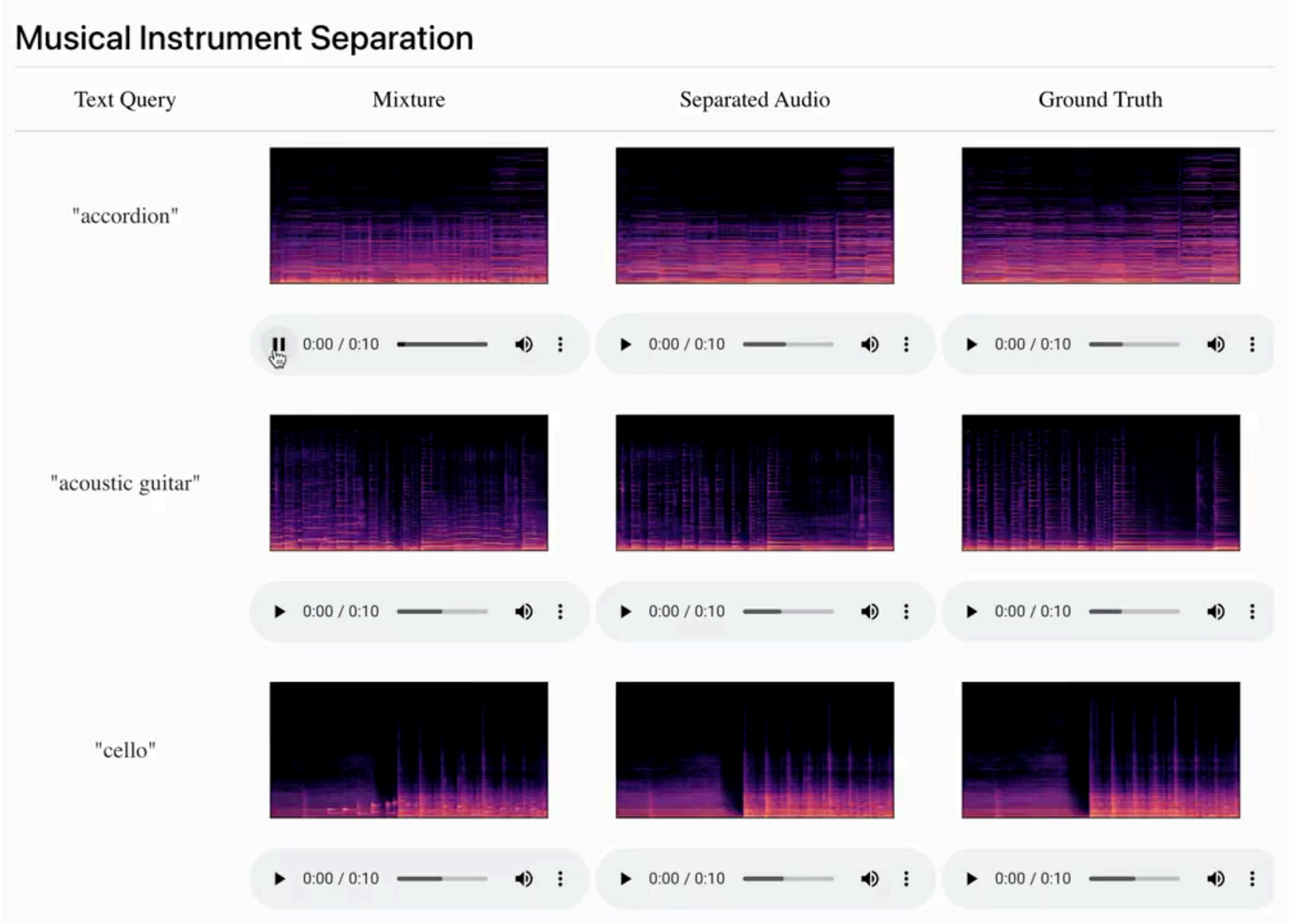
TABLE I
AUDIOSEP TRAINING DATASETS.

	Caption	Label	Video	Num. clips	Hours
AudioSet	×	✓	✓	2 063 839	5800
VGGSound	×	✓	✓	183 727	550
AudioCaps	✓	✓	✓	49 768	145
Clotho v2	✓	×	×	4884	37
WavCaps	✓	×	×	403 050	7568



X Liu, Q Kong, Y Zhao, H Liu, Y Yuan, Y Liu, R Xia, Y Wang, MD Plumbley, W Wang, “Separate Anything You Describe”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024

Language-Queried Audio Source Separation

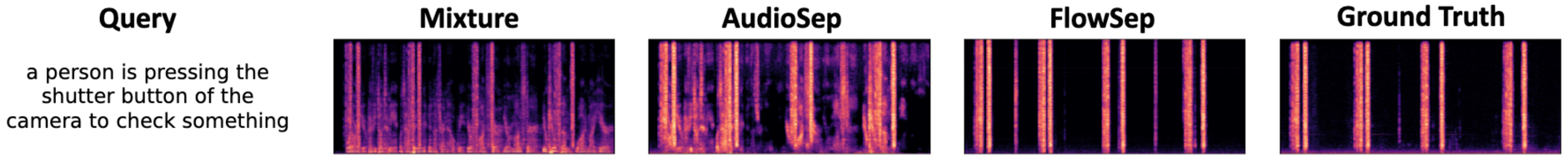


X Liu, Q Kong, Y Zhao, H Liu, Y Yuan, Y Liu, R Xia, Y Wang, MD Plumbley, W Wang, “Separate Anything You Describe”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024

