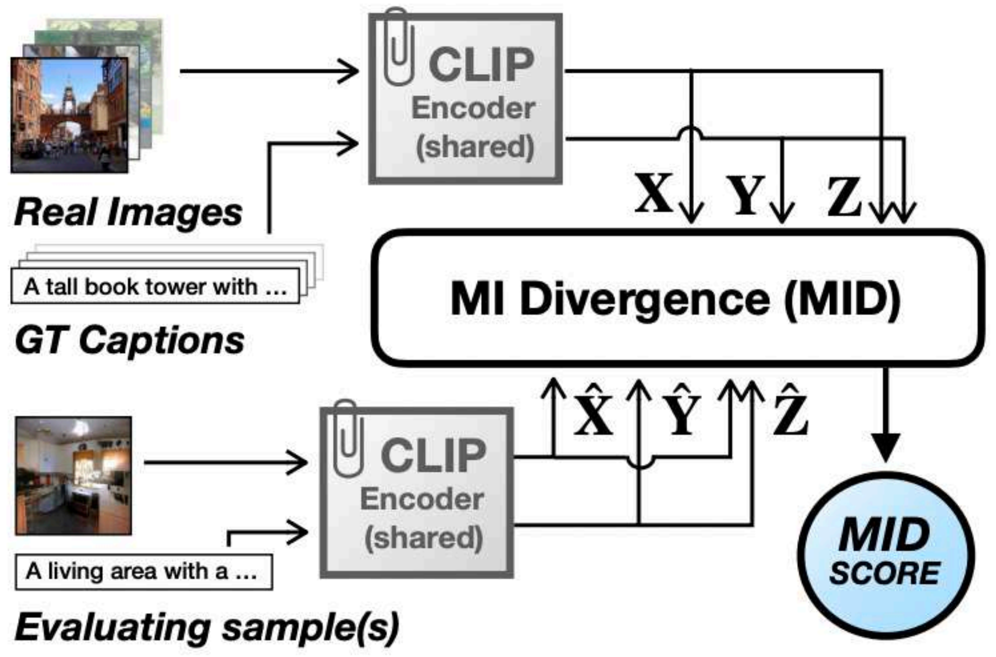
Audio Generation from Visual Contents

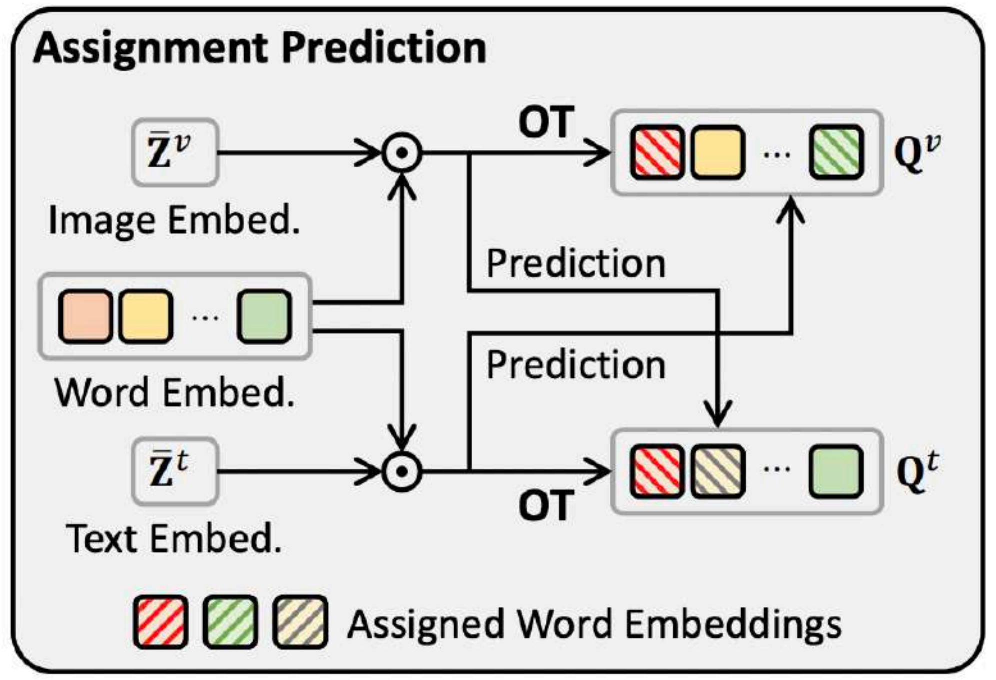
Jiyoung Lee
ML Research

NAVER AI LAB

Recent works



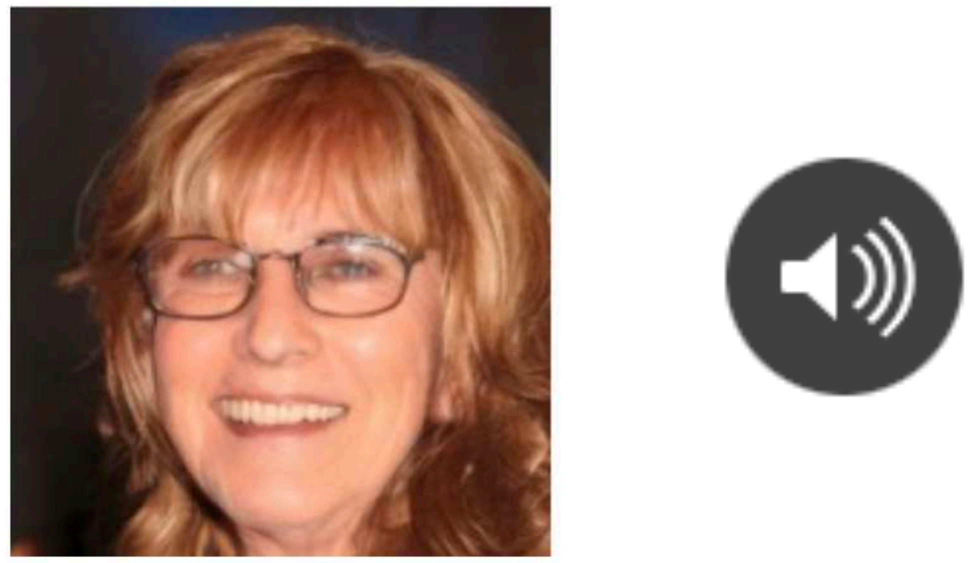
MID (NeurIPS'22)



VLAP (ICLR'24)

