

Diffusion Models in Audio Processing

Dr. Xubo Liu

Invited Talk at University of Cambridge, 14 October, 2026

Short Bio

Dr. Xubo Liu

- Research Scientist, Meta Superintelligence Labs, London (25 - now)
- Head of Speech, Stability AI, London (24)
- PhD in Vision, Speech and Signal Processing, University of Surrey (20-24)
- Research interests: machine listening, multimodal learning, generative AI, human-machine interaction
- Personal website: <https://liuxubo717.github.io/>; Email: liuxubo717@gmail.com

Diffusion Models

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

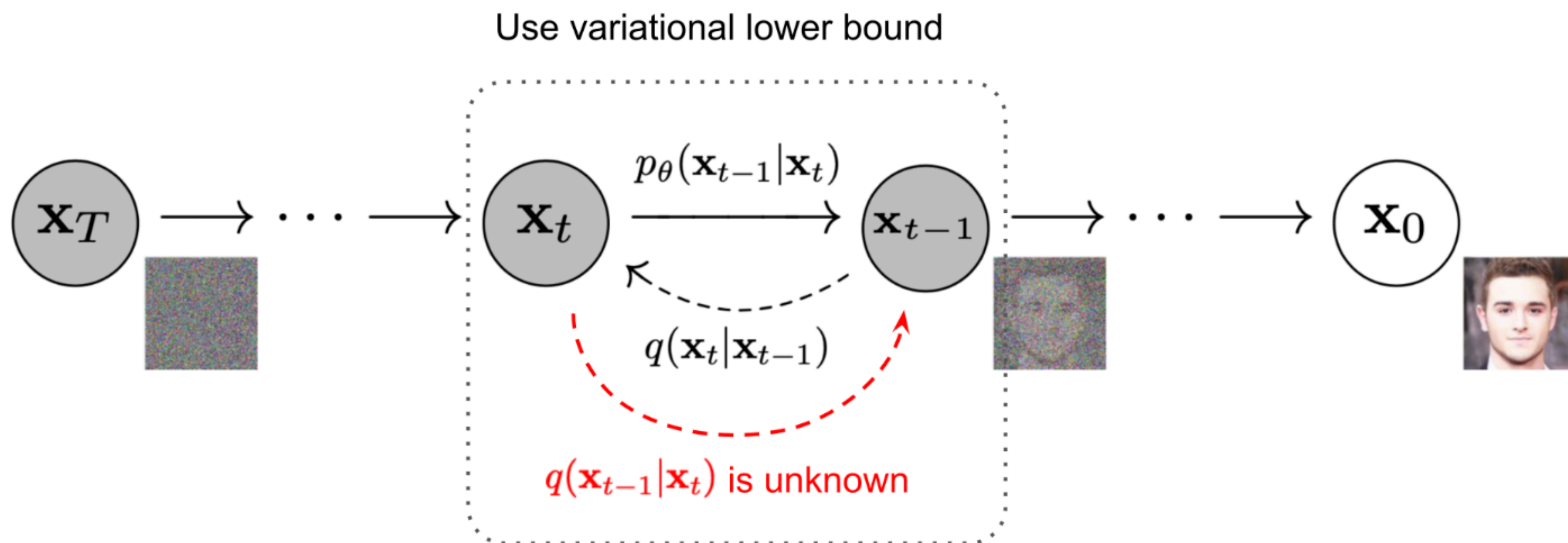
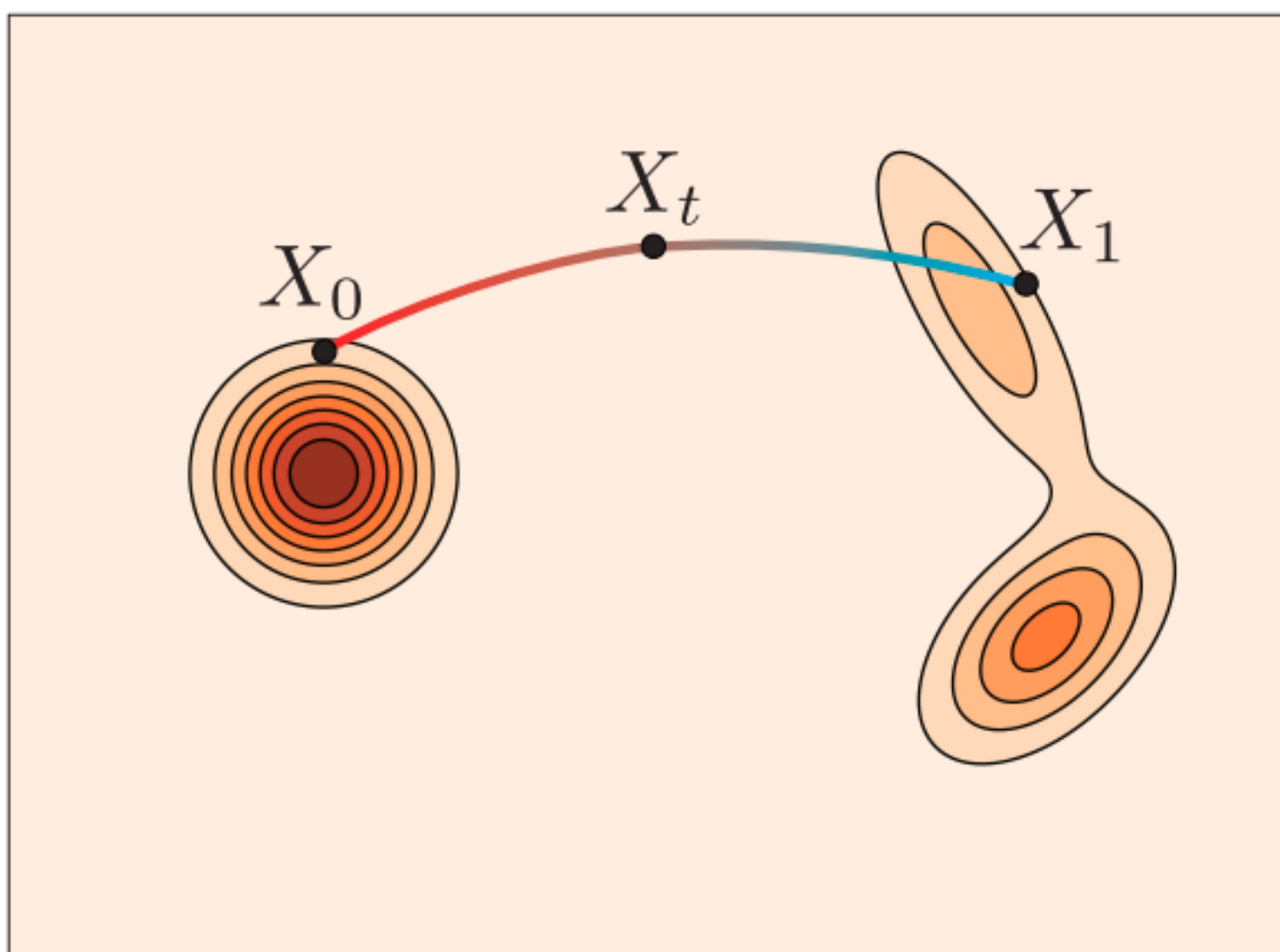


Figure 2: The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: [Ho et al. 2020](#) with a few additional annotations)

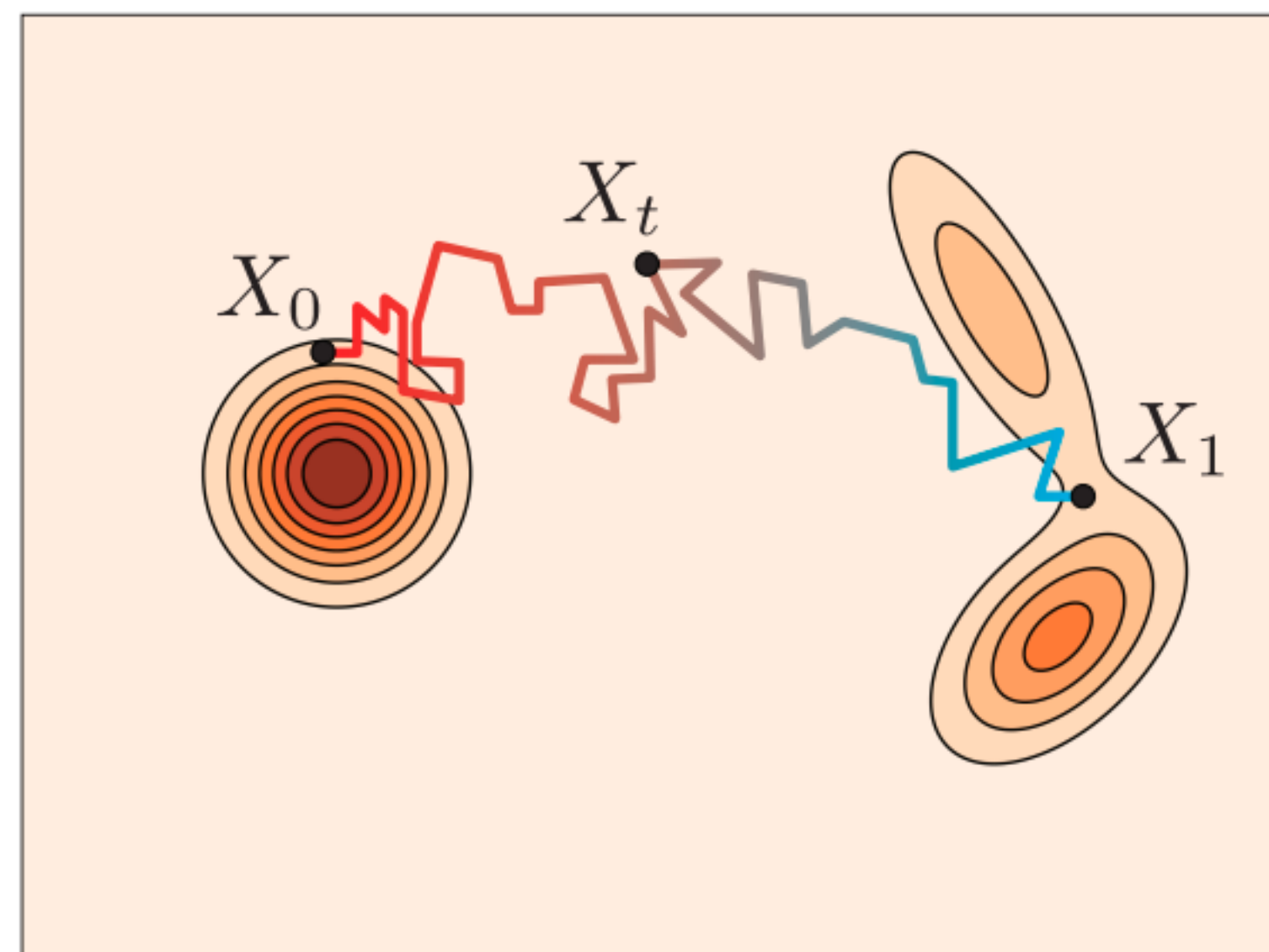
Flow Matching

$$x_t = (1 - t)x_0 + tx_1$$

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, x_0, x_1} \left[\left\| v_\theta(x_t, t) - \frac{dx_t}{dt} \right\|^2 \right]$$



(a) Flow



(b) Diffusion

Talk Outline

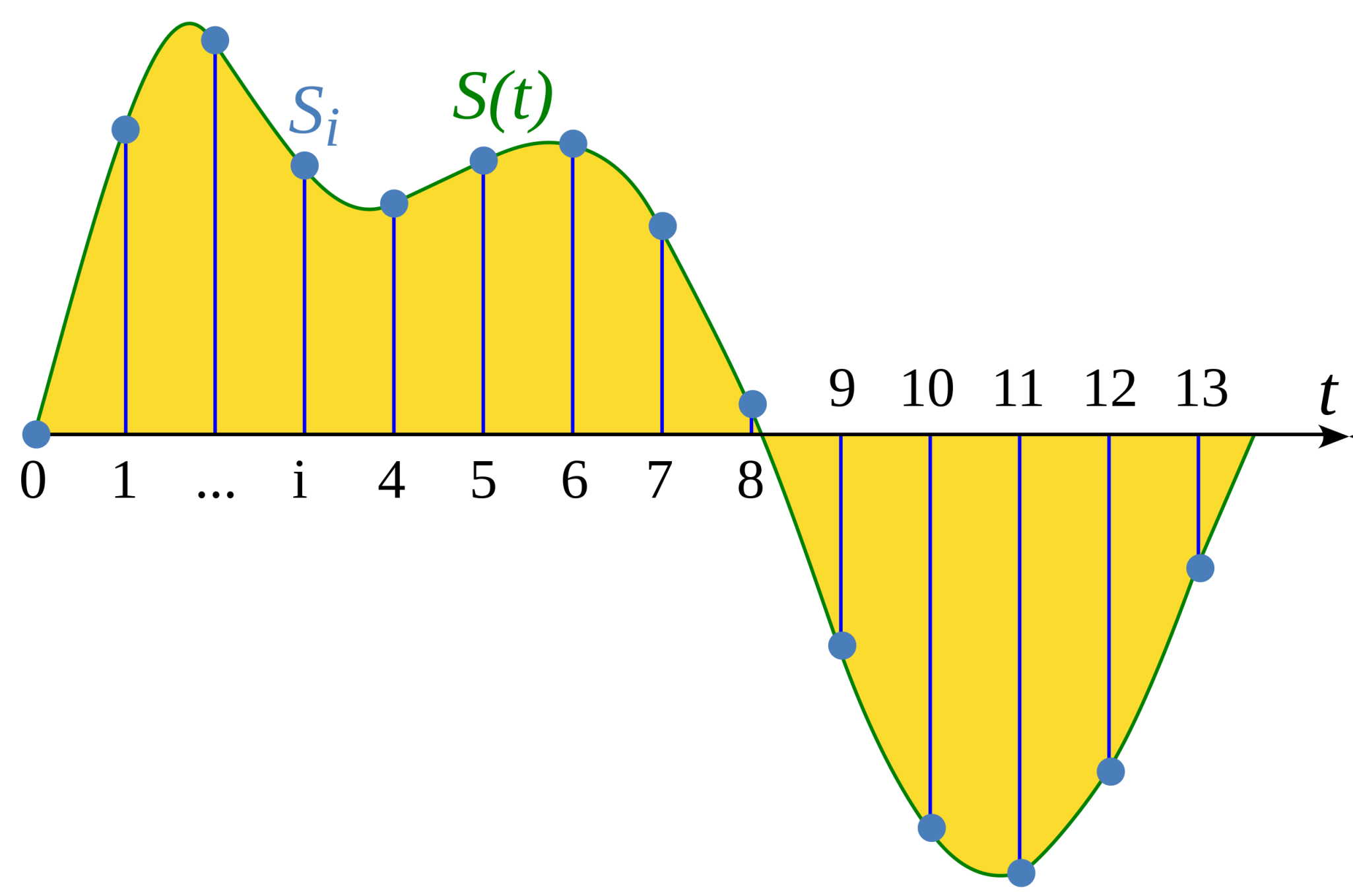
Diffusion Models for Audio Processing

- Audio Representations
- Audio Generation:
 - e.g., Text-to-Audio Generation
- Source Separation:
 - e.g., Language-Queried Audio Source Separation
- Representation Learning:
 - e.g., Audio Tokenisers