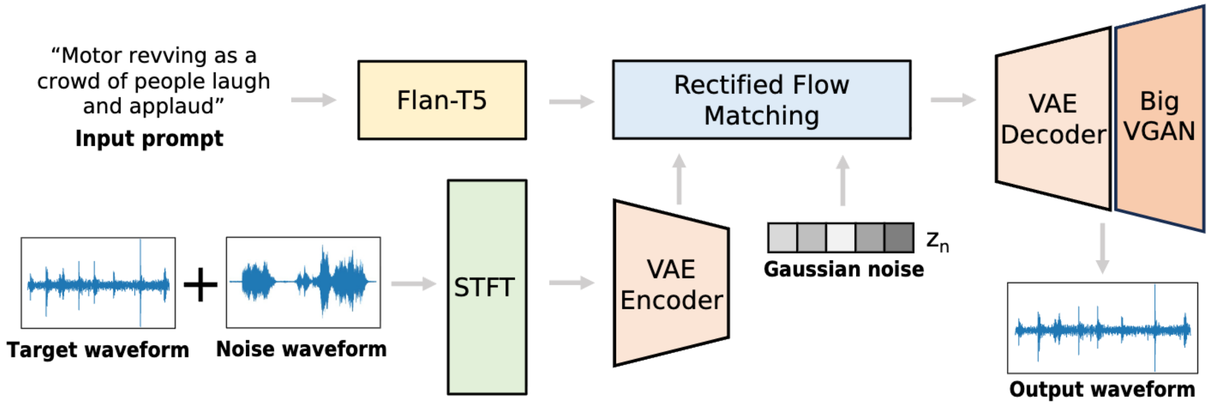
Language-Queried Audio Source Separation

FlowSep - LASS with Rectified Flow Matching

- Masking-based discriminative methods often leads to excessive or incomplete separation



- Using Diffusion-based generative approach for high-quality audio source separation



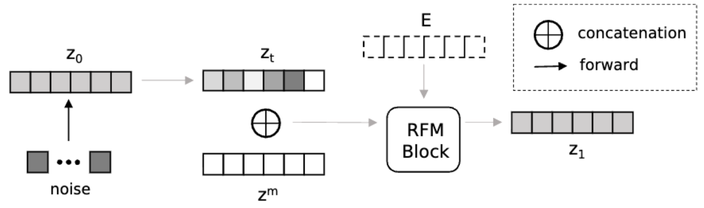


Fig. 2. The channel-concatenation conditioning mechanism

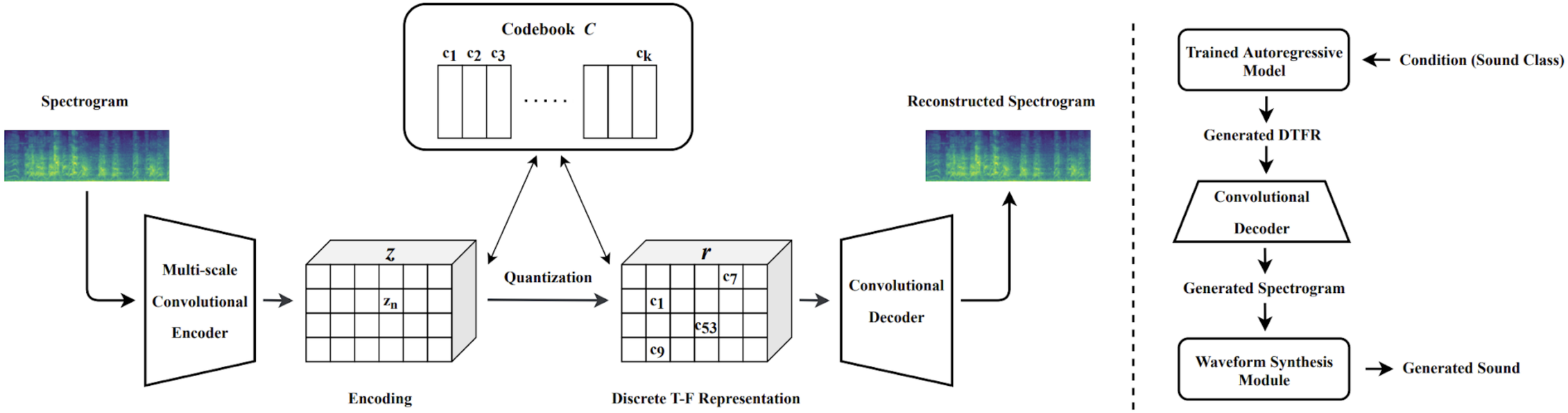
Fig. 1. The architecture of FlowSep. FlowSep consists of four main components: (1) a FLAN-T5 encoder for text embedding; (2) a VAE for encoding and decoding mel-spectrograms; (3) an RFM module for generating audio features within the VAE latent space; (4) a BigVGAN vocoder to generate the waveform.