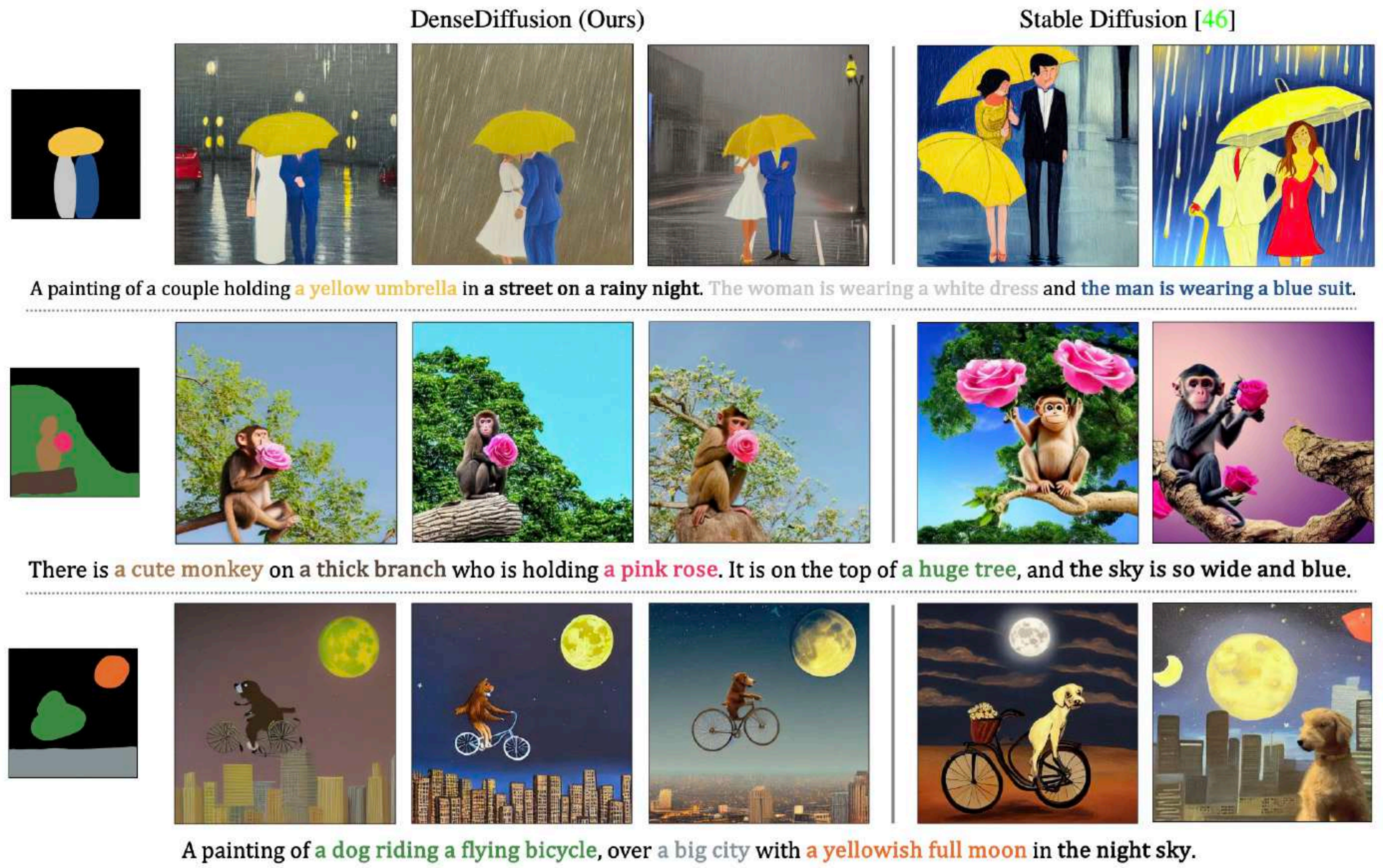
Imaginary Voice: Face-styled Diffusion Model for Text-to-Speech

Case 6 (Virtual face brought from Stable Diffusion)