Audio Representations

Audio Representations

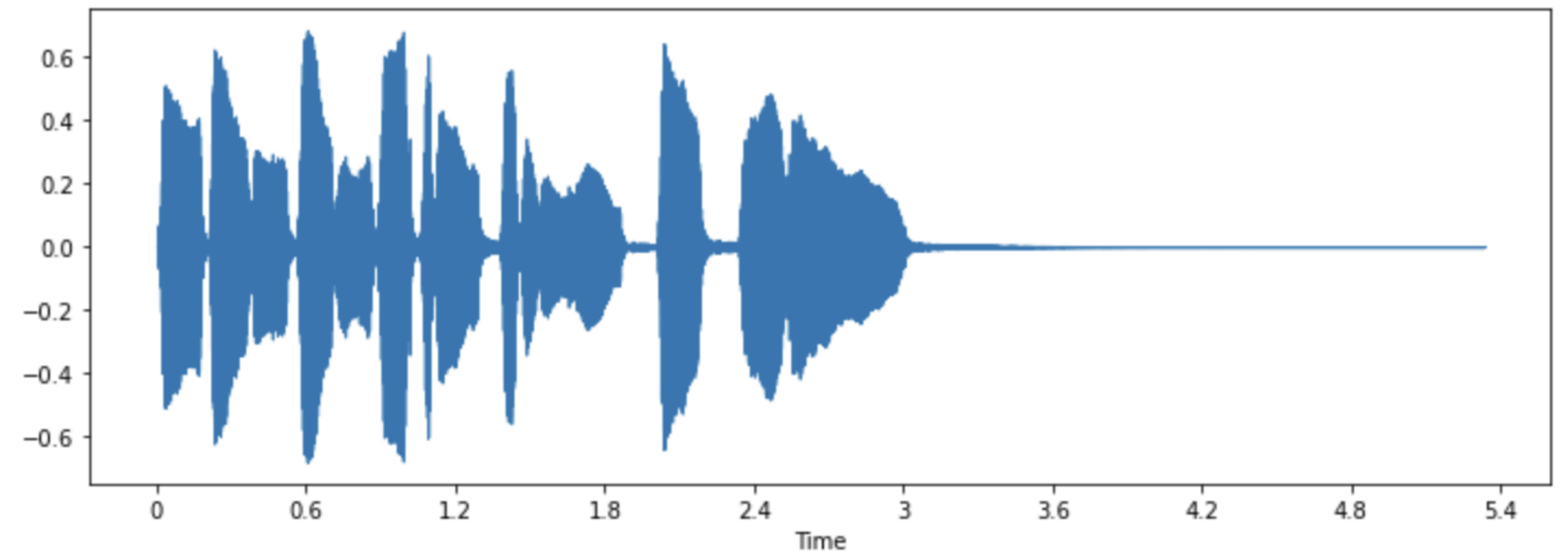
Audio as a waveform



Sampling rate: the number of samples taken in one second and is measured in hertz (Hz)

Waveform (1-Dimensional)

- The direct plot of “vibration”.
- ☒ Original format
- ☐ Hard to read
- ☐ High-dimensional



Waveform (1-D): Time-domain representation of audio

Audio Representations

Audio as a Spectrogram

Short-Time Fourier Transform

$$X(m, k) = \sum_{n=0}^{N-1} x[n + mH]w[n]e^{-j2\pi kn/N}$$

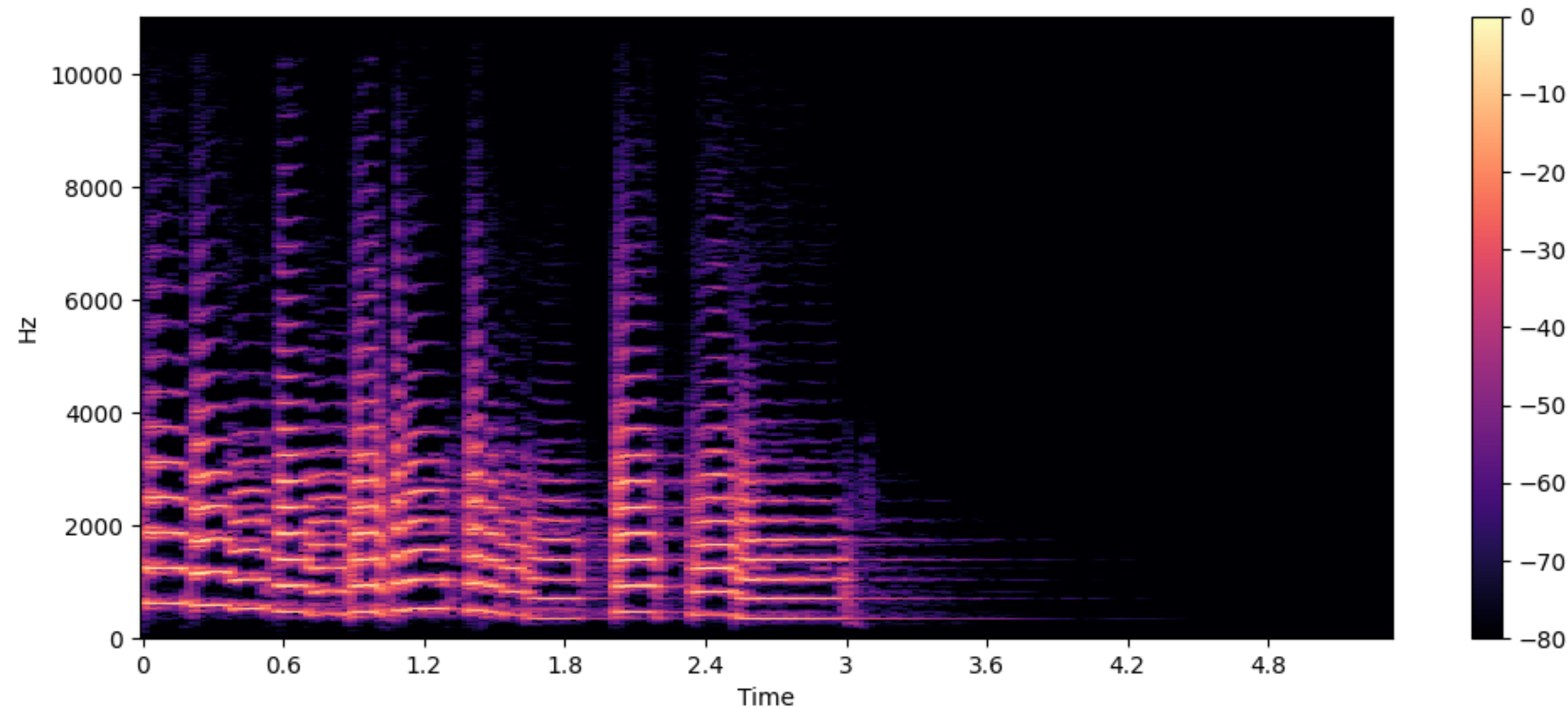
$x[n]$: audio waveform

$w[n]$: analysis window

N : FFT size

H : hop size

m, k : time-frame and frequency-bin indices



The **Short-Time Fourier Transform (STFT)** represents audio in the time–frequency domain. The complex STFT retains both magnitude and phase, allowing waveform reconstruction through **inverse STFT (iSTFT)**.

Audio Representations

Audio as a Mel-Spectrogram

Mel Filterbank Projection

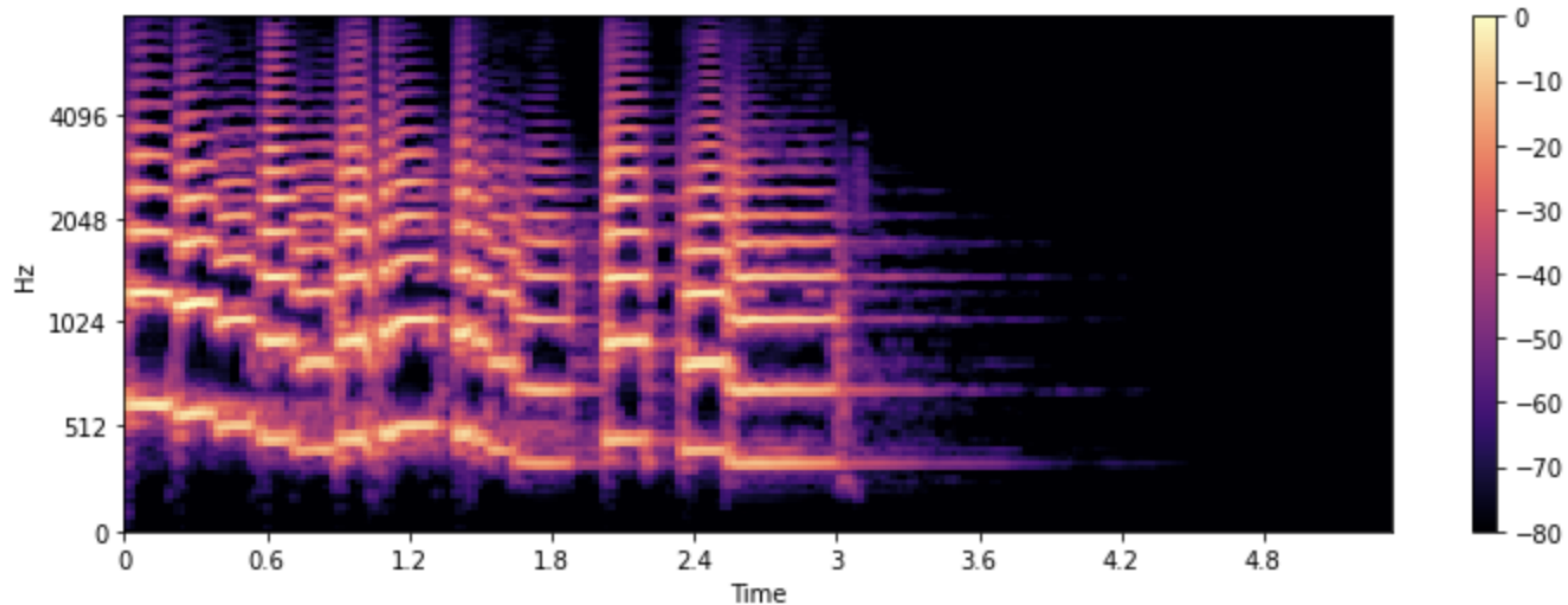
$$M(m, b) = \sum_{k=0}^{N/2} F_b(k) |X(m, k)|^2$$

$X(m, k)$: complex STFT

$F_b(k)$: mel filter weights

$M(m, b)$: mel-band energy

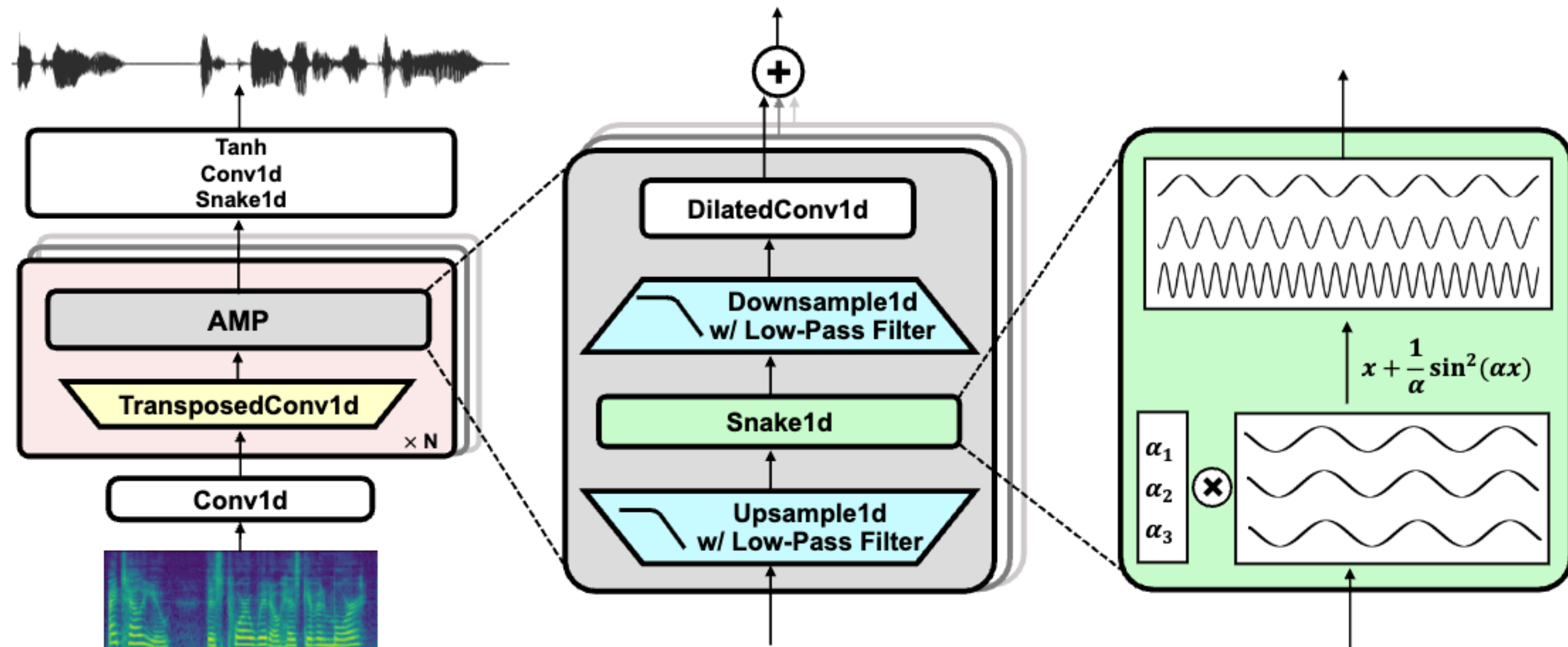
b : mel-frequency bin index



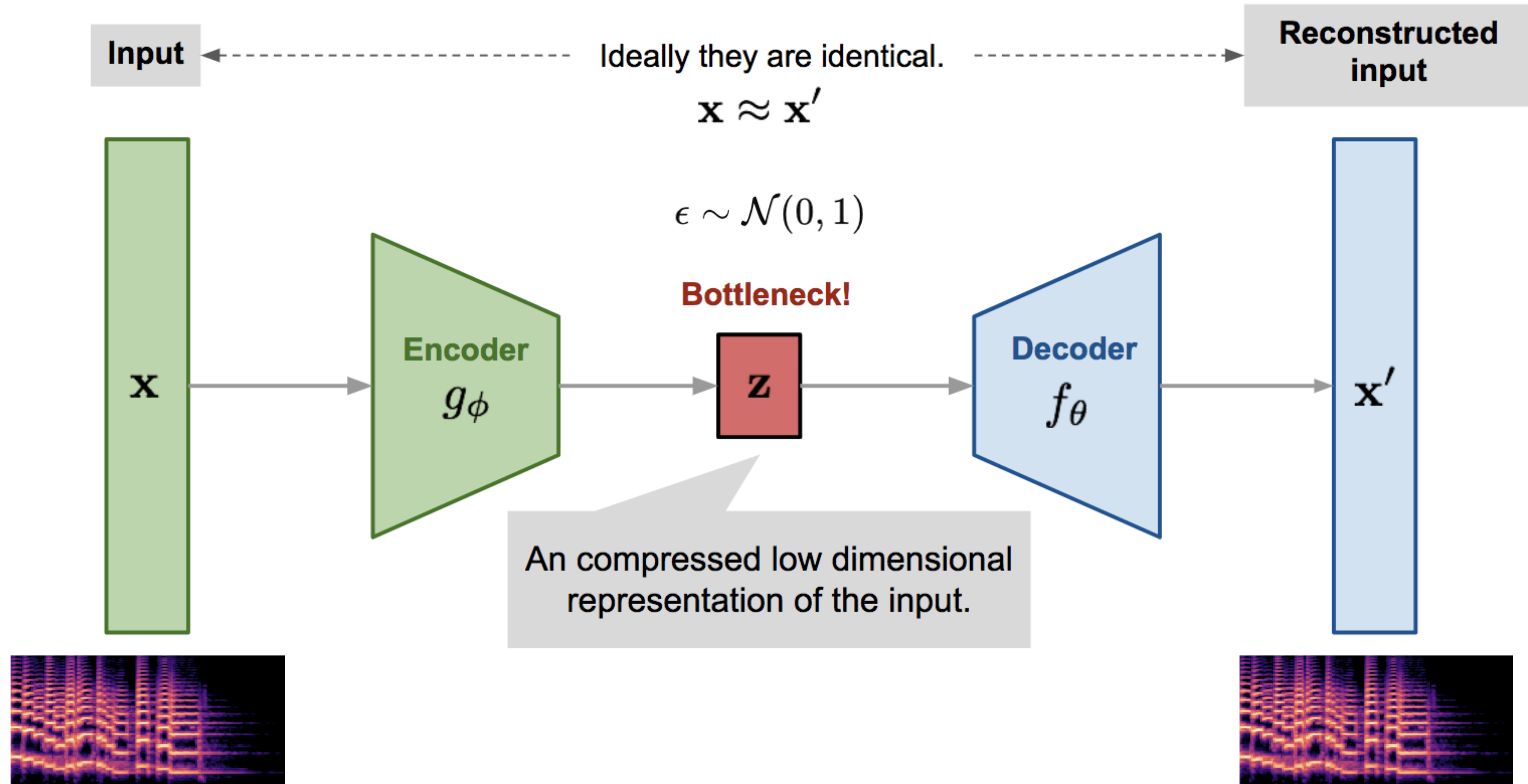
A **mel-spectrogram** represents audio energy over time on a perceptually motivated frequency scale, obtained by applying a mel filter bank to the power spectrogram. Since phase information is typically discarded, it cannot be directly inverted back to the waveform.