Y Yuan, X Liu, H Liu, MD Plumbley, W Wang, "FlowSep: Language-Queried Sound Separation with Rectified Flow Matching", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025

Audio Generation

Language Modeling-based General Audio Synthesis

- Treat the audio generation as a **language modelling** task in the **Time-Frequency (TF) domain**

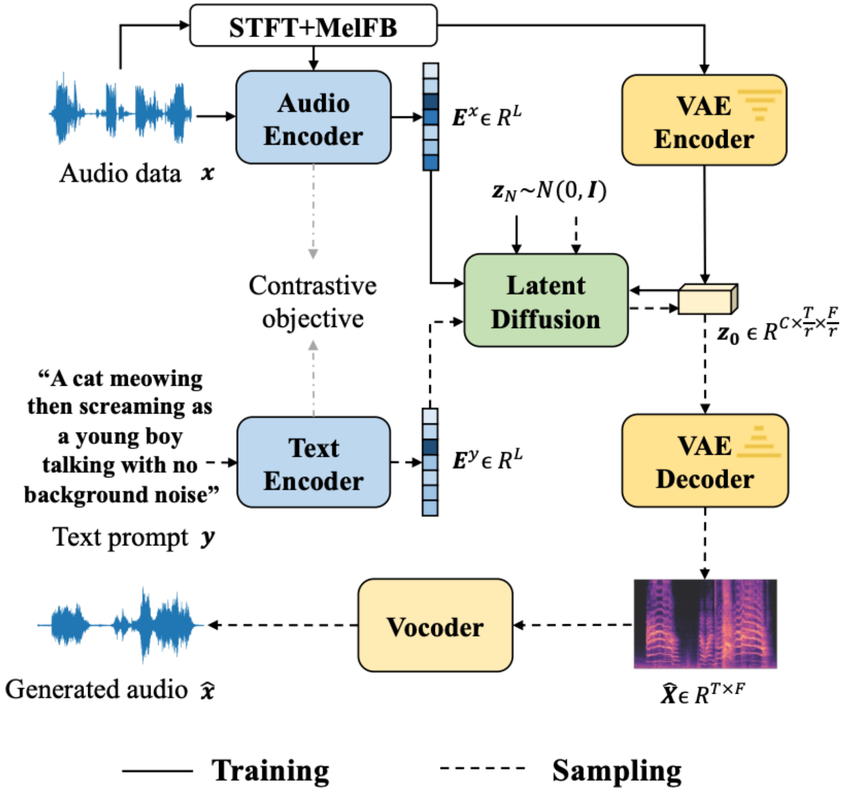


X Liu, T Iqbal, J Zhao, Q Huang, MD Plumbley, W Wang, “Conditional Sound Generation using Neural Discrete Time-Frequency Representation Learning”. In 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.

Text-to-Audio Generation

AudioLDM: Text-to-Audio Generation with Latent Diffusion Models (LDMs)

- **Key Components:**
  - Mel-spectrogram VAE
  - Contrastive Language-Audio Pretraining (CLAP) Encoders
  - **Latent Diffusion Models**
  - Mel-to-Waveform Vocoder

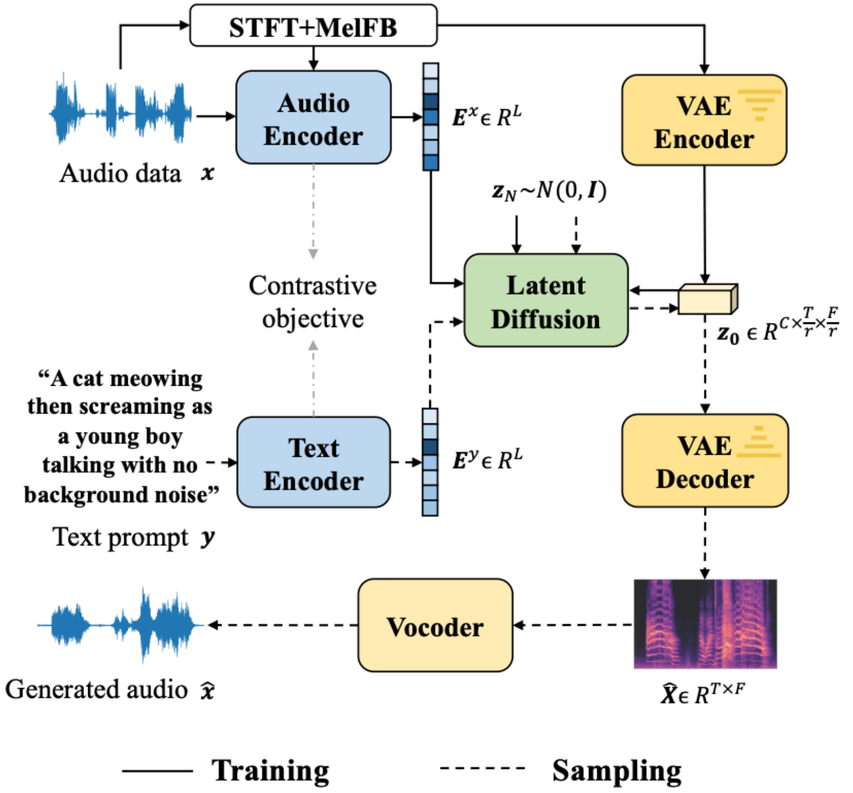


H Liu, Z Chen, Y Yuan, X Mei, X Liu, D Mandic, W Wang, MD Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models". In Proceedings of the International Conference on Machine Learning (ICML), 2023