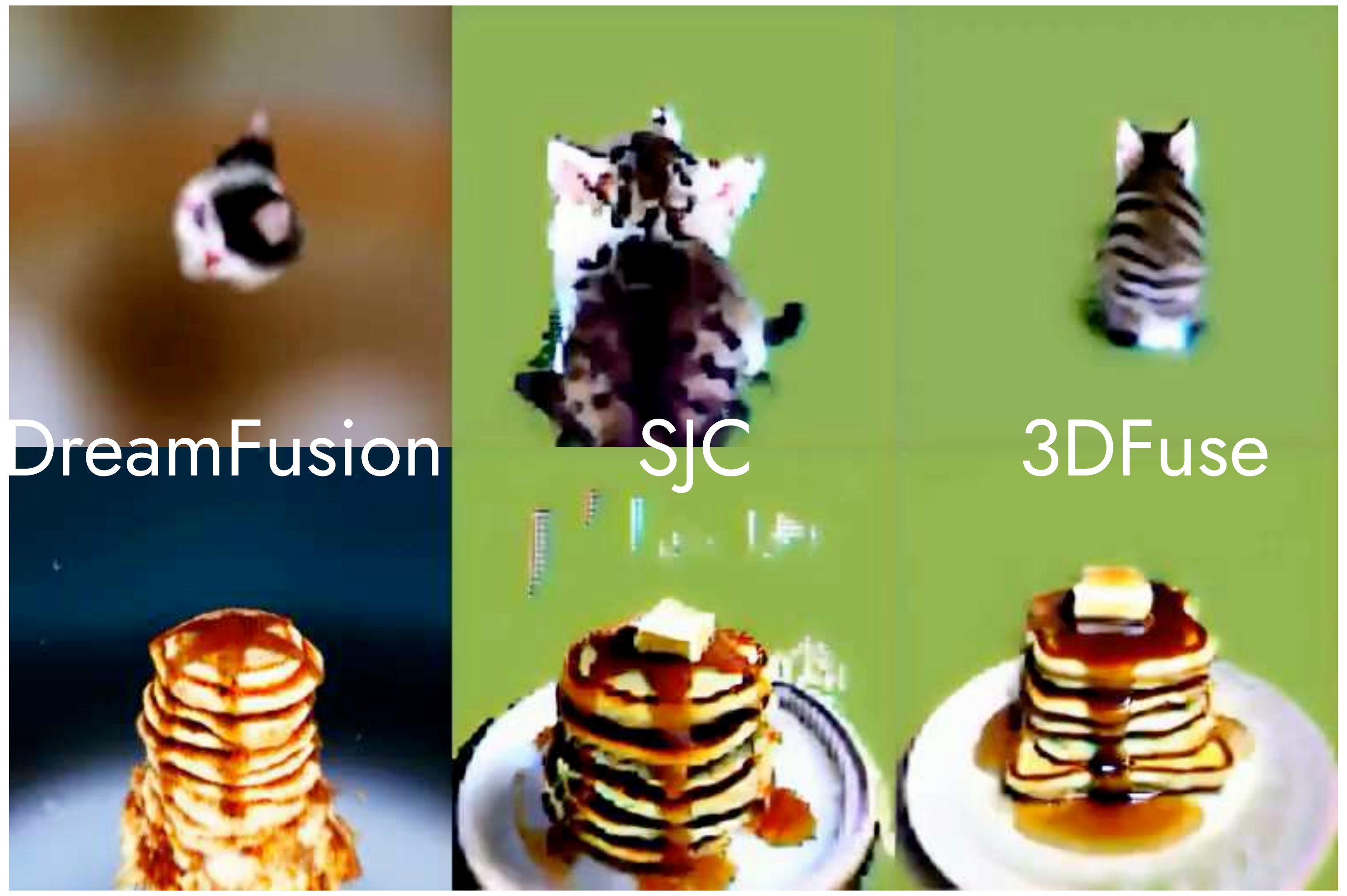


Text: And one or two men were allowed to mend clothes and make shoes. The rules made by the Secretary of State were hung up in conspicuous parts of the prison.

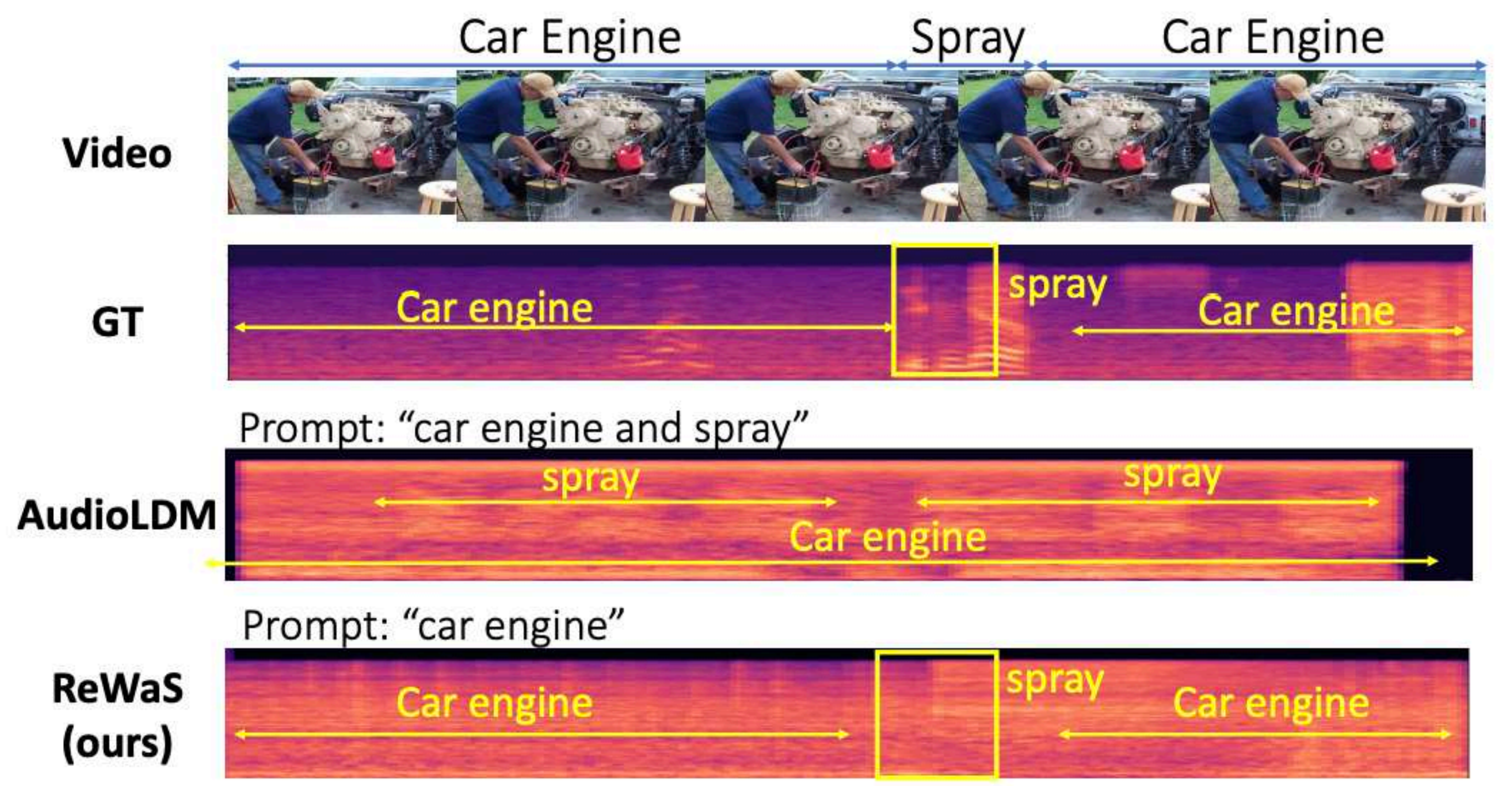
Face-TTS (ICASSP'23)



DenseDiffusion (ICCV'23)



3DFuse (ICLR'24)



ReWaS (NeurIPSW'24)

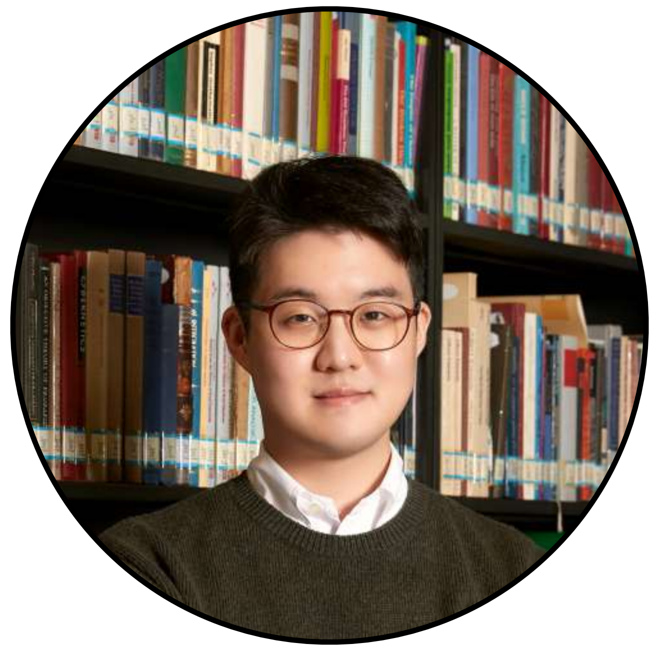
Today contents

- Speech generation from face image and text
  - Imaginary Voice: Face-styled Diffusion Model for Text-to-Speech (Face-TTS)
- Sound generation from video and text
  - Read, Watch, and Scream! Sound Generation from Text and Video (ReWaS)