Mel-to-Waveform Vocoder

An example: BigVGAN



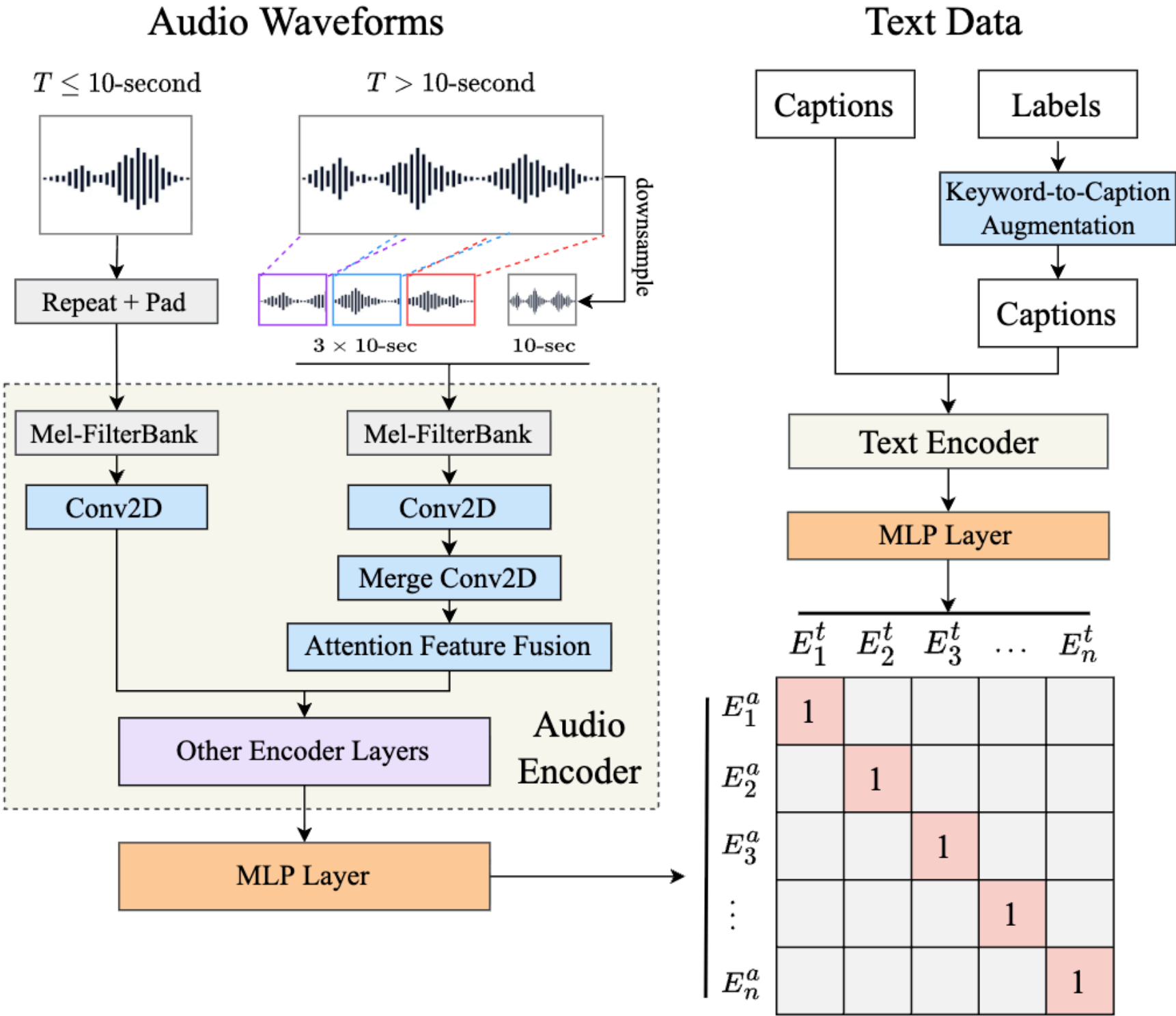
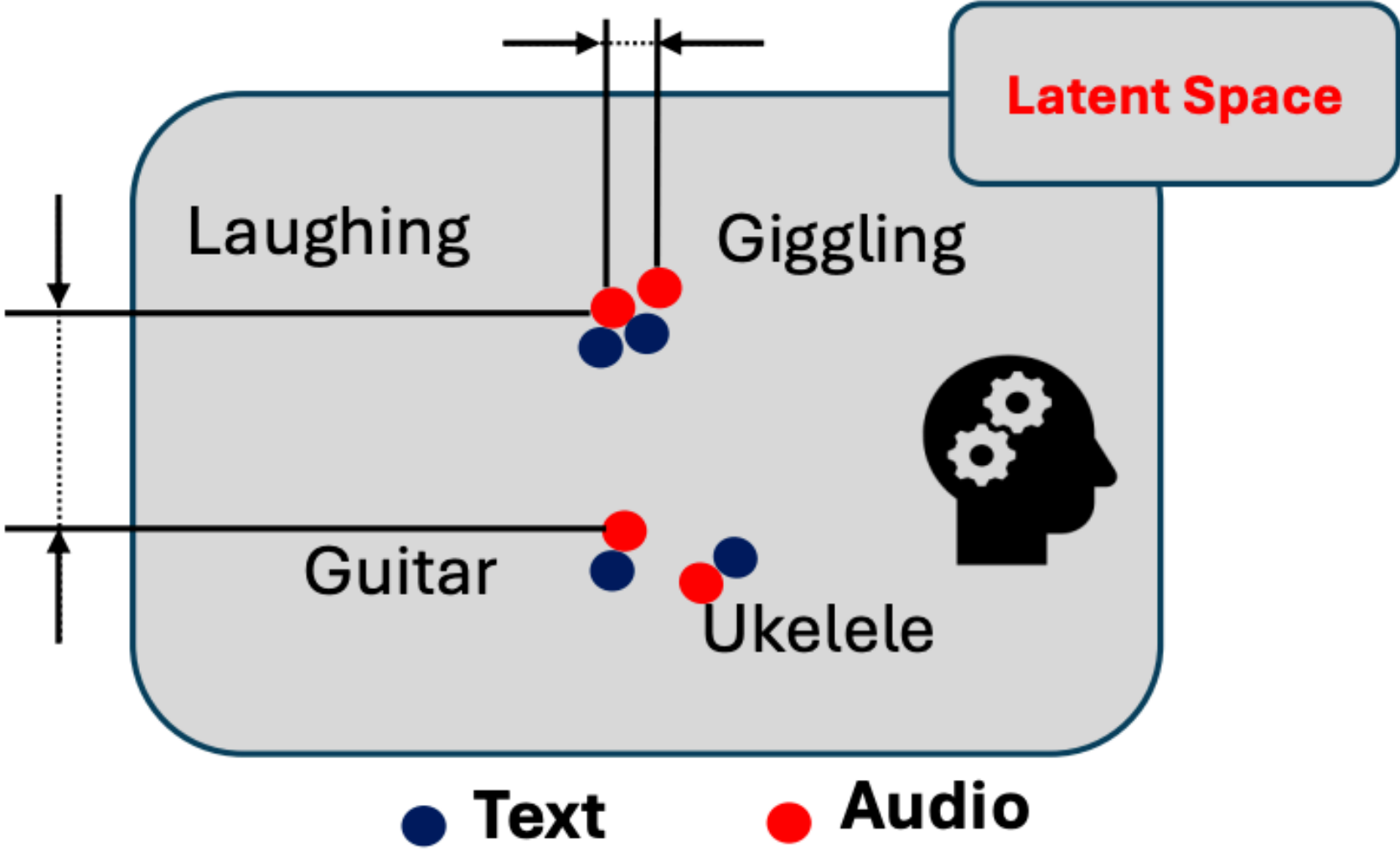
Variational Autoencoder (VAE) for Audio



Contrastive Language-Audio Pretraining (CLAP)

$$\mathcal{L}_{\text{CLAP}} = \frac{1}{2} (\mathcal{L}_{a \rightarrow t} + \mathcal{L}_{t \rightarrow a})$$

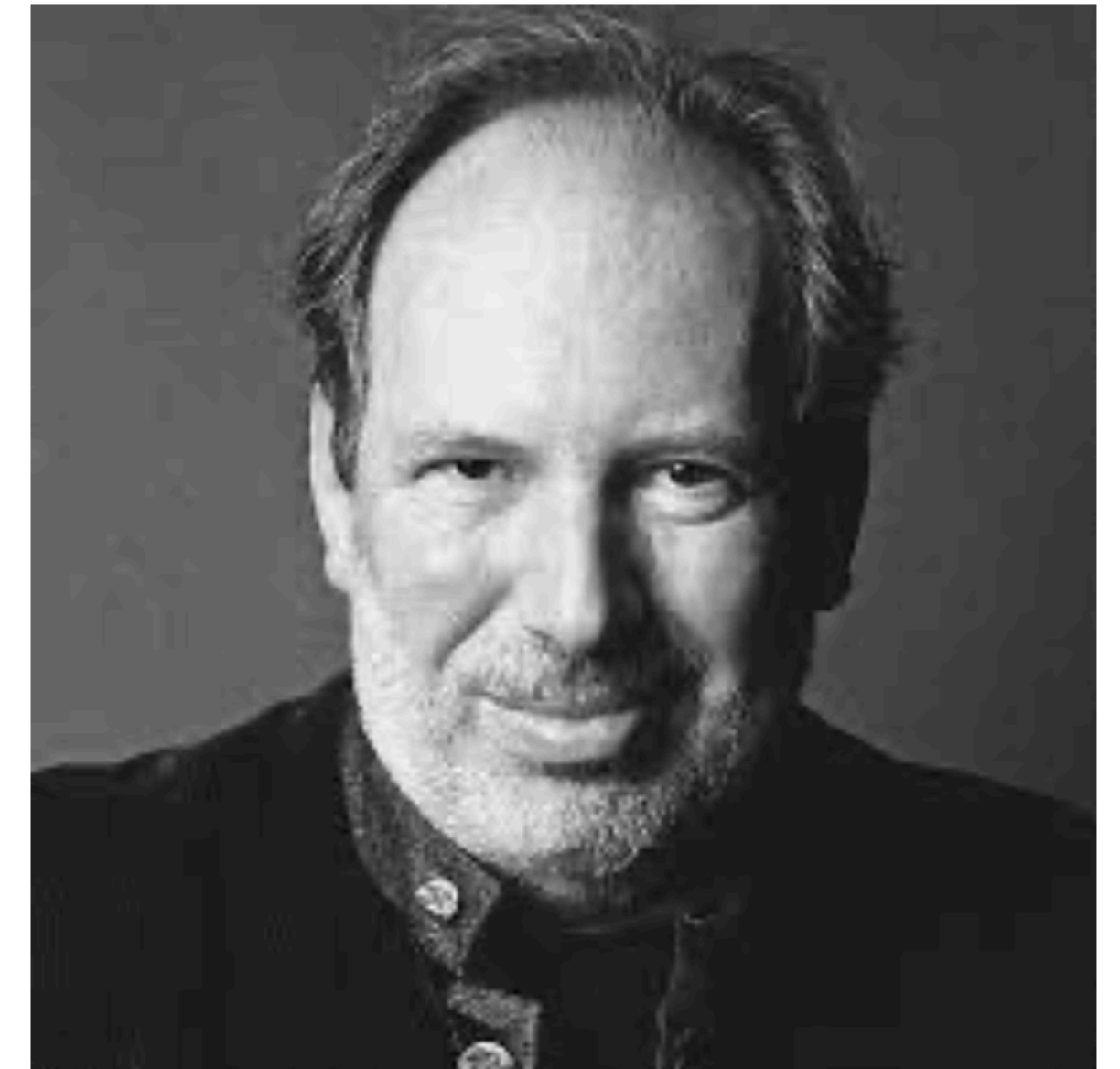
- Learning an aligned space between language and audio



Audio Generation

Audio Generation

- Text-to-Speech: Generating speech from text. e.g., “Hello, how are you today?”
- **Text-to-Audio: Generating general sounds from textual descriptions.**
 - *e.g., “A dog barking in the distance while rain falls heavily.”*
- 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.*
 - ...
- Data Augmentations & Simulations