Text-to-Audio Generation

AudioLDM: Text-to-Audio Generation with Latent Diffusion Models (LDMs)

- Self-supervised text-to-audio generation
- Data/resource-efficient approach (e.g., training with single GPU)
- Enable zero-shot tasks (e.g., inpainting, style transferring, super-resolution)



Project Page: https://audioldm.github.io/

GitHub: https://github.com/haoheliu/AudioLDM

H Liu, Z Chen, Y Yuan, X Mei, X Liu, D Mandic, W Wang, MD Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models". In Proceedings of the International Conference on Machine Learning (ICML), 2023

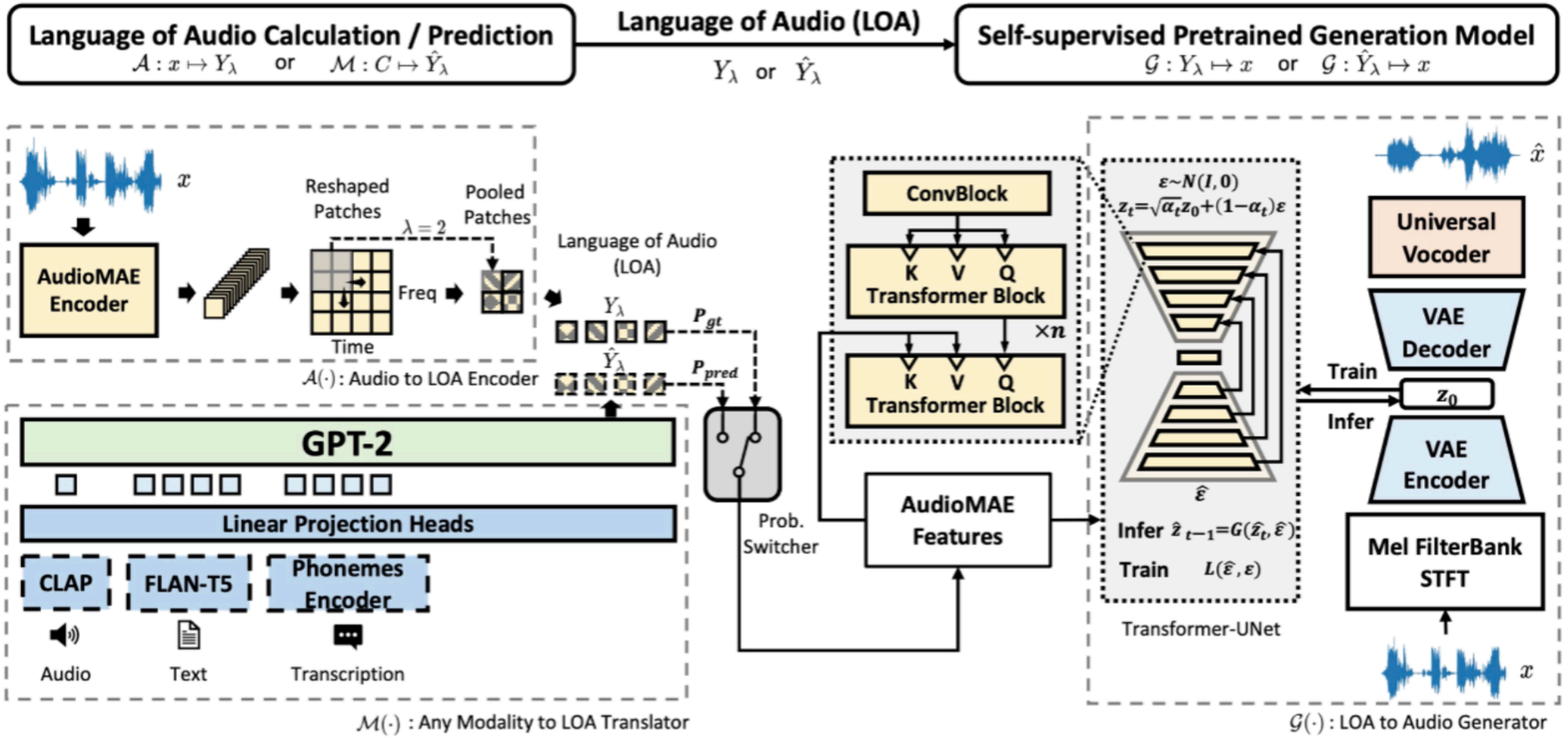
Text-to-Audio Generation

AudioLDM: Text-to-Audio Generation with Latent Diffusion Models (LDMs)

H Liu, Z Chen, Y Yuan, X Mei, X Liu, D Mandic, W Wang, MD Plumbley, “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models”. In Proceedings of the International Conference on Machine Learning (ICML), 2023