Speech Generation from Face and Text

Imaginary Voice: Face-styled Diffusion Model for Text-to-Speech

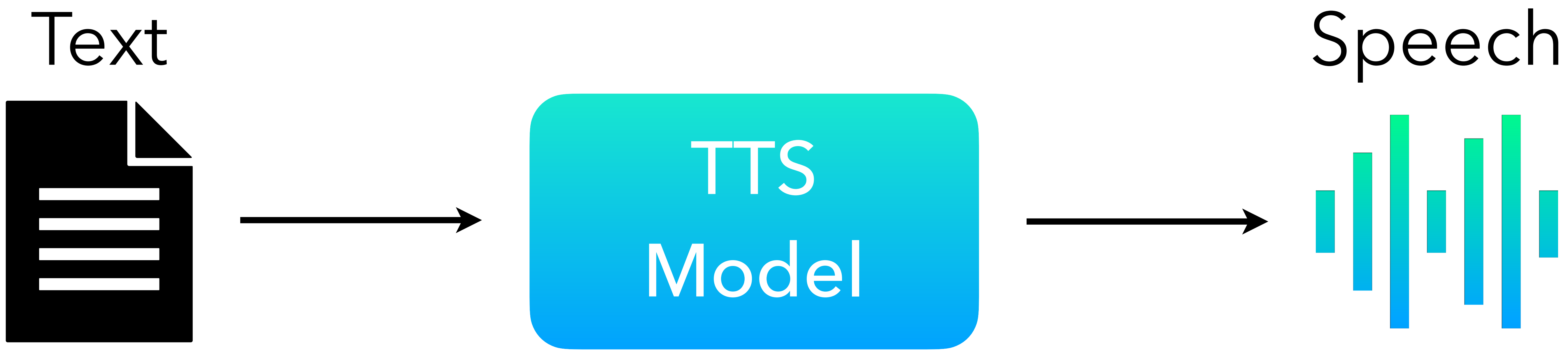
ICASSP 2023

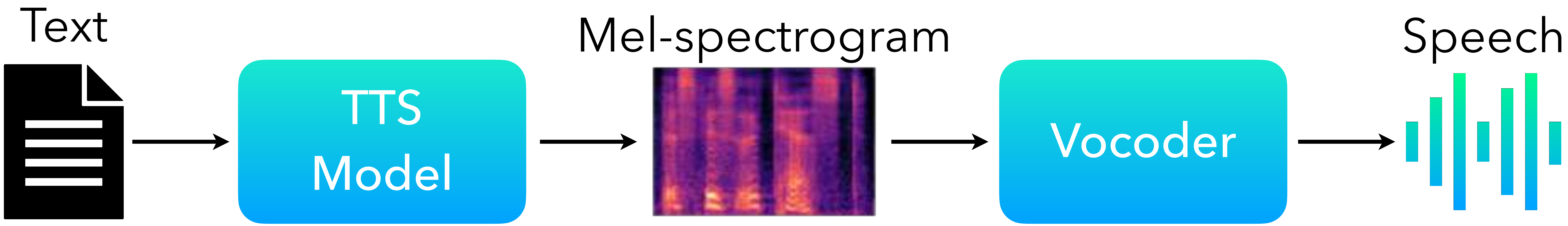

Jiyoung Lee Joon Son Chung Soo-Whan Chung

https://facetts.github.io/

Text-to-Speech

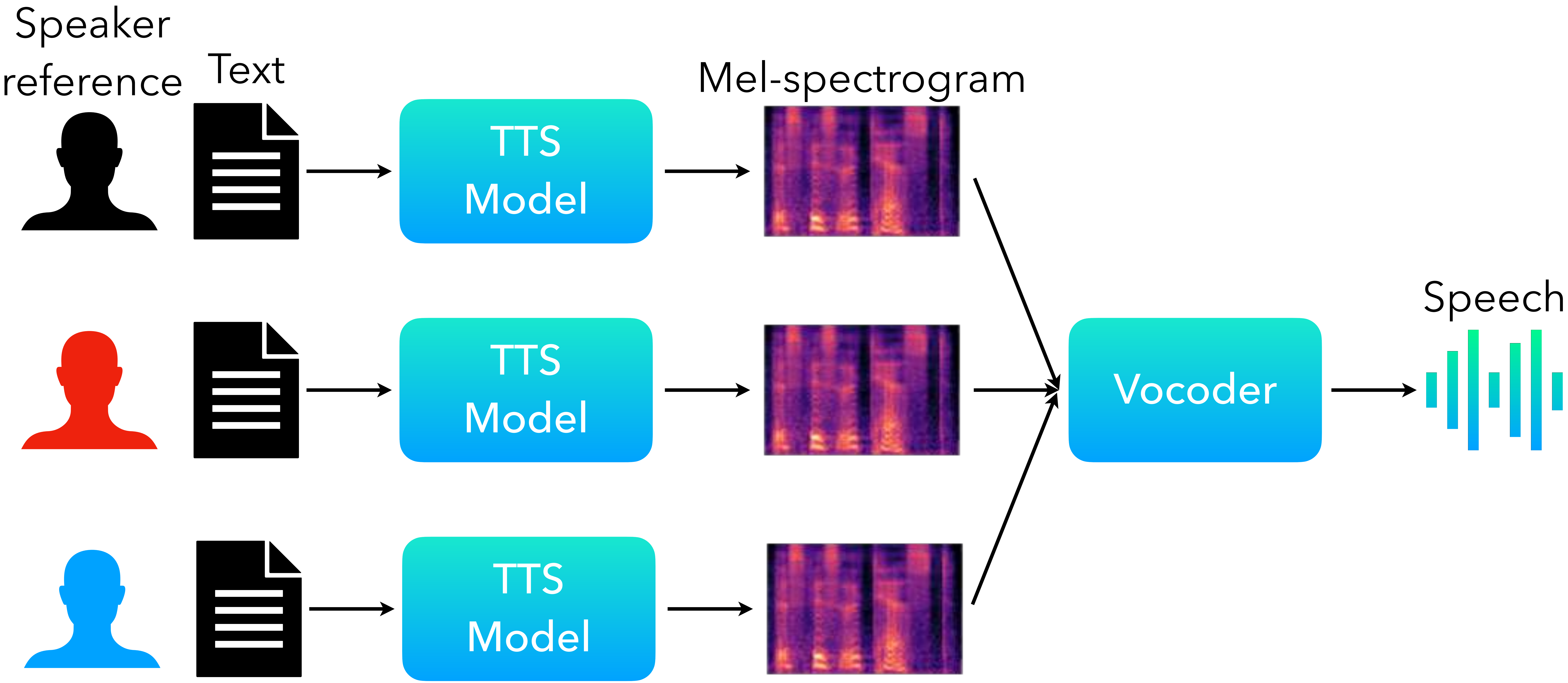


Text-to-Speech

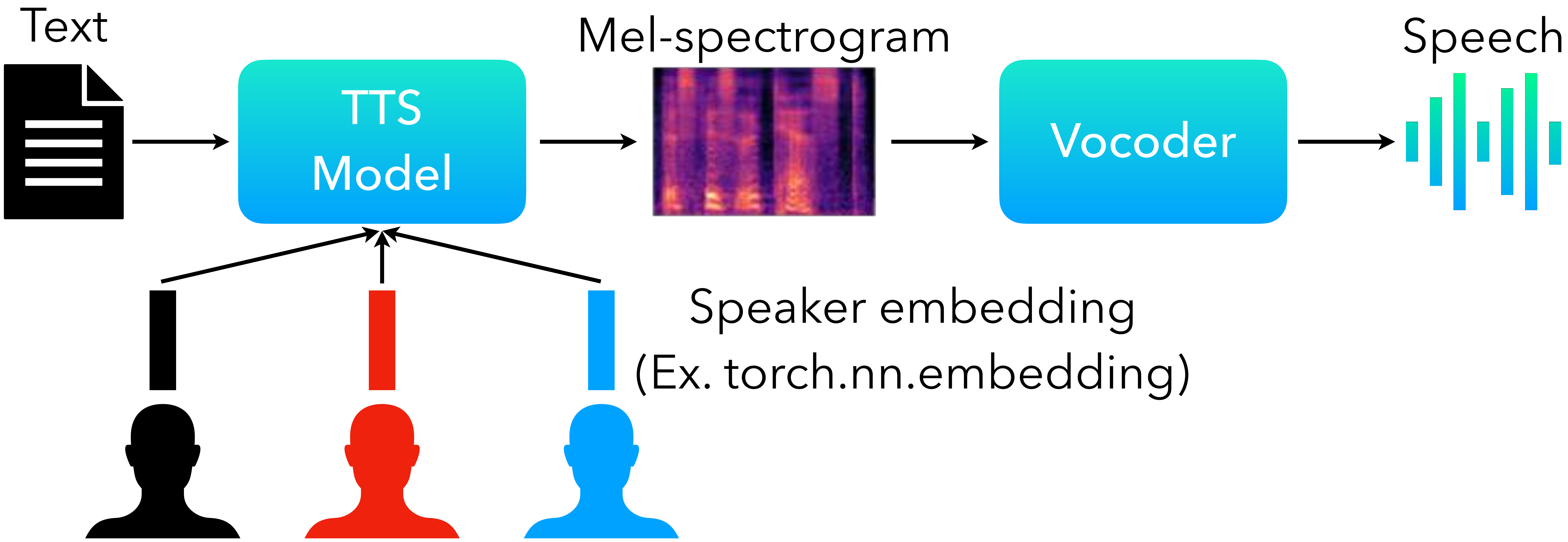


WaveGlow (R. Prenger et al., 2019)
HiFi-GAN (J. Kong et al., 2020)
BigVGAN (S. Lee et al., 2022)