Sound is often the unsung hero of the movie world
- Hans Zimmer

Text-to-Audio Generation

<https://audioldm.github.io/>

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

- **Key Components:**

- **Contrastive Language-Audio Pretraining (CLAP)**

- Align audio and text in one space

- **Mel-spectrogram VAE**

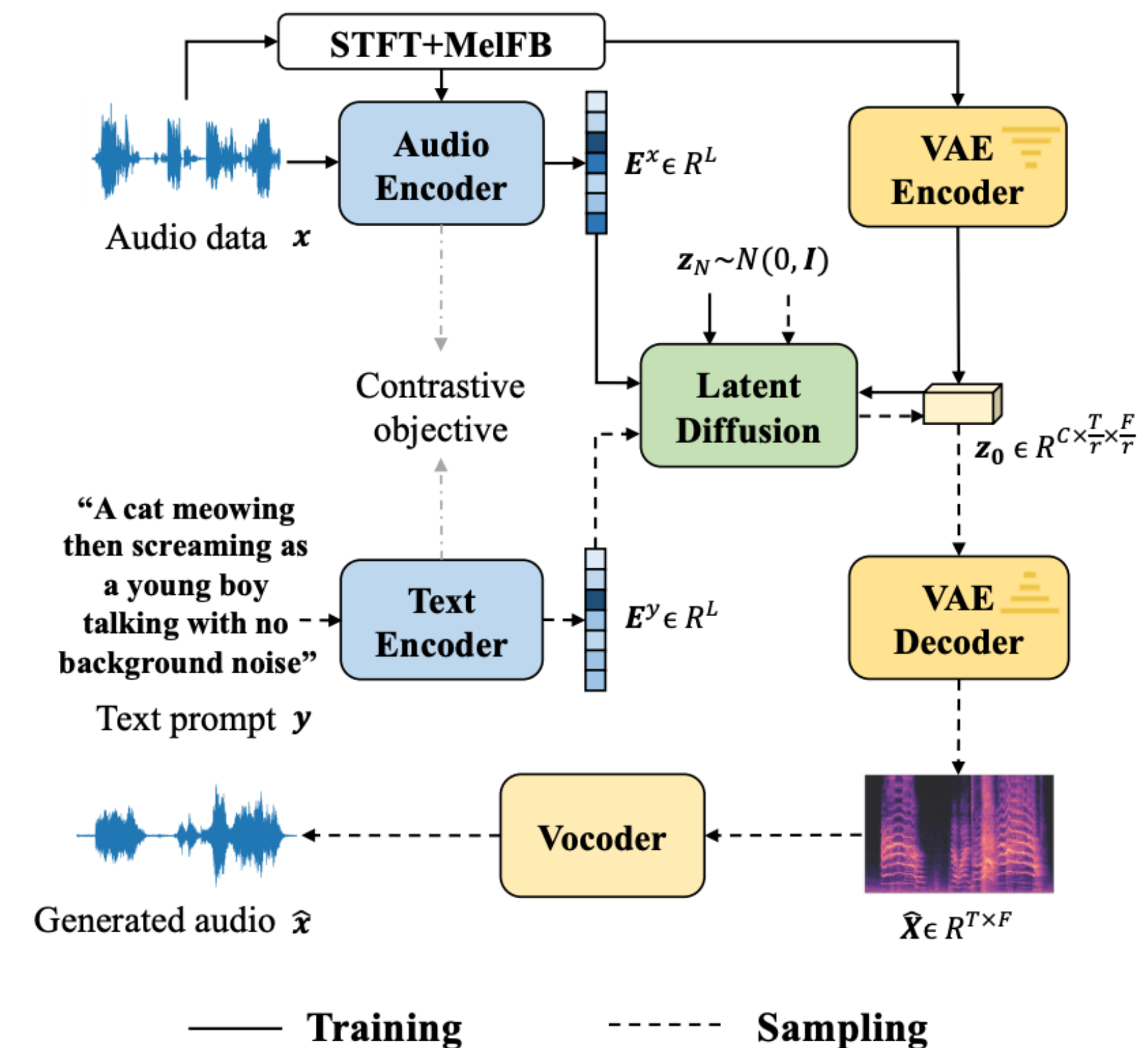
- Latent representations

- **Latent Diffusion Models**

- Learn to generate VAE latent conditioned on CLAP embeddings

- **Mel-to-Waveform Vocoder**

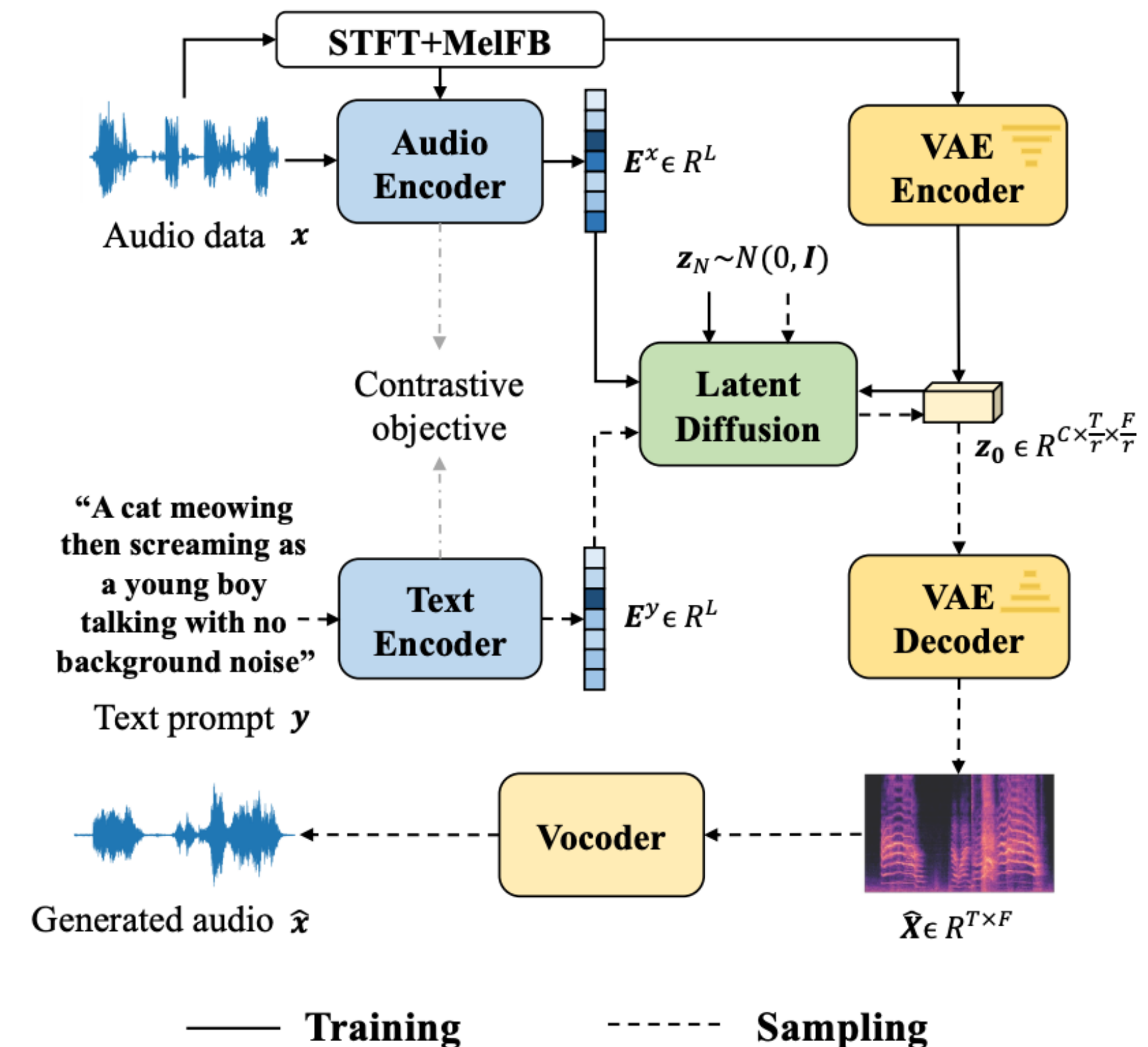
- Reverse Mel back to waveform



Text-to-Audio Generation

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

- **Self-supervised text-to-audio generation:**
 - Enabling audio-only semi-supervised training approach
- **Why?**
 - Easily scale training data, since we're not limited by the need of text captions
 - Increase the robustness to noisy text descriptions

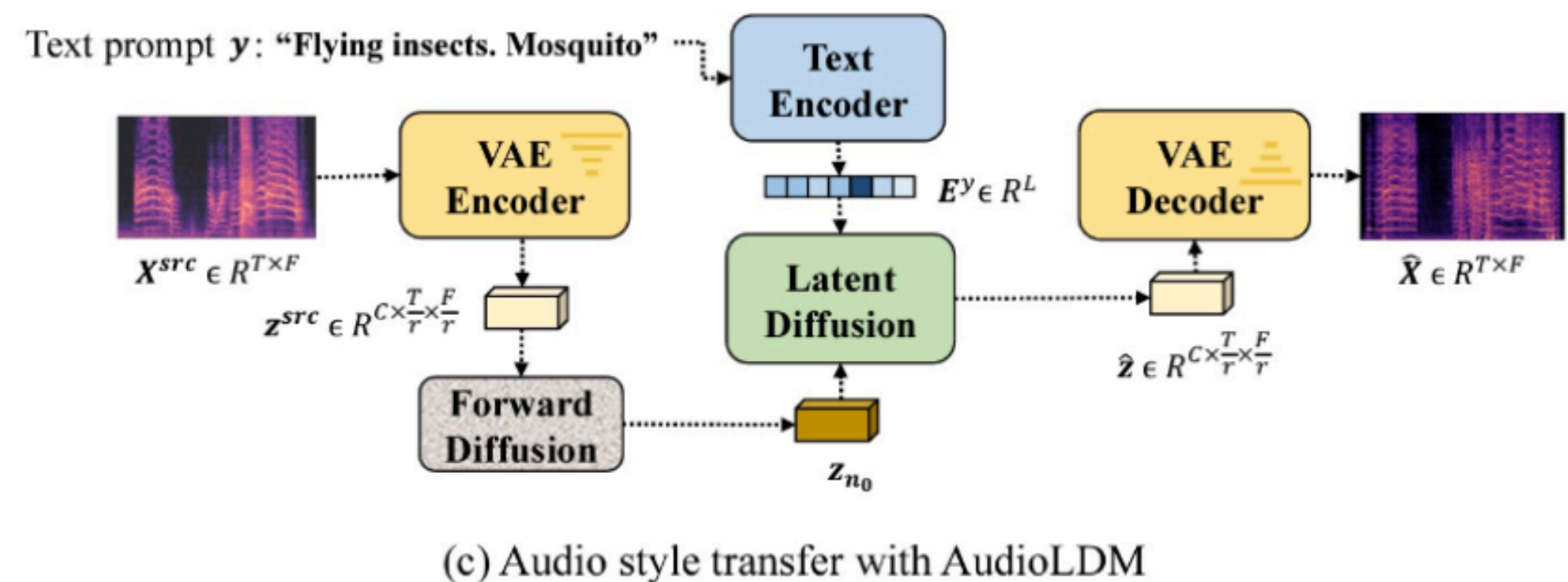
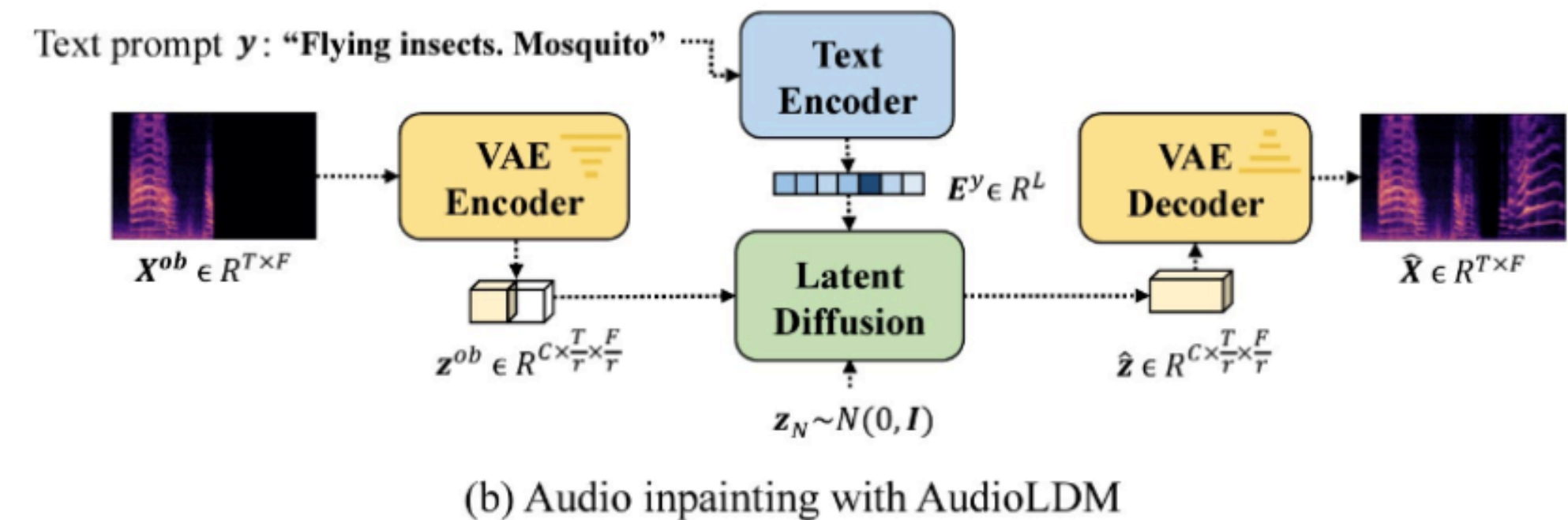


Text-to-Audio Generation

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

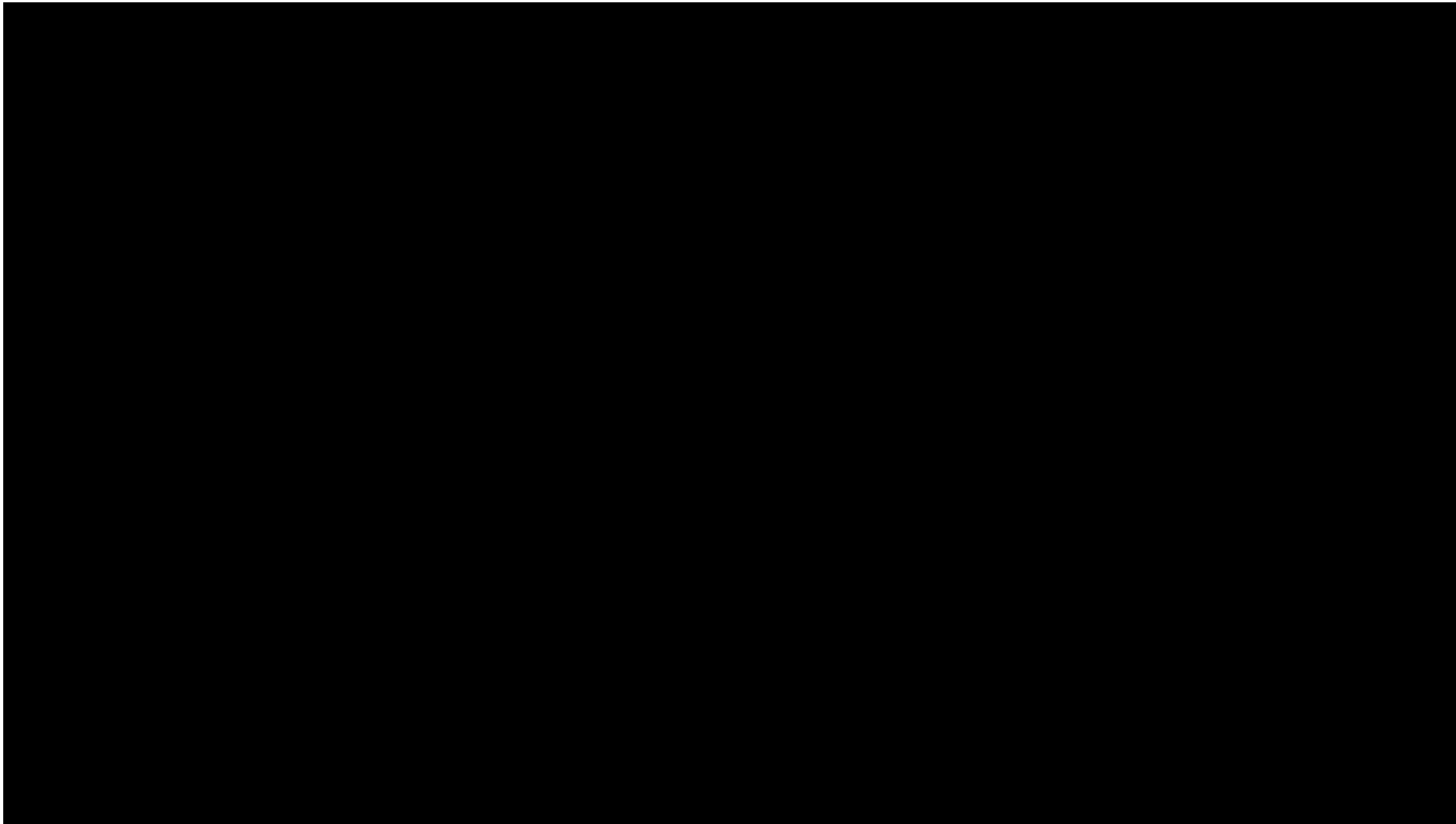
Zero-shot down stream tasks

- Audio style transfers
 - Corrupt -> Reverse Diffusion
- Audio inpainting
- Audio super-resolutions
 - Provide temporal hint during sampling
- Audio style transfer
 - Provide frequency hint during sampling