Text-to-Audio Generation

AudioLDM2: Combining LLMs with Diffusion



H Liu, Y Yuan, X Liu, X Mei, Q Kong, Q Tian, Y Wang, W Wang, Y Wang, MD Plumbley, "AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024

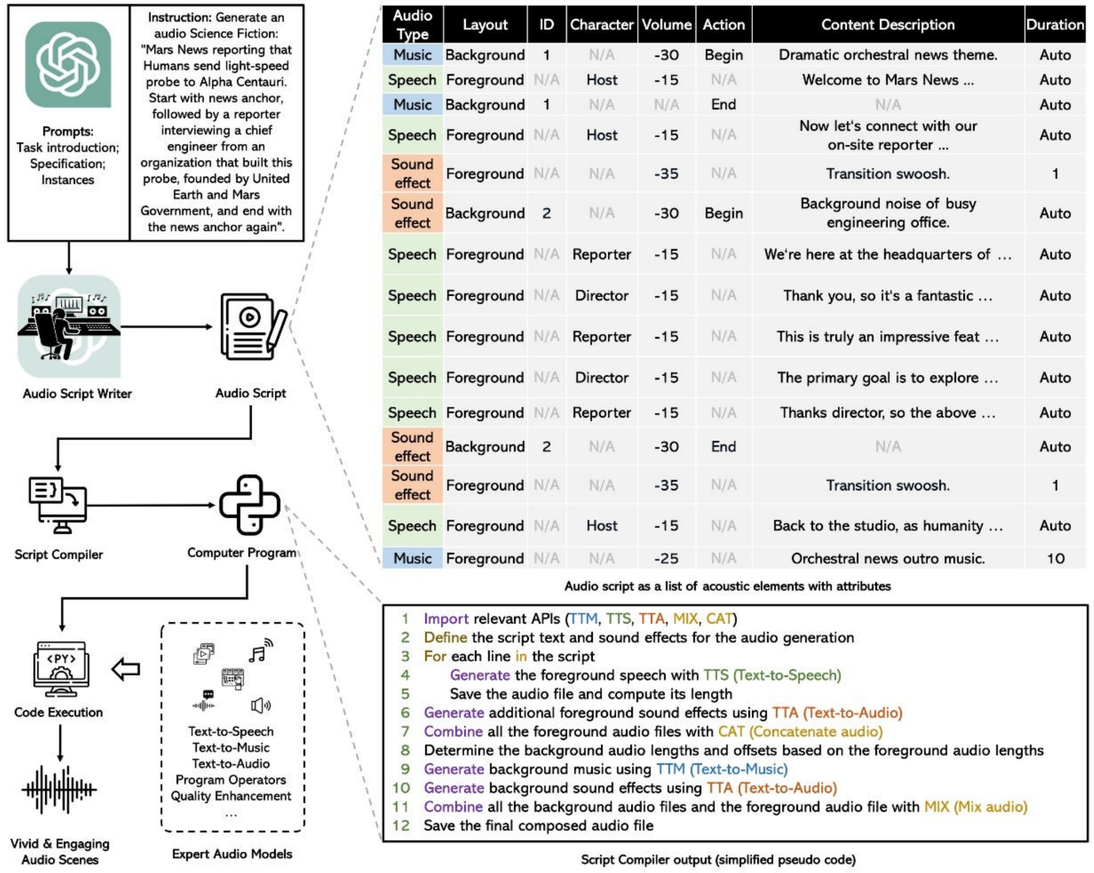
Audio Storytelling

WavJourney: Compositional Audio Content Creation with LLMs

Create audio storytelling with:

- Personalized speakers
- Lifelike speech
- Immersive music
- Impactful sound effects

<u>All simply from texts!</u>



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)

H Liu, Y Yuan, X Liu, X Mei, Q Kong, Q Tian, Y Wang, W Wang, Y Wang, MD Plumbley, "AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024