Multi-speaker Text-to-Speech



Multi-speaker Text-to-Speech

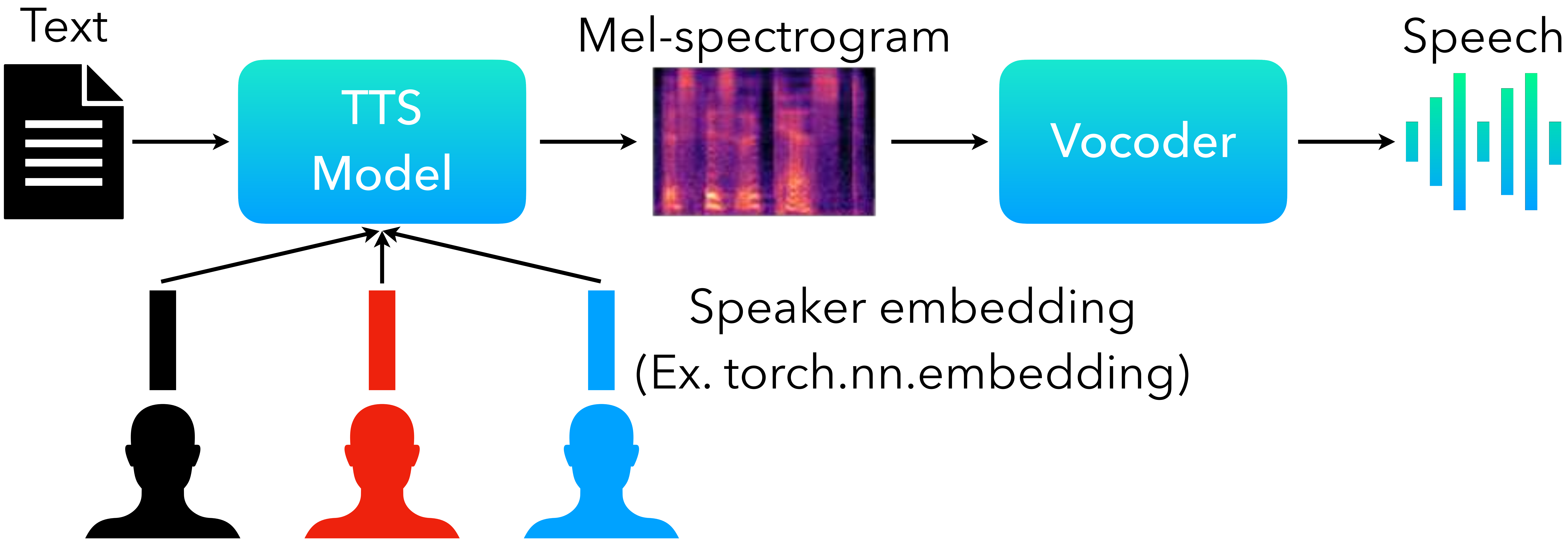


Multi-speaker Text-to-Speech

😭 *Clear voice recordings of each speaker are much needed*

😭 *We cannot create style variations for each speaker*

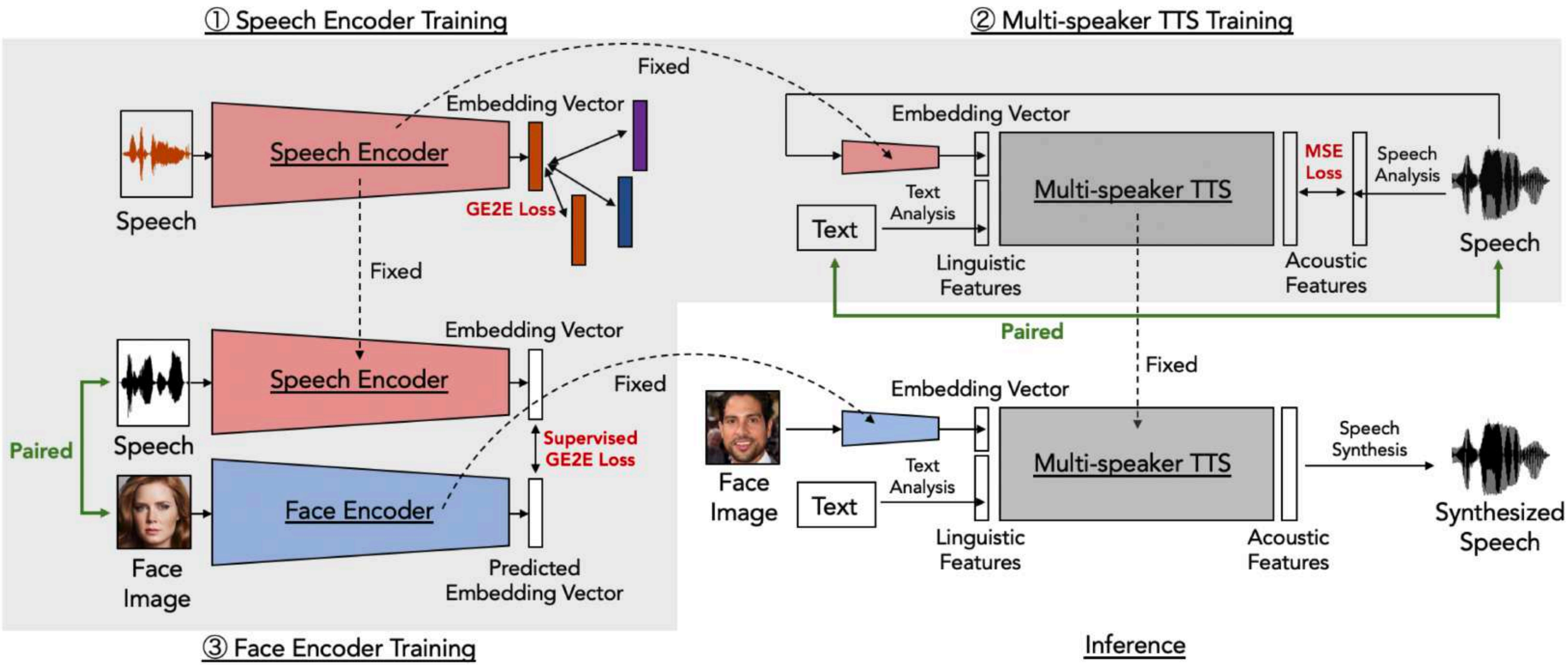