Text-to-Audio Generation

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



Language Modelling vs Diffusion

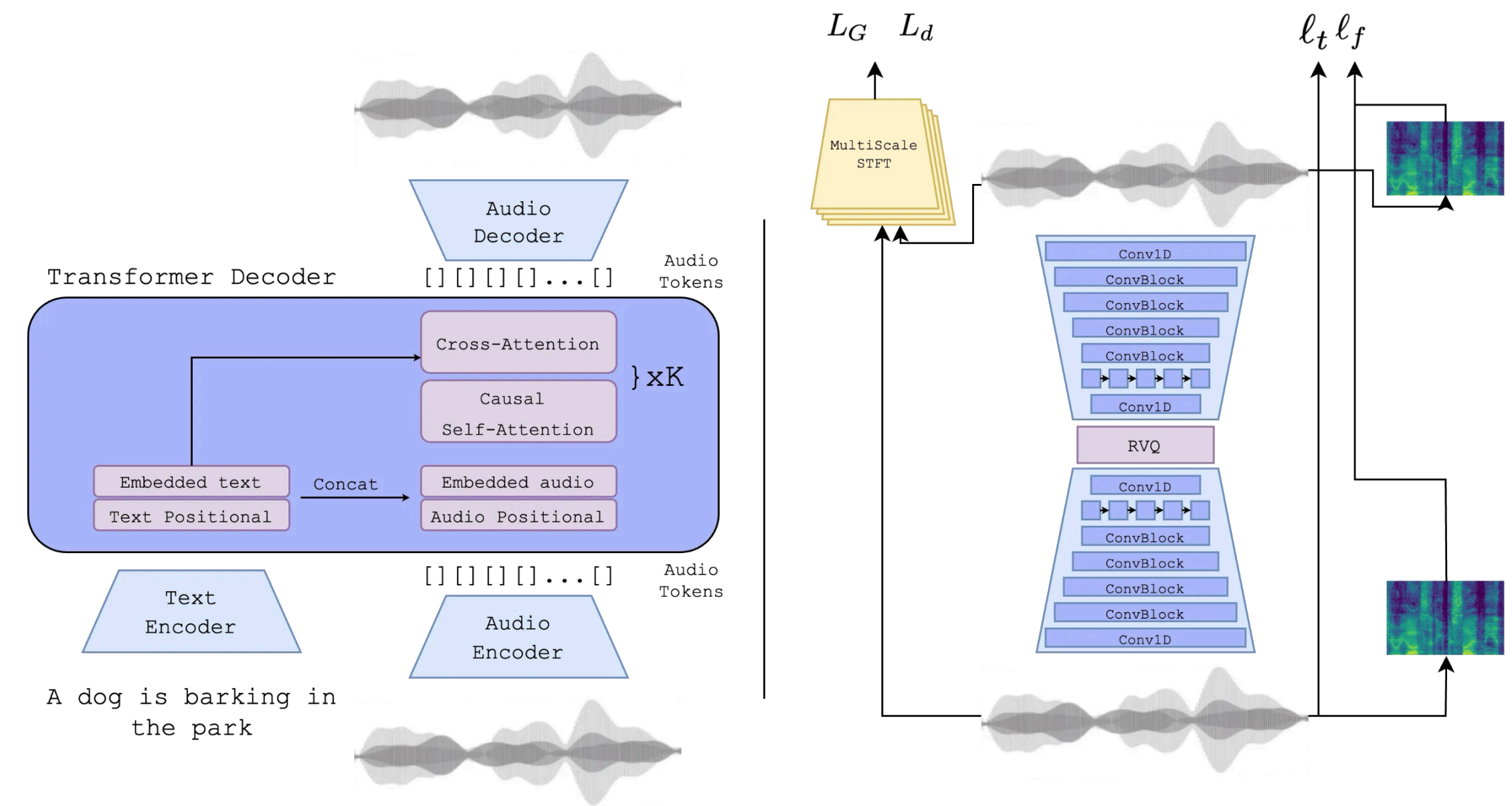
Auto-regressive language modeling

-  Explicit modeling of temporal dependencies.
-  Enjoy the advance of recent LLM development.
-  Good in-context learning performance.
-  Long generation sequence/lack of parallelism
-  Long range dependencies
-  Error propagation

Diffusion-based approach

-  Stable
-  State-of-the-art generation quality
-  Flexible formulation for manipulation, interpolation, etc.
-  Do not explicitly model temporal dependencies
-  Less flexible on duration

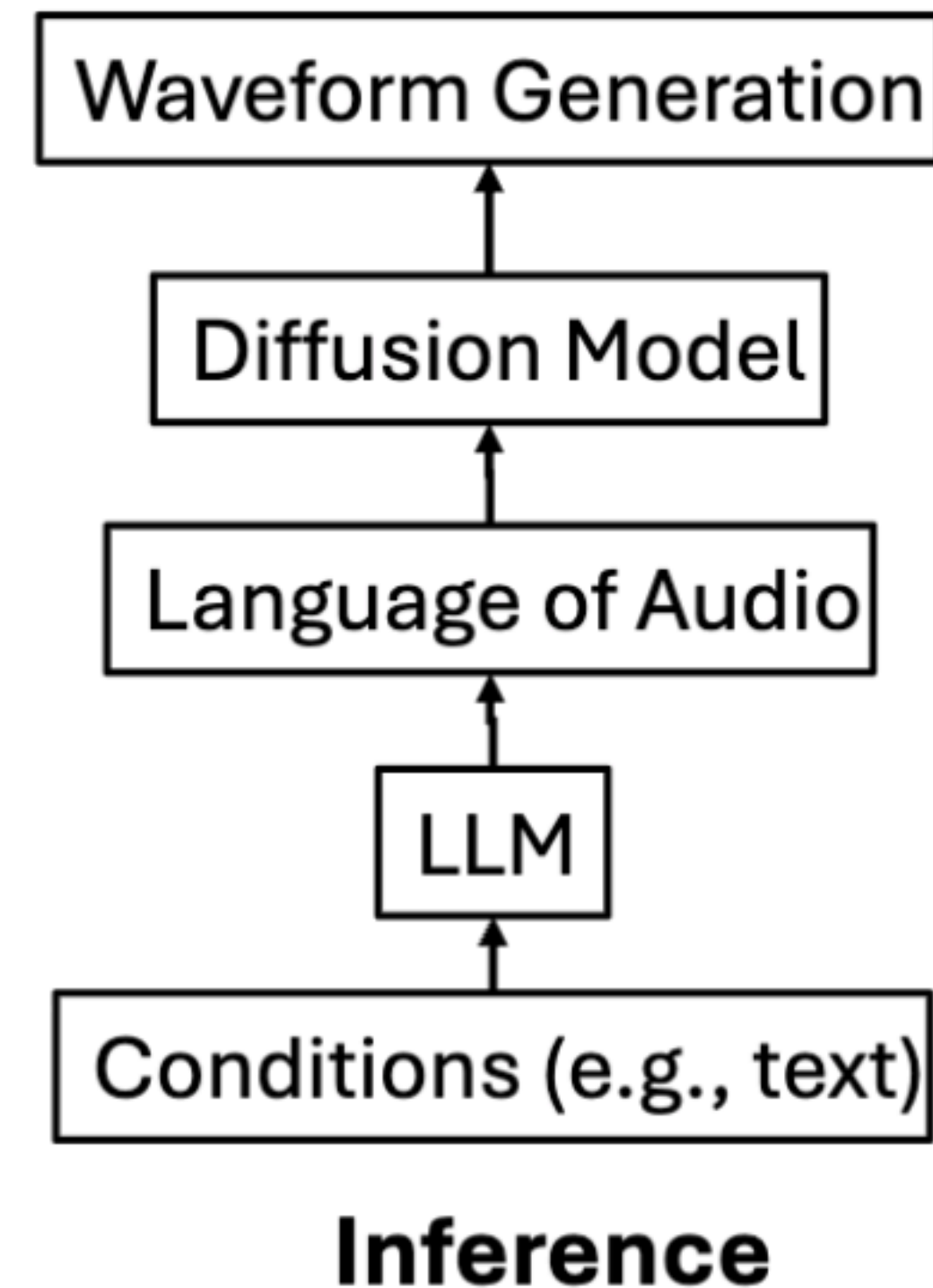
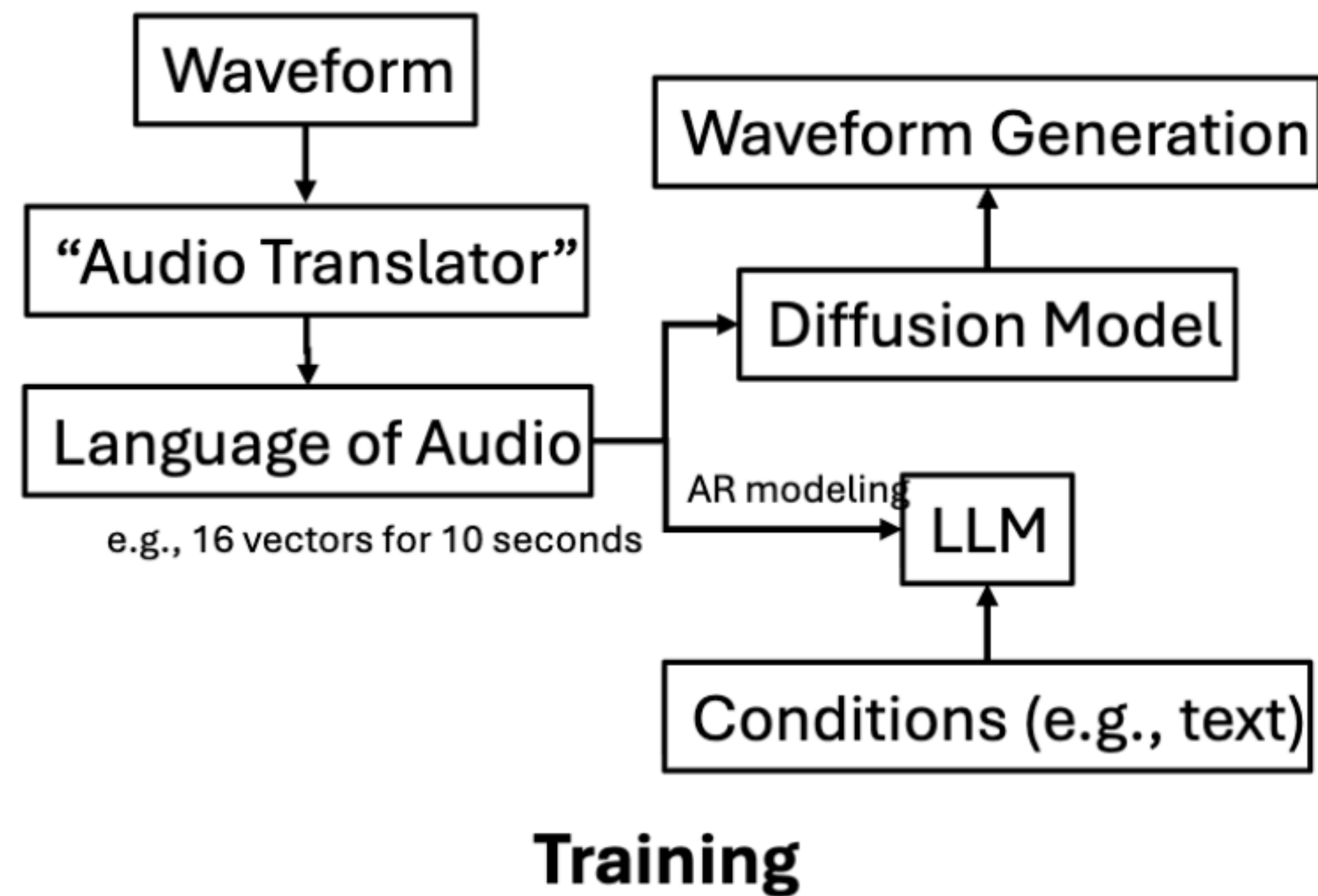
Can we utilise both advantages from LM and Diffusion?



AudioGen (Kreuk et al. 2022)

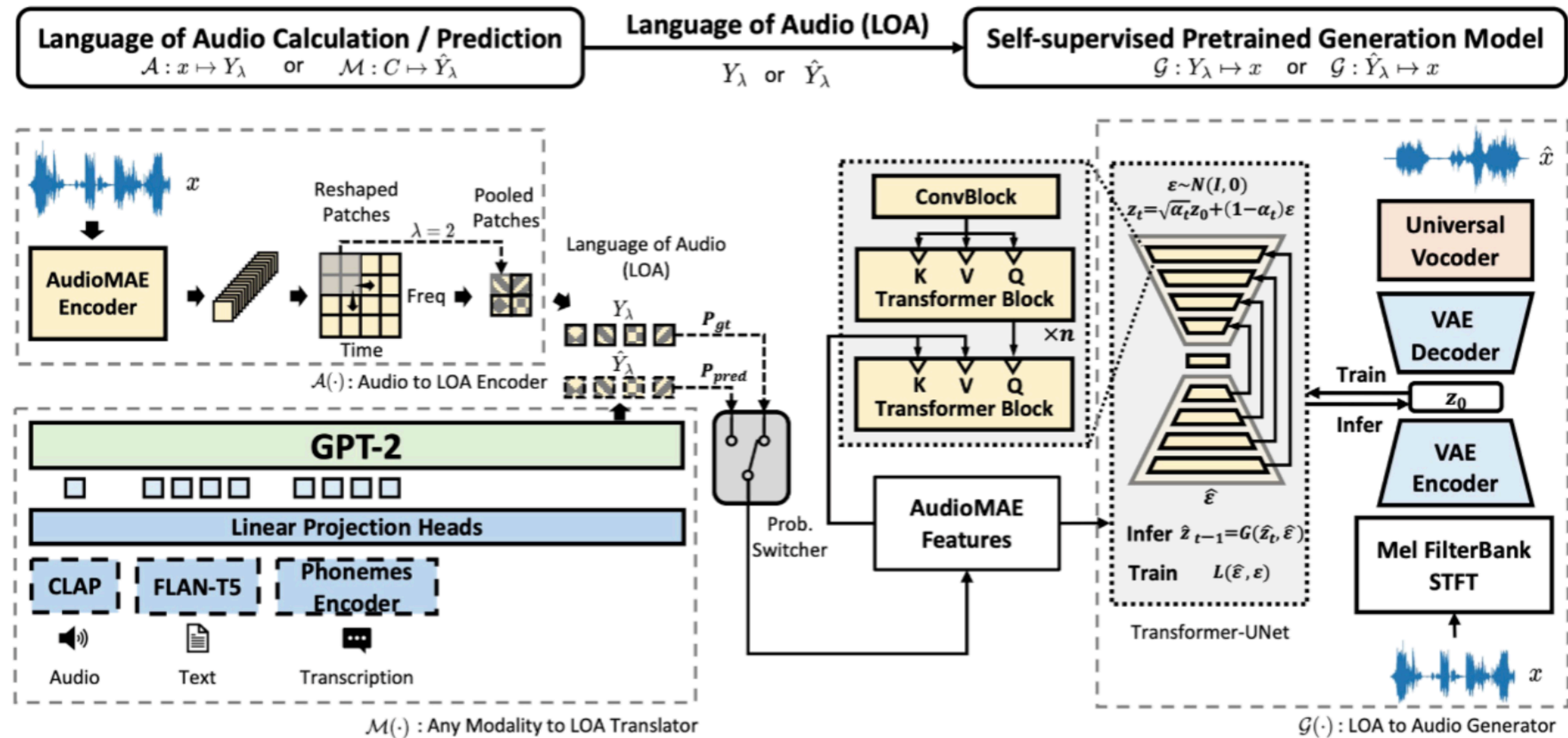
Language Modelling + Diffusion

1. Auto-regressive modelling of the semantic audio sequence
2. Reconstruct semantic audio sequence to waveform with Diffusion