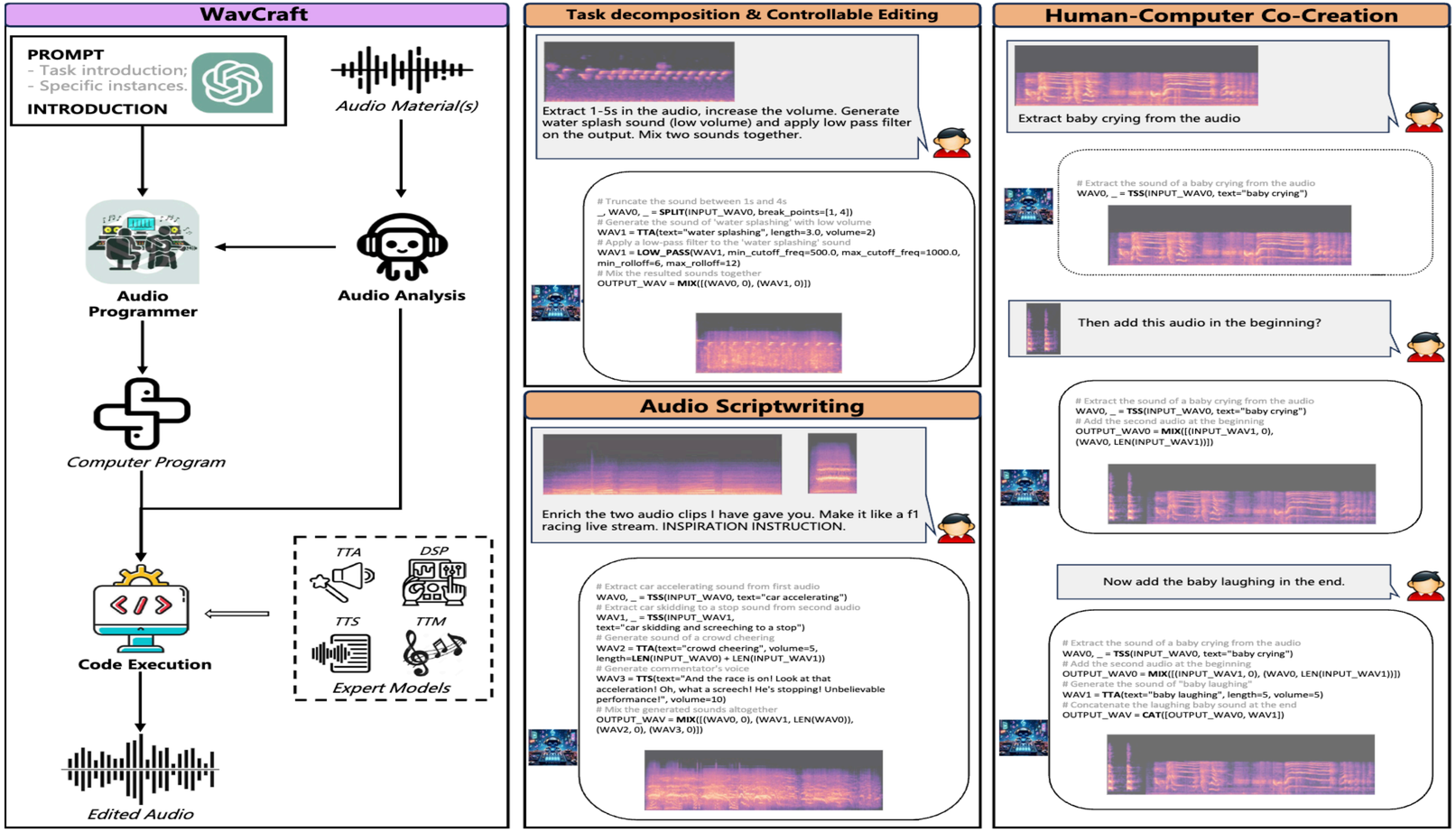
Audio Storytelling

WavJourney: Compositional Audio Content Creation with LLMs

X Liu, Z Zhu, H Liu, Y Yuan, M Cui, Q Huang, J Liang, Y Cao, Q Kong, MD Plumbley, W Wang "WavJourney: Compositional Audio Creation with Large Language Models". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2025

Audio Editing

WavCraft: Agentic Audio Editing



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

Representation Learning (Audio Tokenisation)

Audio Tokenisation

Source Disentanglement Audio Codec (SD-Codec)

Combining audio coding and audio source separation in latent space

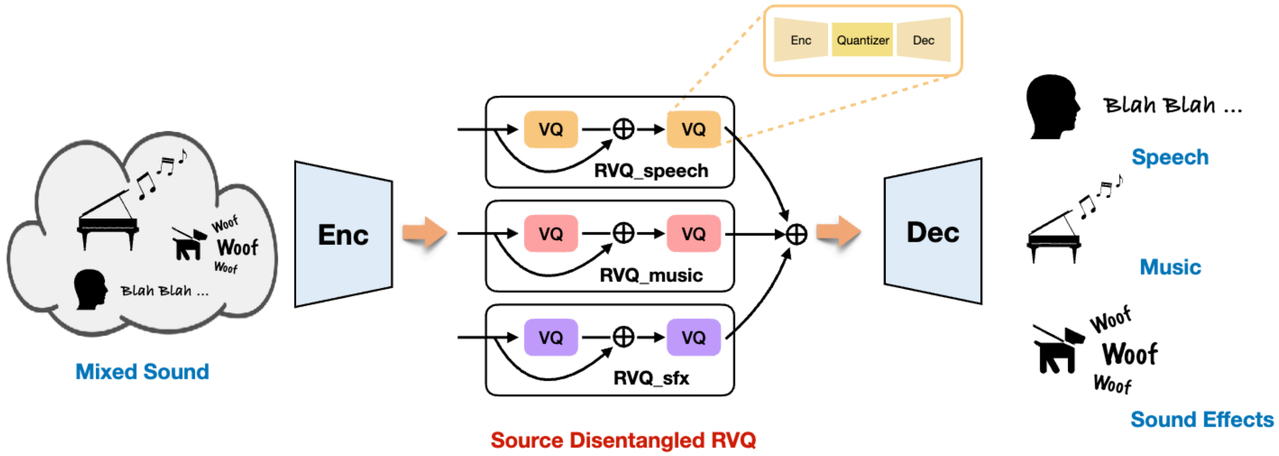


Fig. 1. Source Dissentangled Neural Audio Codec (SD-Codec)