😭 *We cannot generate **new** speaker's speech without voice (zero-shot setting)*



“What if face images can be used
for enrollment instead of speech signals?”

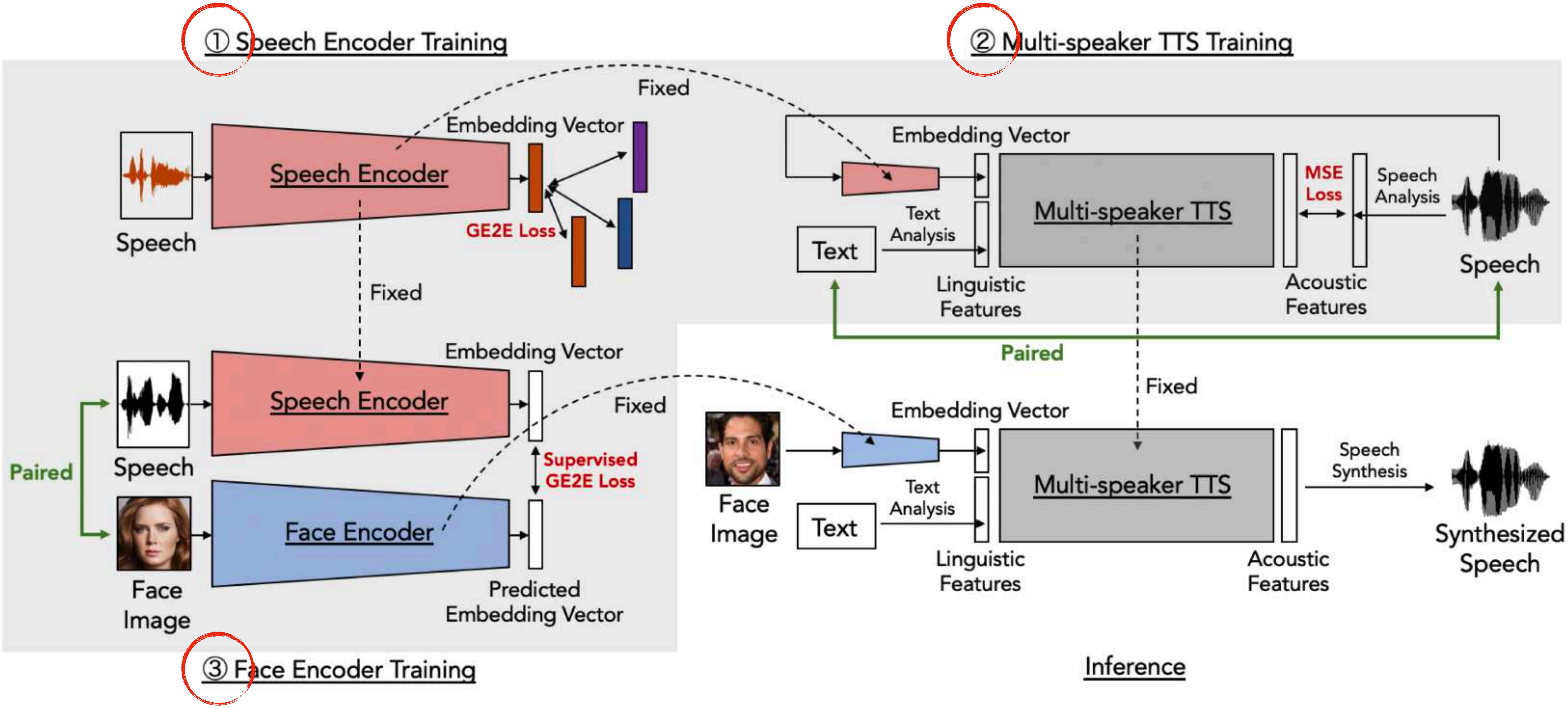
Previous work

Face2Speech (Goto et al., 2020)