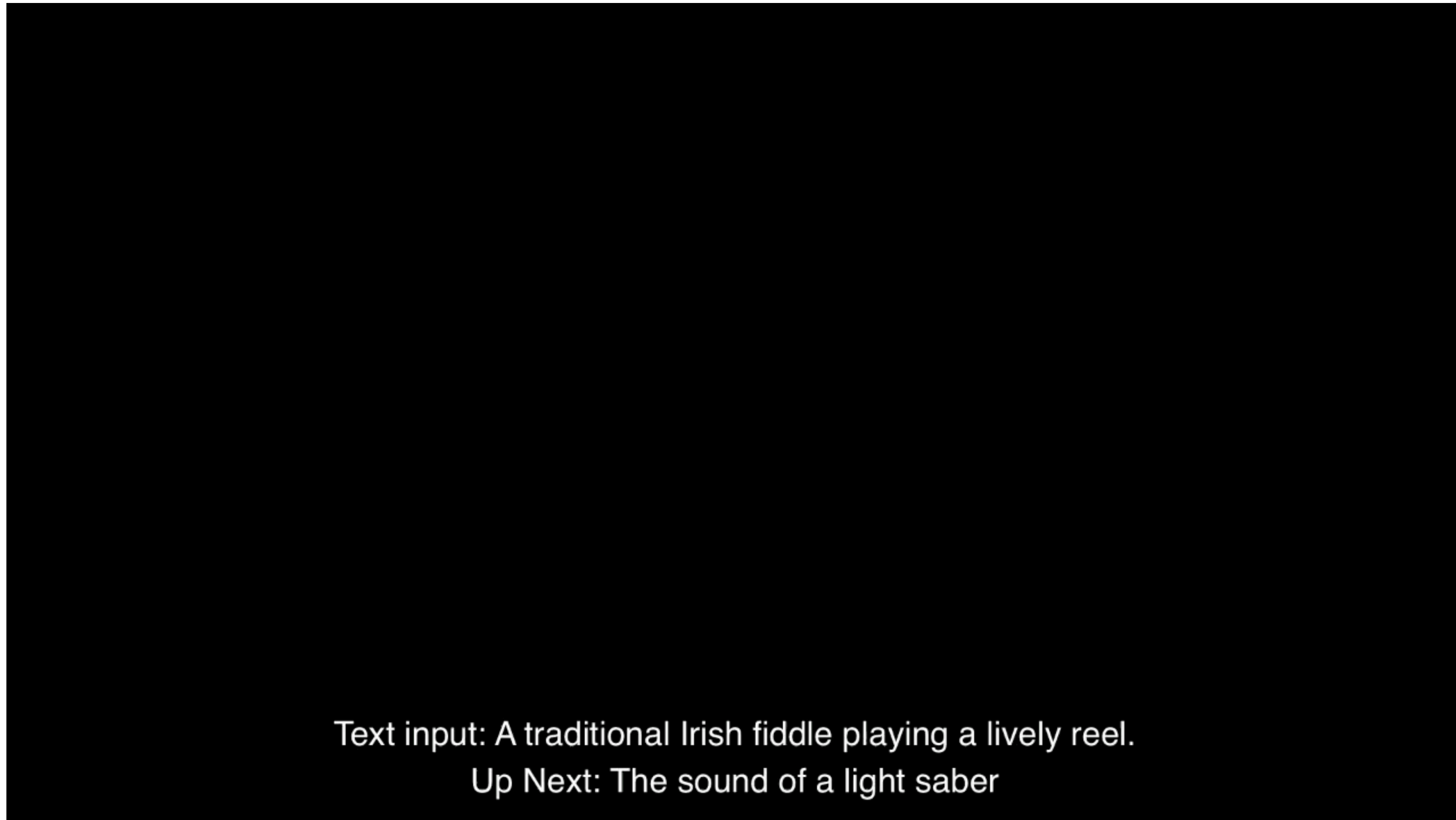
Text-to-Audio Generation

AudioLDM2: Combining LLMs with Diffusion



AudioLDM 2

Text-to-Audio/Music/Speech Generation



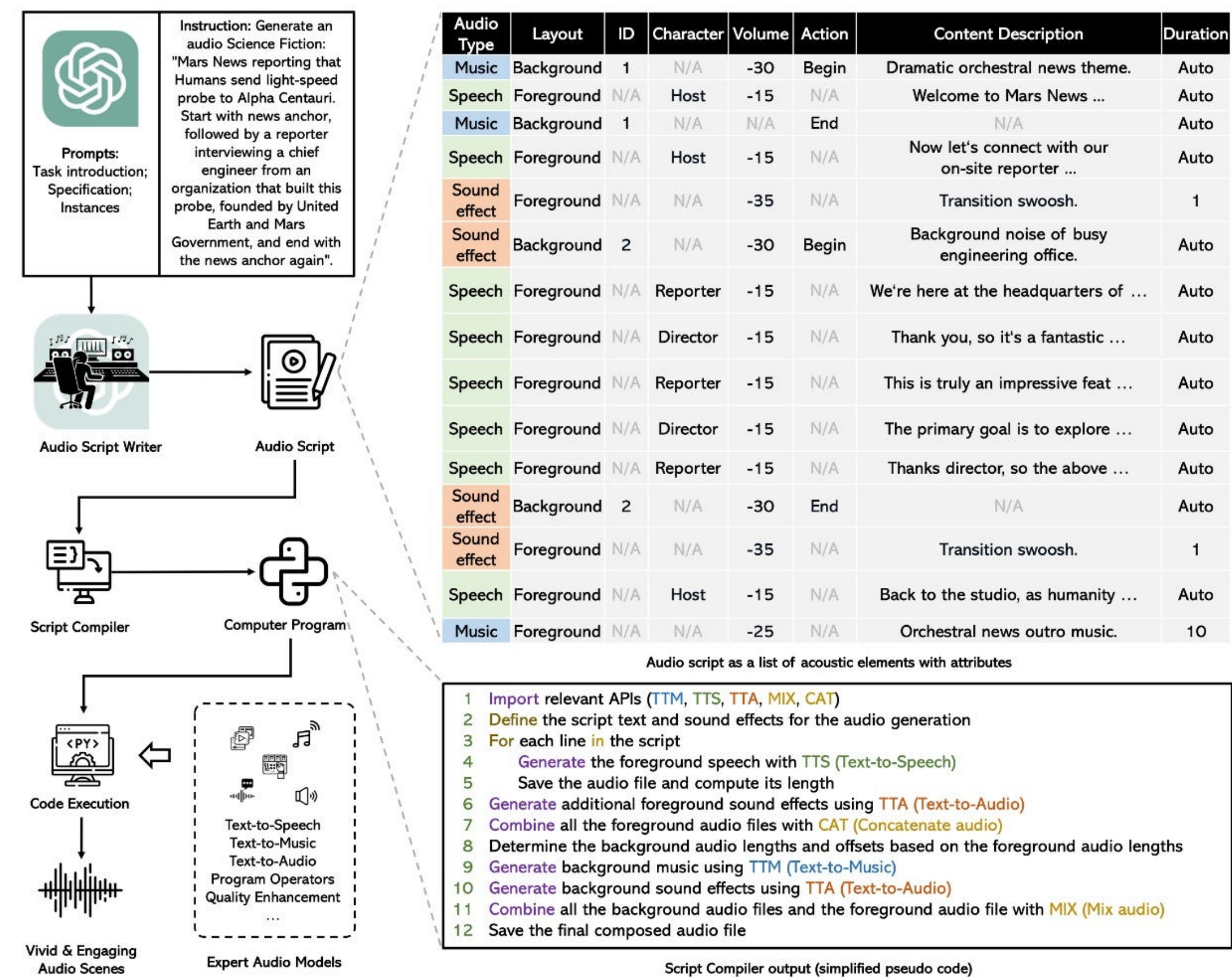
Audio Storytelling

WavJourney: Compositional Audio Content Creation with LLMs

Create audio storytelling with:

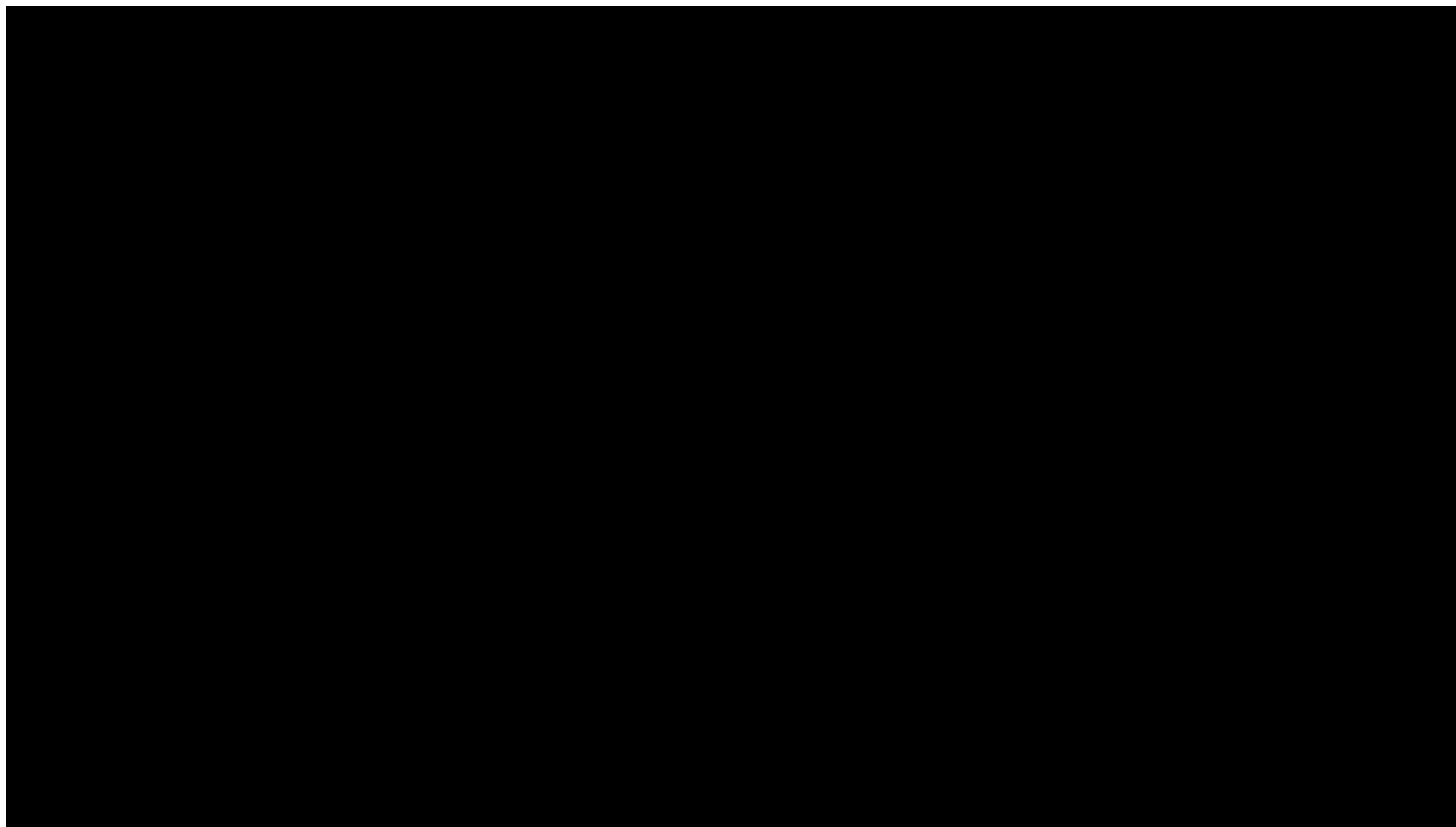
- Personalized speakers
- Lifelike speech
- Immersive music
- Impactful sound effects

All simply from texts!



Audio Storytelling

WavJourney: Compositional Audio Content Creation with LLMs



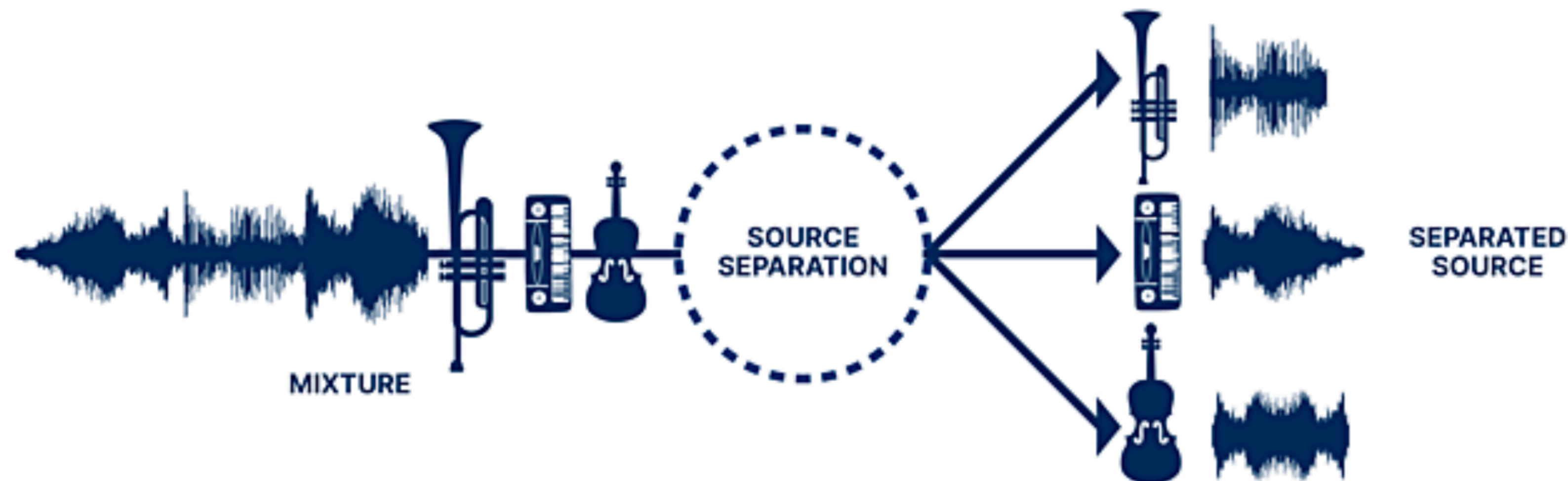
Audio Source Separation

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

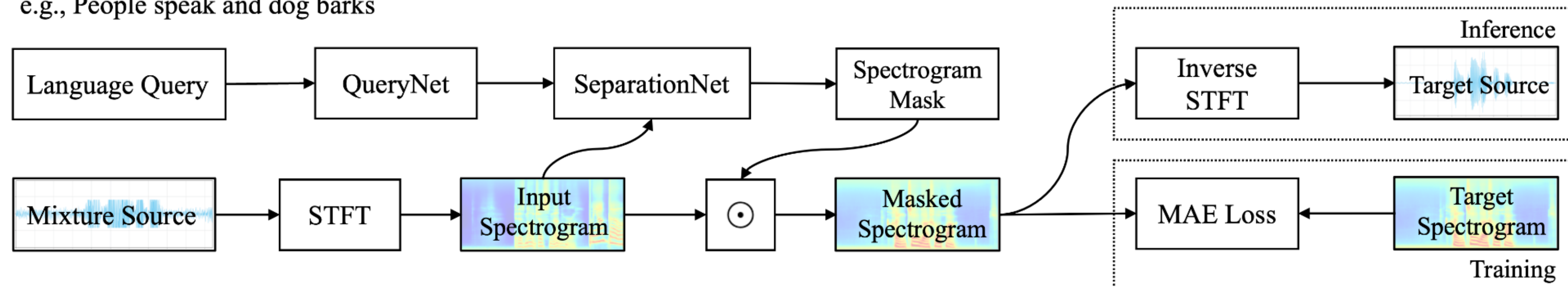
Language Query: A bird is chirping under the thunder storm