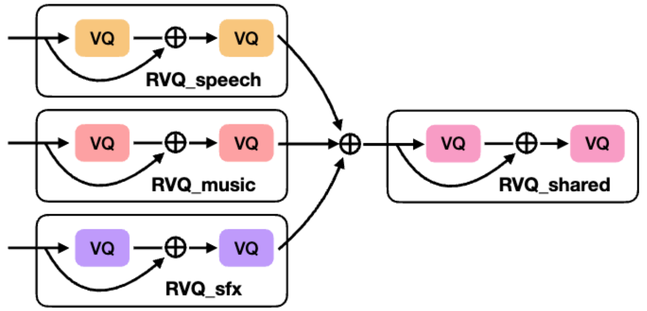


Fig. 2. SD-Codec with shared codebooks ($R = 4$, $S = 2$)

X Bie, X Liu, G Richard, "Learning Source Disentanglement in Neural Audio Codec". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025

Audio Tokenisation

Scaling up **transformer-based speech tokenizers** (e.g., 1B) with an FSQ bottleneck can achieve state-of-the-art speech reconstruction quality at ultra-low bitrates, as low as 400–700 bps

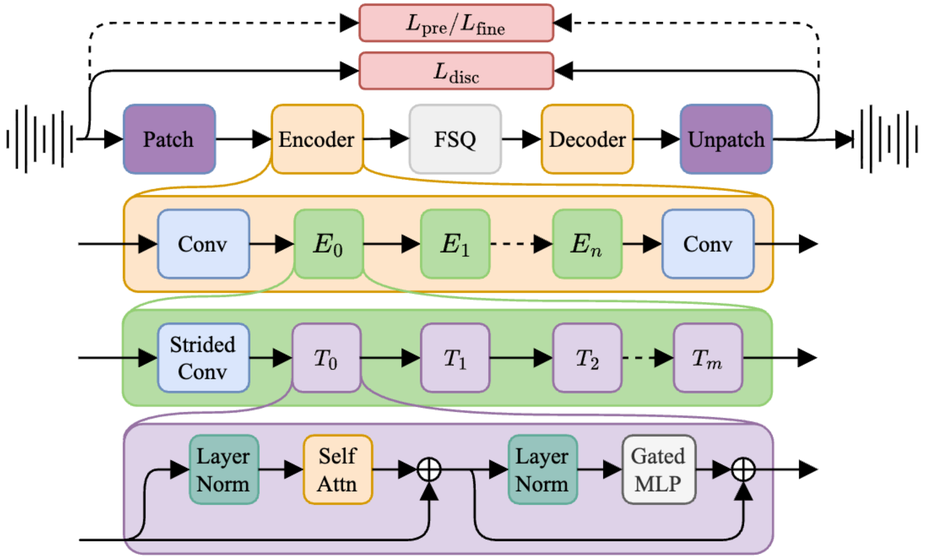


Figure 1: Architecture of the proposed model. Detail is shown for the encoder block and sub-blocks. The decoder block is configured identically to the encoder block, with the exception of the strided convolution, which is replaced with its transposed equivalent and moved to the end of the T_m blocks.

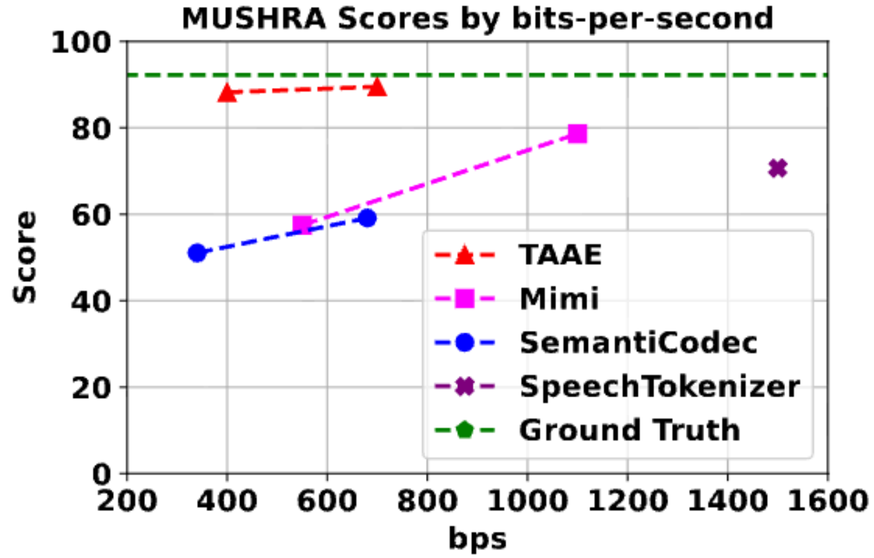
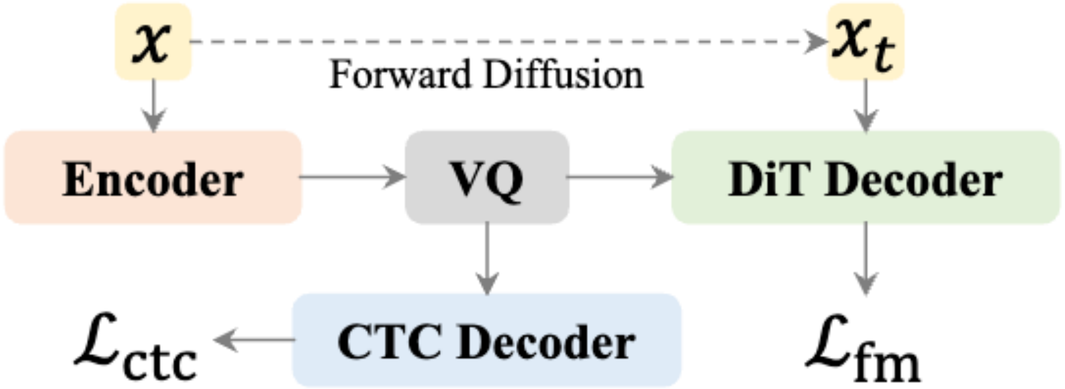