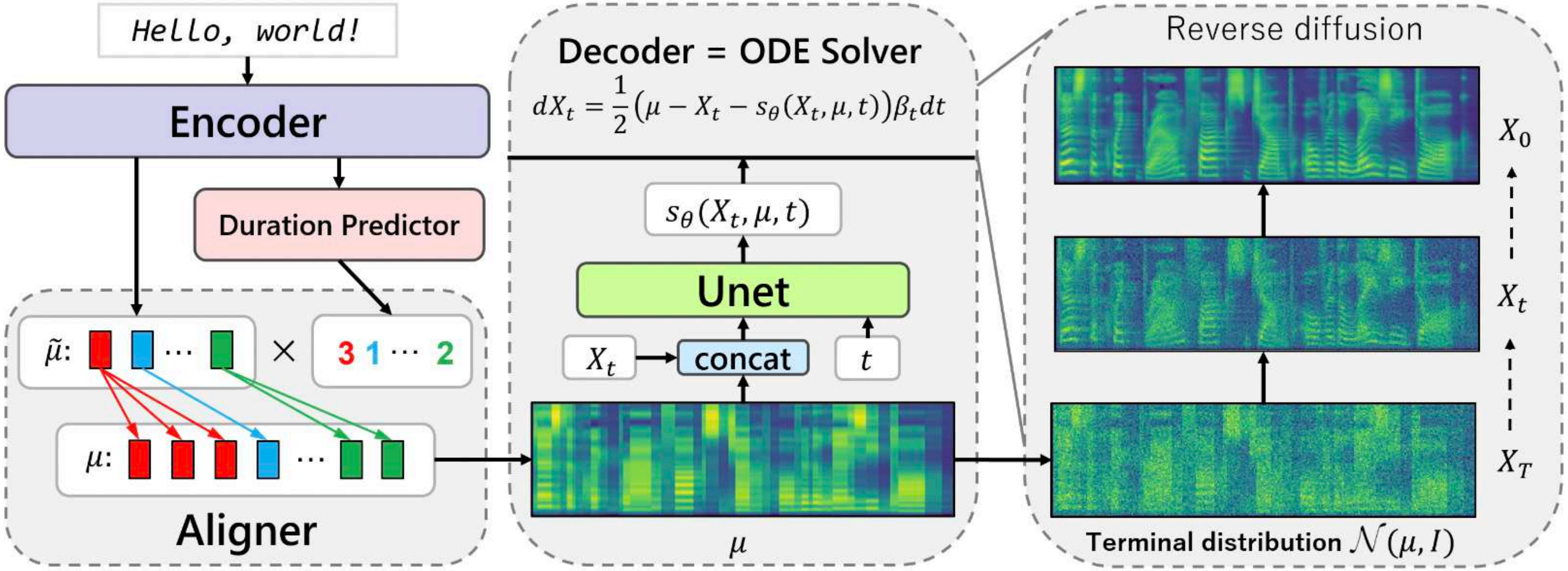
Previous work

Face2Speech (Goto et al., 2020)



Previous work

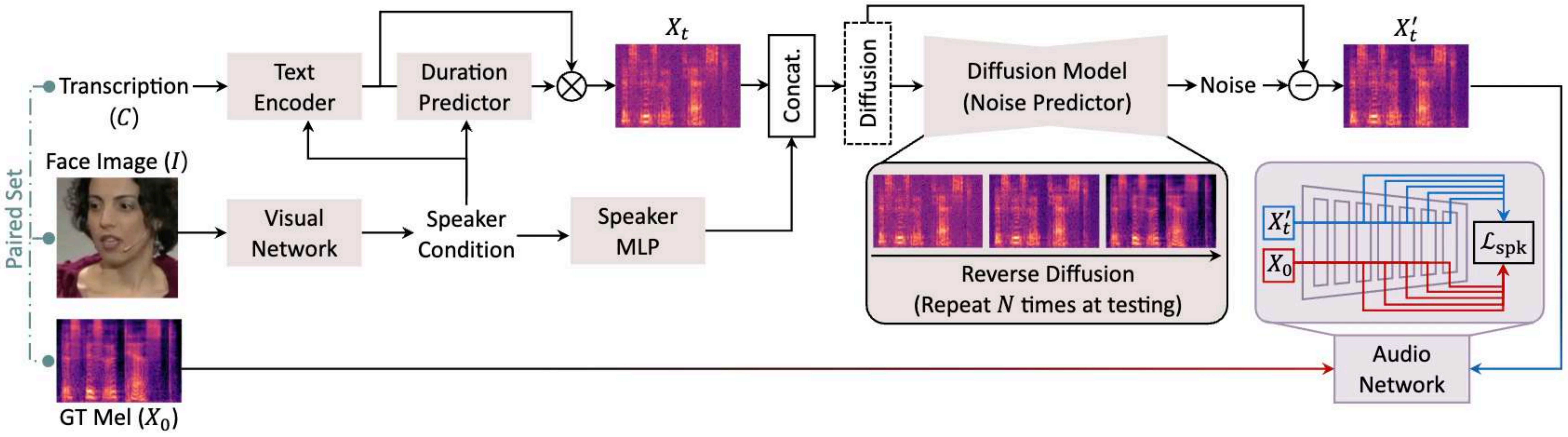
Grad-TTS (Popov et al., 2021)



- Encoder+duration predictor: Encoding text (phonemes) to frame-level features μ and duration d
- Diffusion model (Unet): refine features to generate Mel-spectrogram