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

e.g., People speak and dog barks



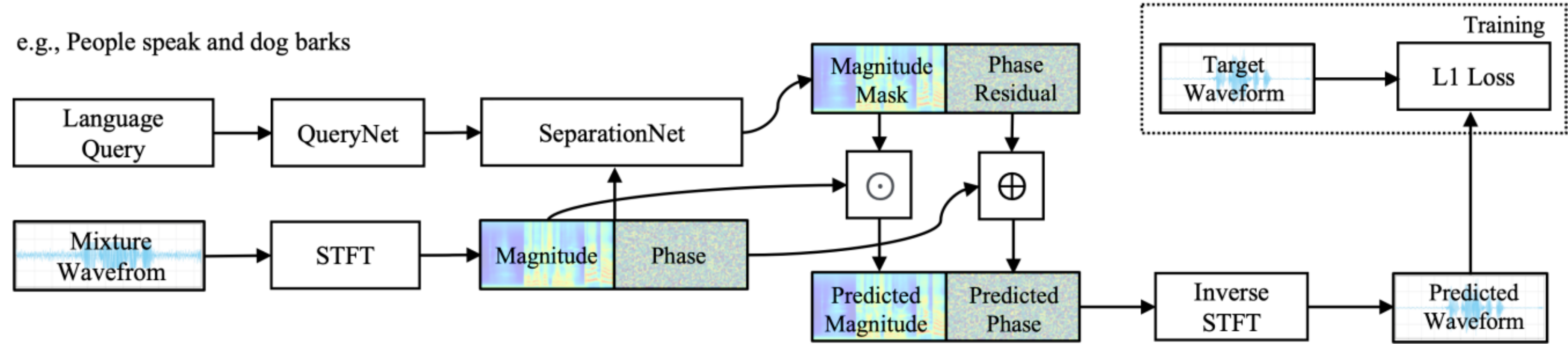
Language-Queried Audio Source Separation

AudioSep - Separate Anything You Describe

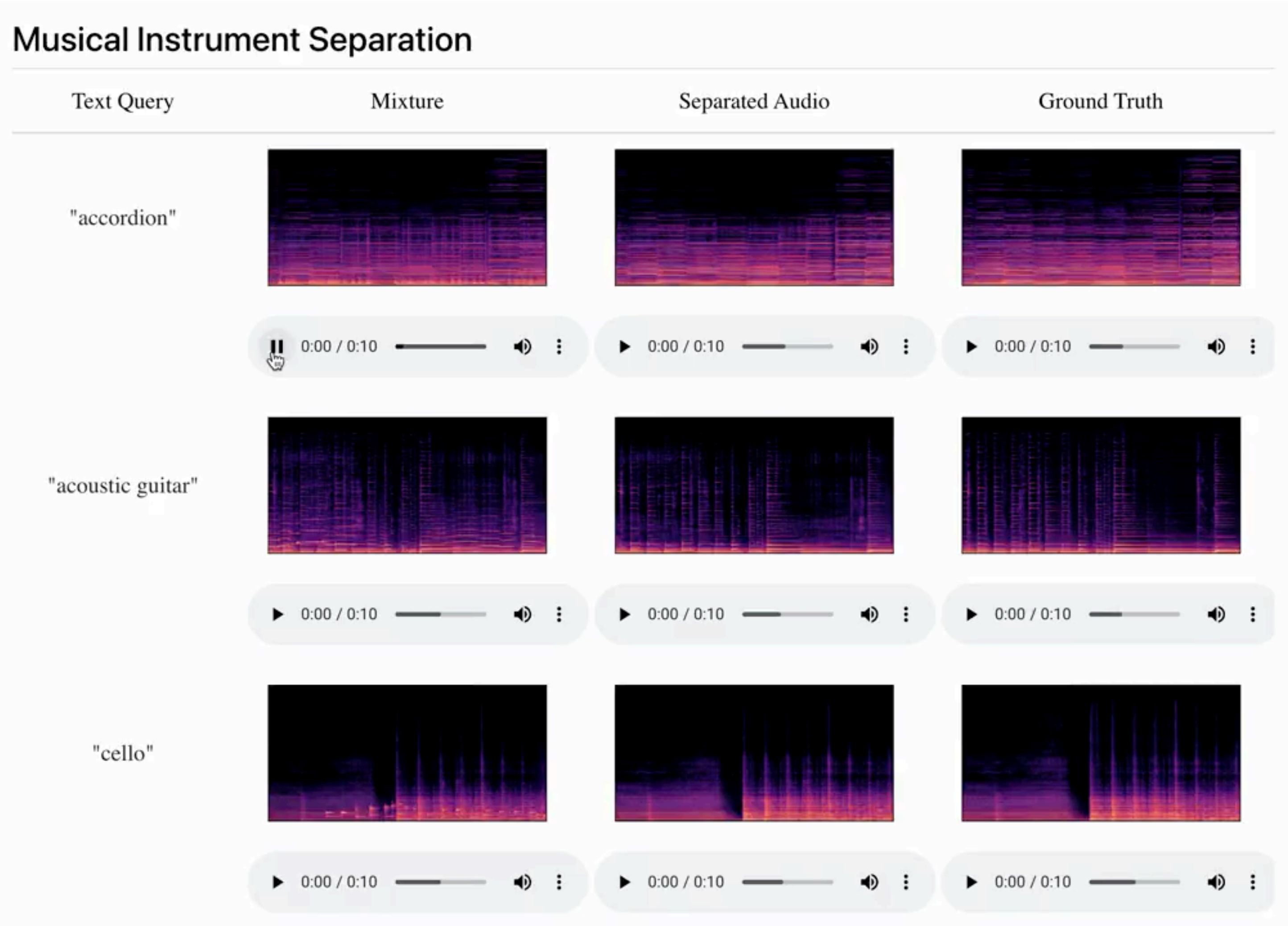
- CLAP/CLIP + ResUNet, scaling with **14,000** hours of multimodal data
- A Foundation model for open-domain sound separation with texts
- Impressive zero-shot performance in separating speech, music, sounds

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



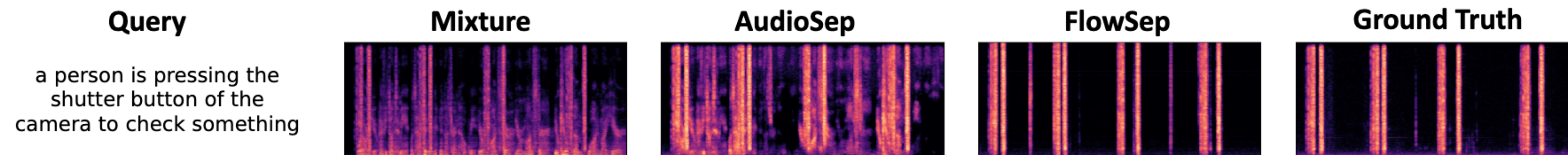
Language-Queried Audio Source Separation



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

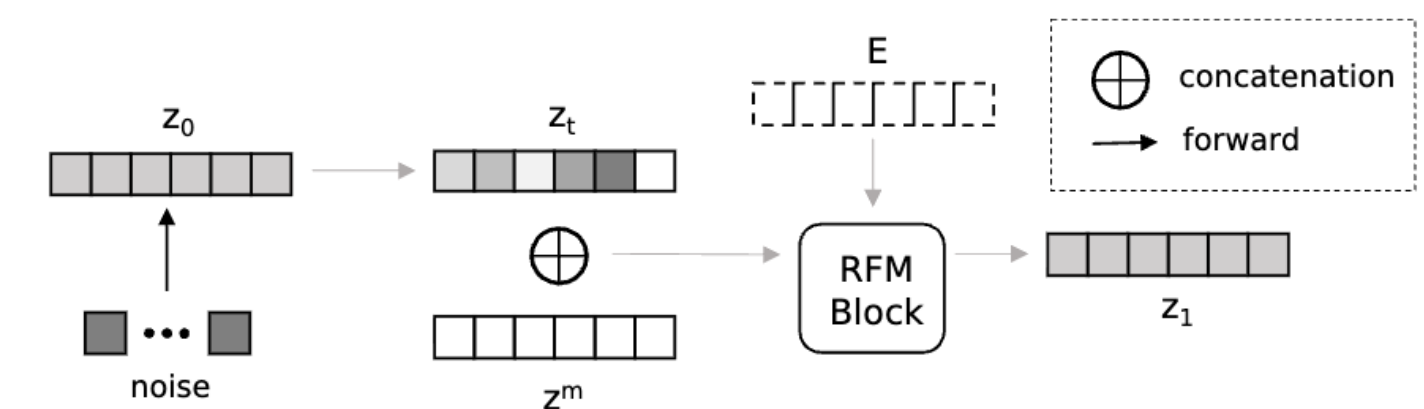
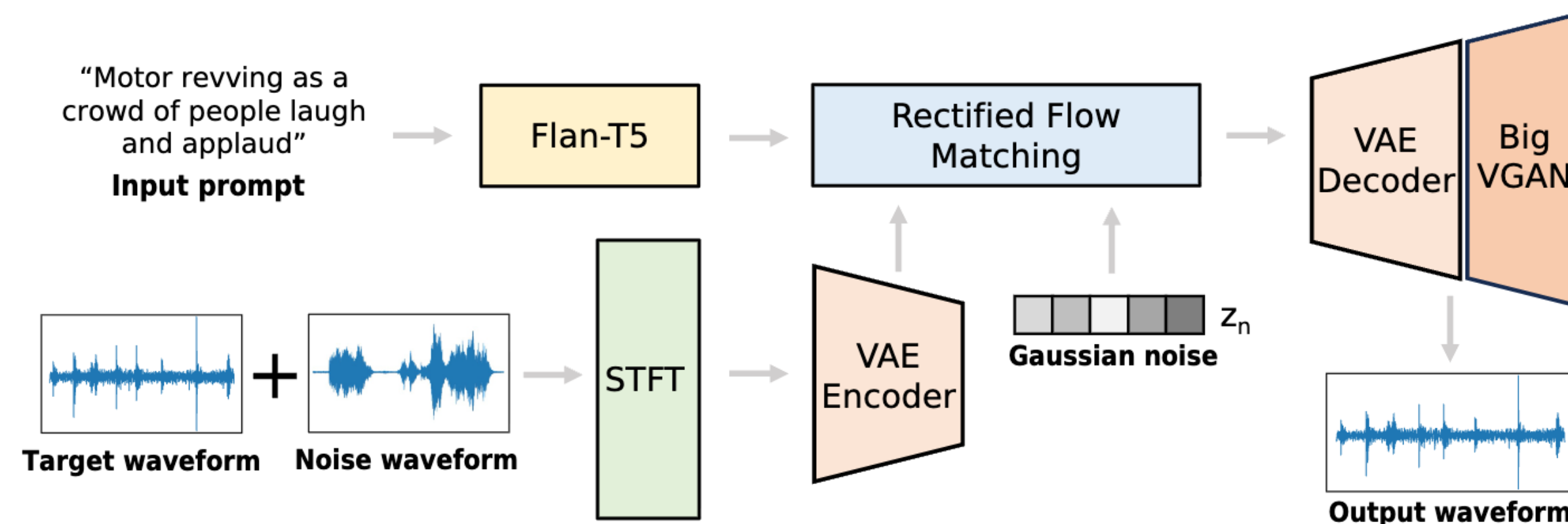


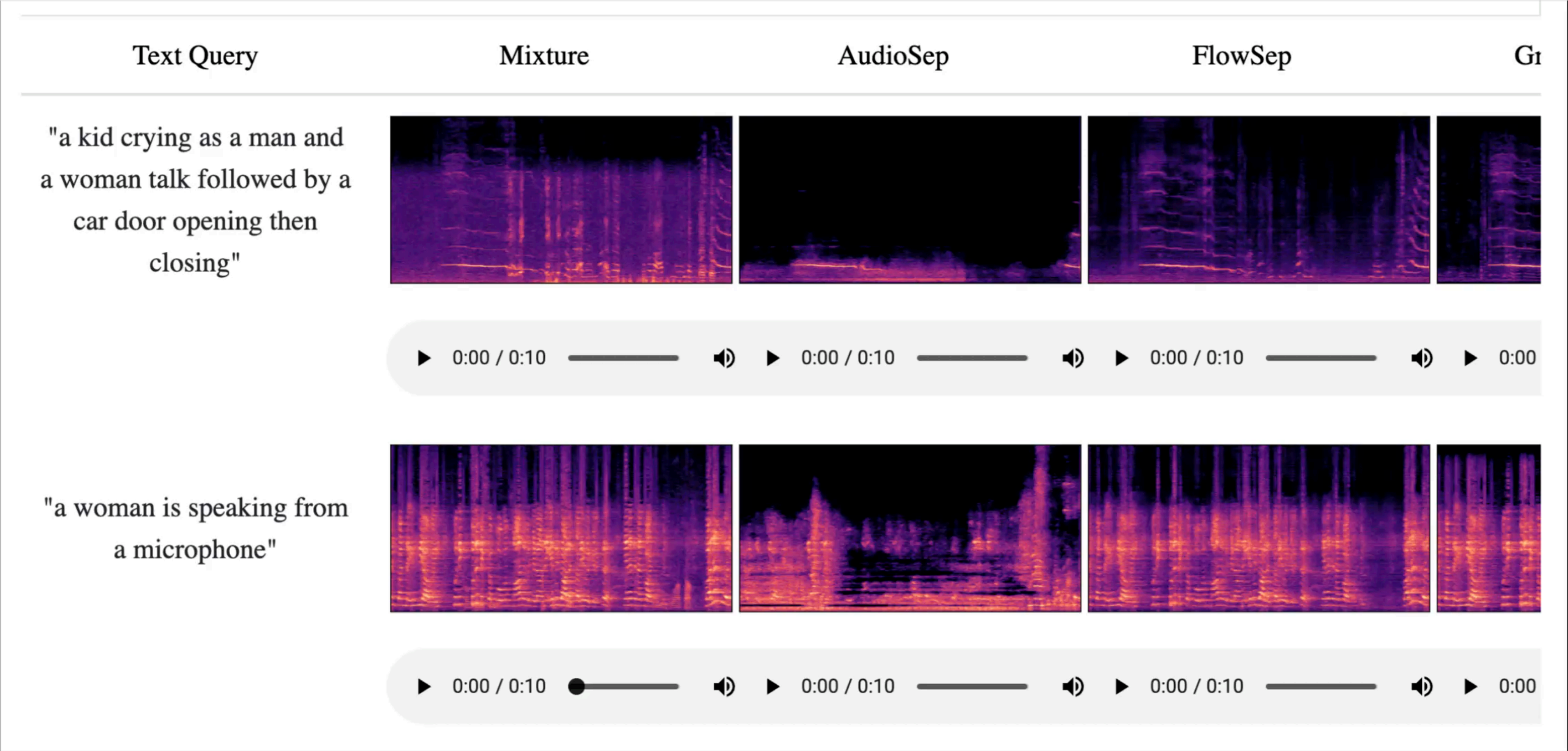
Fig. 2. The channel-concatenation conditioning mechanism

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.