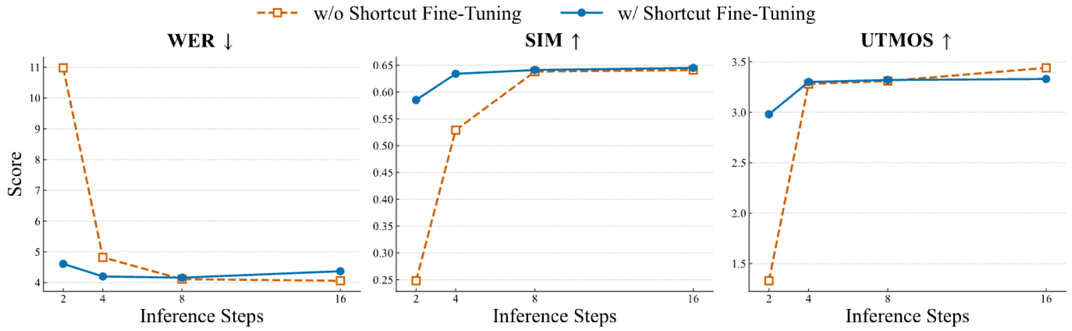
Figure 2: Results of MUSHRA test.

JD Parker, A Smirnov, J Pons, CJ Carr, Z Zukowski, Z Evans, X Liu, "Scaling Transformers for Low-Bitrate High-Quality Speech Coding". International Conference on Learning Representations (ICLR 2025)

Audio Tokenisation

Speech Diffusion Tokeniser (SiTok) - a diffusion autoencoder that jointly learns semantic-rich representations through supervised learning and enables high-fidelity audio reconstruction with diffusion. We scale SiTok to 1.6B parameters and train it on 2 million hours of speech.





$$\mathcal{L}_{\text{total}} = \underbrace{\mathbb{E}_{t,\boldsymbol{x},\boldsymbol{\epsilon}}\left[\|\mathcal{D}_\phi(\boldsymbol{x}_t, t, \boldsymbol{z}_q) - (\boldsymbol{x} - \boldsymbol{\epsilon})\|\right]}_{\text{Reconstruction Loss}} + \lambda_{\text{ctc}} \underbrace{\text{CTC}(\mathcal{D}_{\phi_{\text{ctc}}}(\boldsymbol{z}_q), \boldsymbol{y})}_{\text{CTC Loss}} + \underbrace{\mathcal{L}_{\text{vq}}}_{\text{VQ Loss}}$$

Observation: Shortcut fine-tuning enables efficient low-step inference, retaining high intelligibility and similarity while substantially accelerating decoding.

Y Wang, Z Tang, Y Wang, A Hinsvark, Y Liu, Y Li, K Peng, J Ao, M Ma, M Seltzer, Q He, X Liu, "Scaling Speech Tokenizers with Diffusion Autoencoders". International Conference on Learning Representations (ICLR 2026)

Acknowledgement

- Haohe Liu
- Xinhao Mei
- Jianyuan Sun
- Qiushi Huang
- Yi Yuan
- Jinhua Liang
- Egor Lakomkin
- Christian Fuegen
- Dino Vougioukas
- Yun Wang
- Xiaoyu Bie
- Anton Smirnov
- Yuancheng Wang
- Yin Cao
- Qiuqiang Kong
- Julian Parker
- Jordi Pons
- Turab Iqbal
- Zhongkai Zhu
- Gael Richard
- Emmanouil Benetos
- Mark D. Plumbley
- Wenwu Wang
- ...

Doctoral Research -> Industrial Research

Doctoral Research -> Industrial AI Focus

Academia Focus:

- Novelty
- Clarity
- Scientific Contribution
- Benchmark Performance
- Publication

Create new knowledge

Industrial Focus: create new knowledge

- Impact
- Scalability
- Reliability
- Latency / Cost
- Integration

Turn Knowledge into scalable capabilities

Focus on impact!

Doctoral Research -> Industrial AI Focus

Academia skills:

- Problem formulation
- Literature review
- Experimental rigor
- Communication
- …

Other skill requirements for Industrial AI

- System thinking
- Product and user
- Engineering
- Collaboration
- …

Thank you!