Language-Queried Audio Source Separation

FlowSep - LASS with Rectified Flow Matching

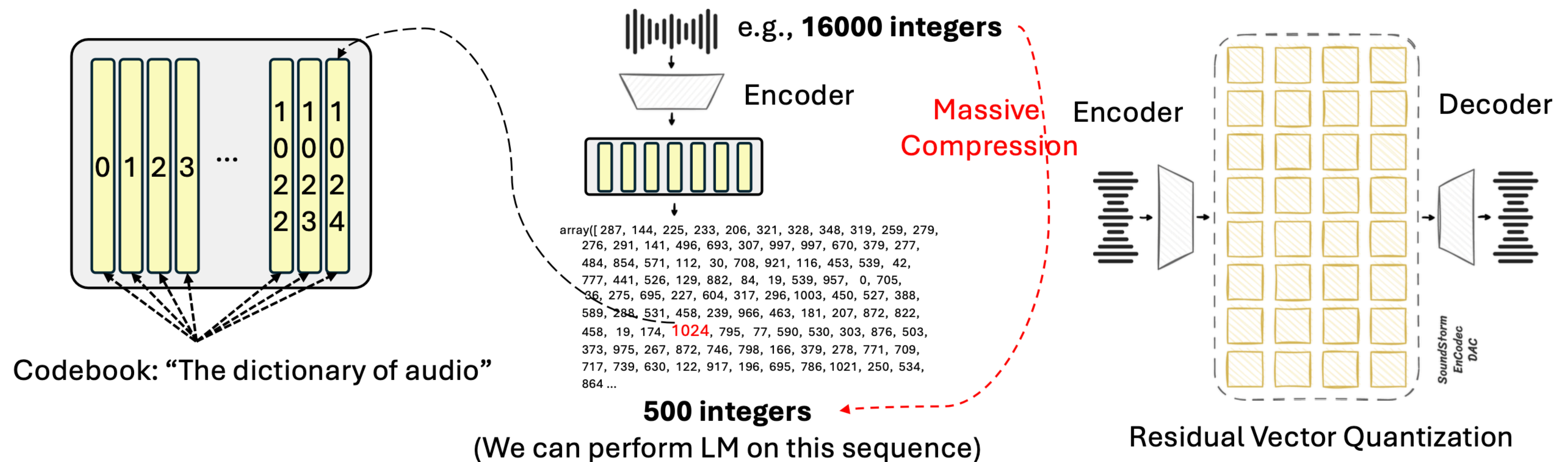
- Open challenges: reduce generative hallucinations



Representation Learning

Neural Audio Codec / Audio Tokenizer

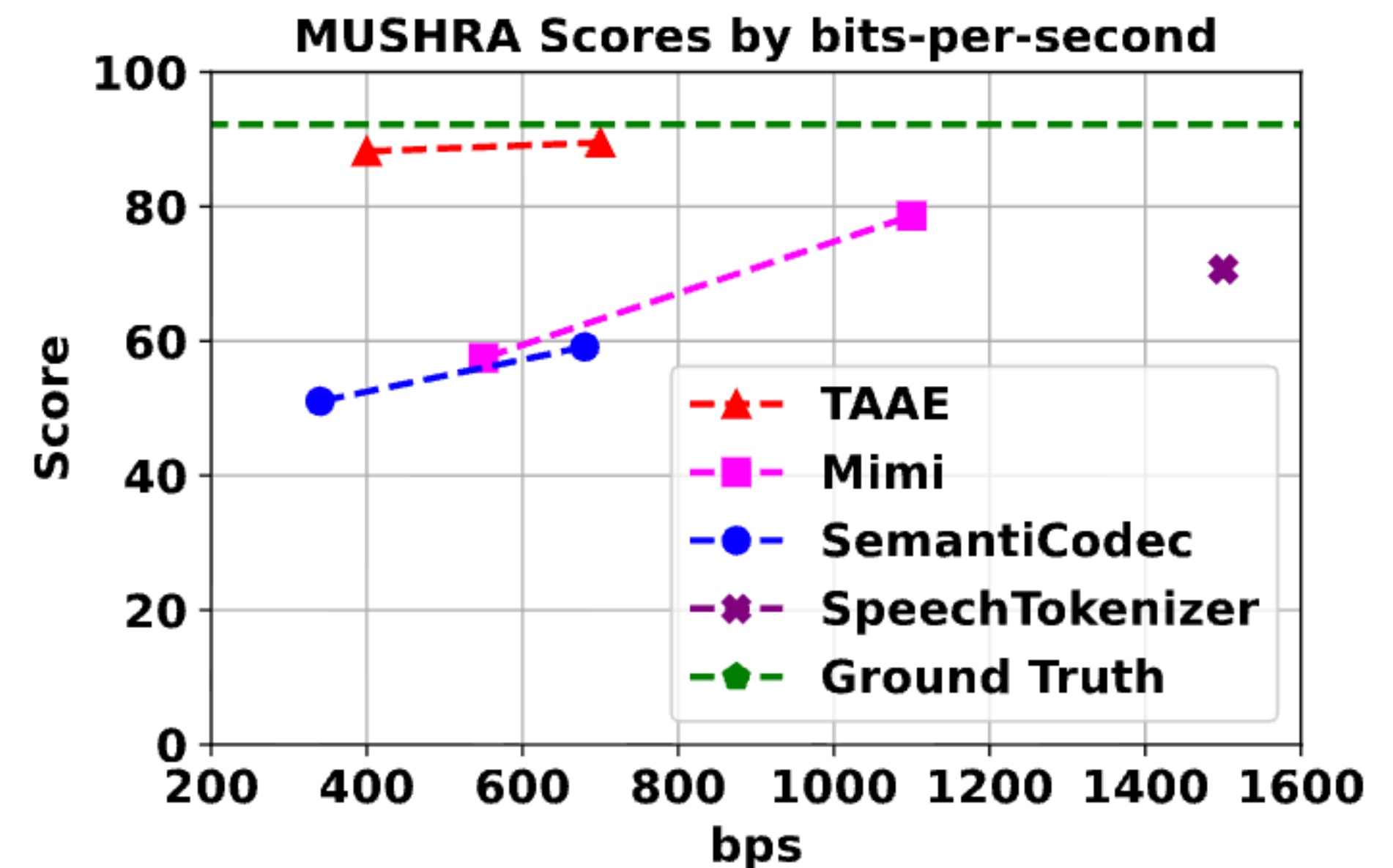
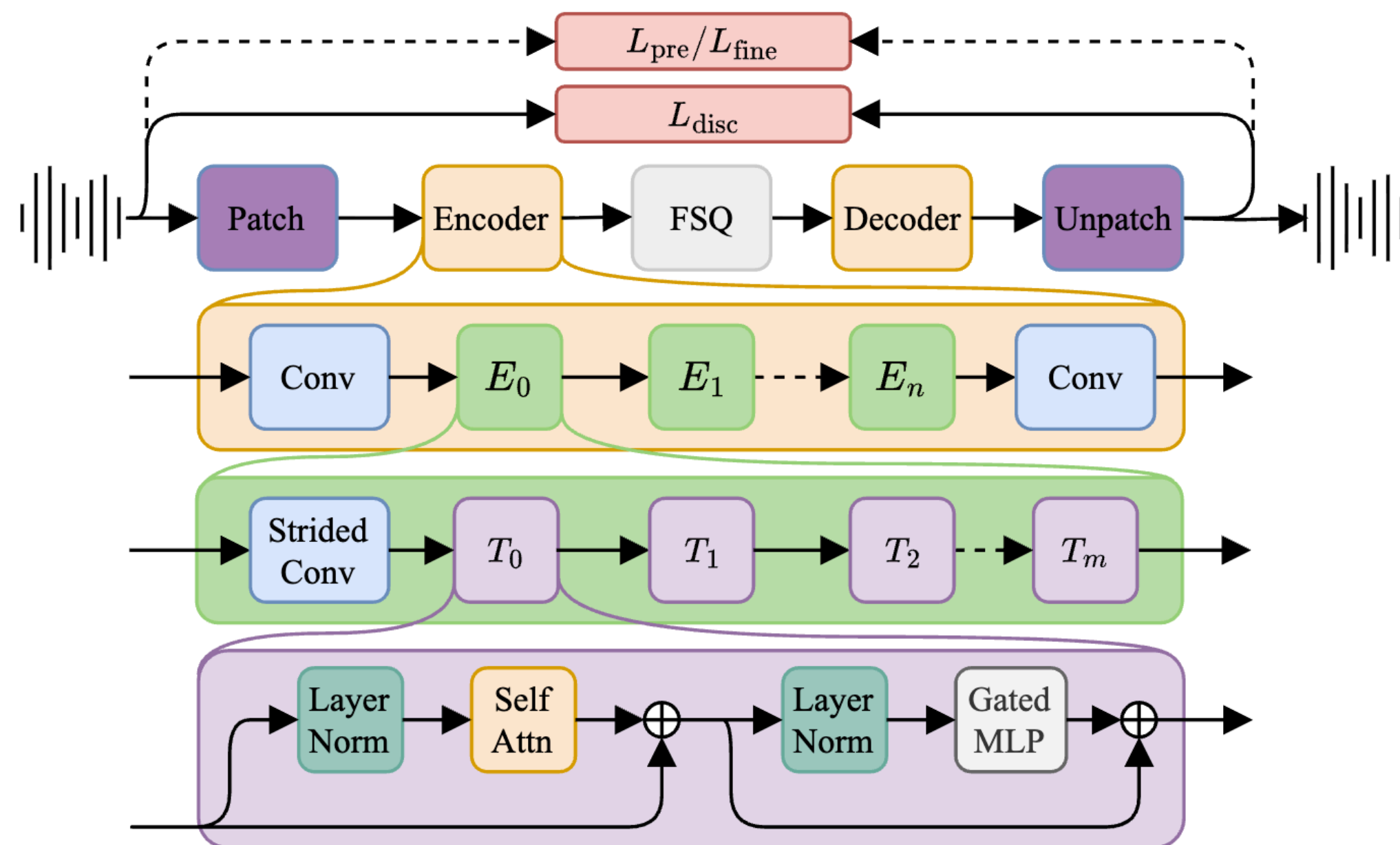
- The “language” of audio
- Compression and discretization of audio signal
- Representative works: Soundstream, Encodec, Descript Audio Codec (DAC), etc.



Transformer Audio Tokenizer github.com/Stability-AI/stable-codec

Limitations of SoTA audio tokenisers: (1) largely based on Convolutional architectures, which can be challenging to scale (2) high bit-rates, too many tokens for downstream language modelling.

Transformer Audio AutoEncoder (TAAE) - first Transformer-based audio codec scaled to 1B parameter that achieves SoTA speech quality at extremely low bit-rates of 400 or 700 bits-per-second.

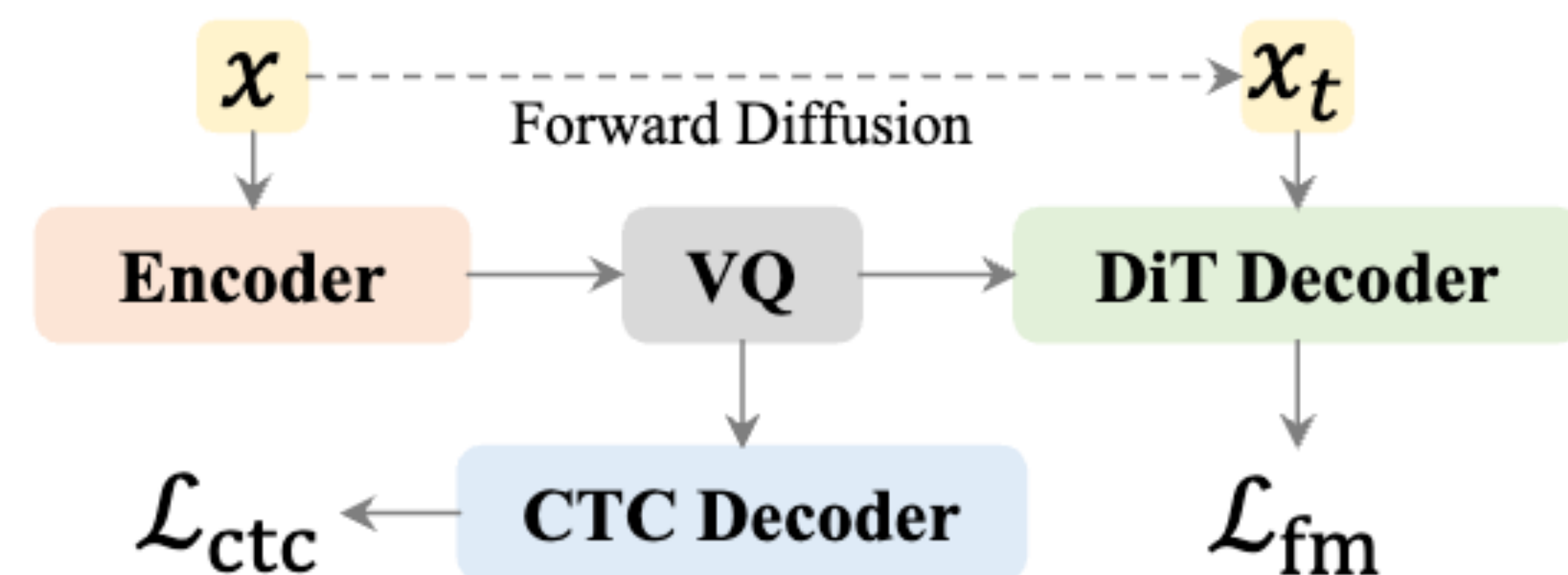


Diffusion Audio Tokenizer

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.

Main takeaways:

- **Diffusion reconstruction loss** improves the trade-off between semantic and acoustic representations under a constrained bottleneck.
- CTC supervision does matter, especially for low-bitrate (12.5 Hz)
- Diffusion Tokenizer is a scalable architecture



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

Open Challenges

- **Sampling Efficiency:** High computational cost and inference latency.
- **Long-Form Generation:** Maintaining temporal coherence over long durations.
- **Fine-Grained Controllability:** Precise control over timing, content, and acoustic attributes.
- **Semantic Consistency:** Aligning generated audio with complex text prompts.
- **Real-Time Generation:** Enabling low-latency, streaming audio synthesis.
- **Evaluation:** Lack of reliable metrics aligned with human perception.

...

Acknowledgements

Dr. Haohe Liu (University of Surrey, now at Meta)

Mr. Yi Yuan (University of Surrey), Mr. Zhongkai Zhu

Dr. Julian D. Parker (Stability AI, now at Spotify), Mr. Anton Smirnov (Stability AI)

Dr. Yuancheng Wang (Internship at Meta, now at Bytedance Seed)

Prof. Qiuqiang Kong (CUHK), Prof. Yin Cao (Institute of Acoustics, CAS)

Prof. Wenwu Wang (University of Surrey), Prof. Mark D. Plumbley (University of Surrey, now at KCL)

...

Some materials in these slides are adapted from:

<https://github.com/haoheliu/haoheliu.github.io/tree/main/pdf>

Thank you!

- Personal website: <https://liuxubo717.github.io/>
- Email: liuxubo717@gmail.com