Overall framework

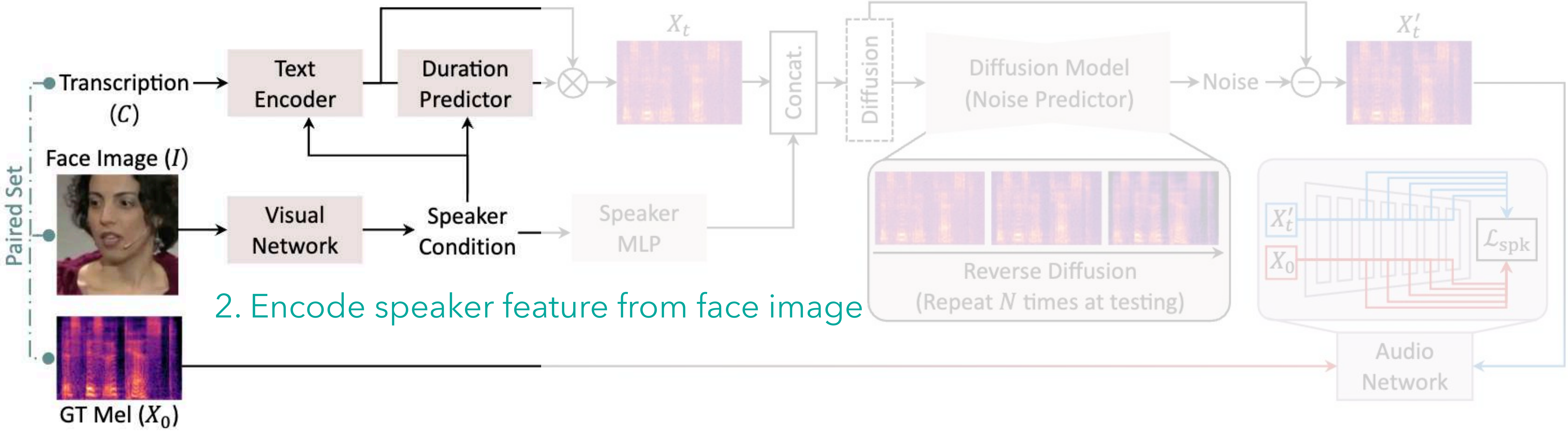
Face-TTS (proposed method)



Overall framework

Face-TTS (proposed method)

1. Encode text feature and predict mean and variance for duration

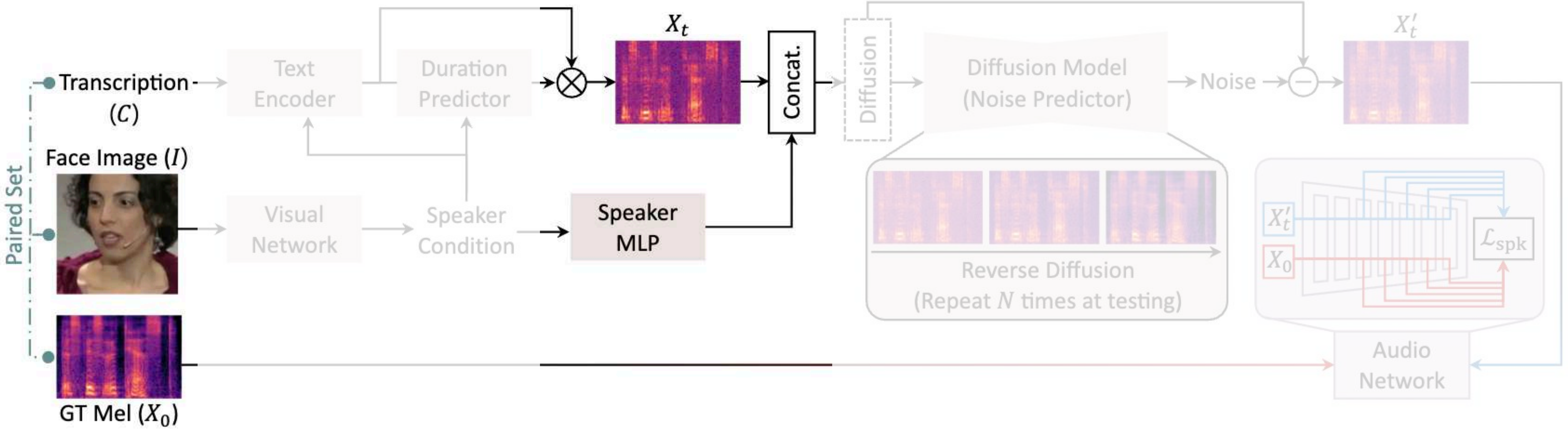


2. Encode speaker feature from face image

Overall framework

Face-TTS (proposed method)

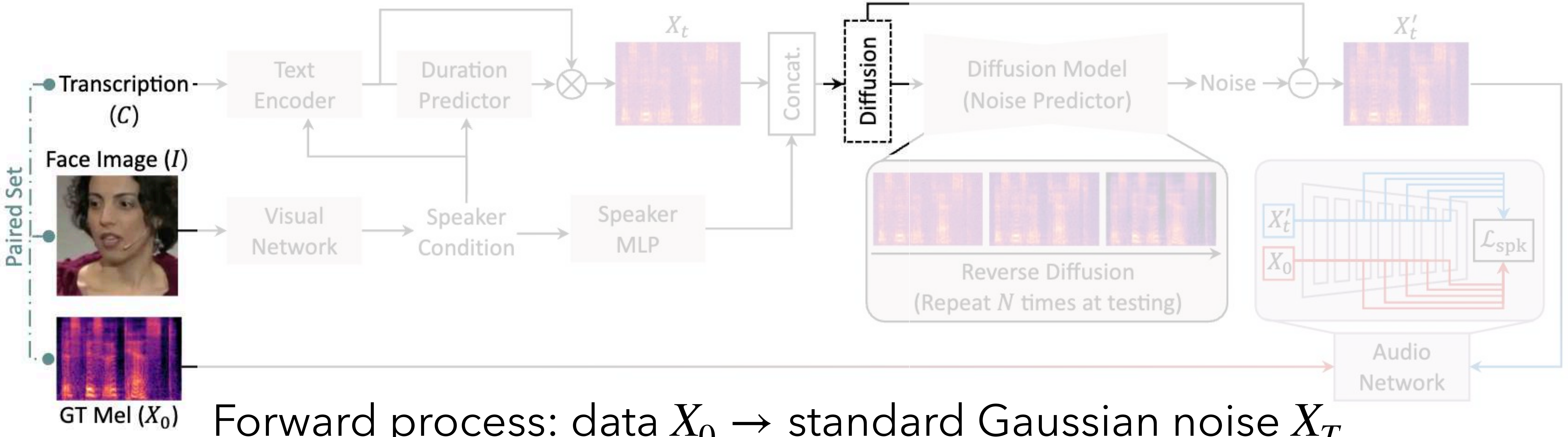
3. Concatenate speaker feature to acoustic feature (X_t)



Overall framework

Face-TTS (proposed method)

4. [**Training**] add noise (diffusion process)



Forward process: data X_0 → standard Gaussian noise X_T

$$dX_t = -\frac{1}{2}X_t\beta_t dt + \sqrt{\beta_t}dW_t$$

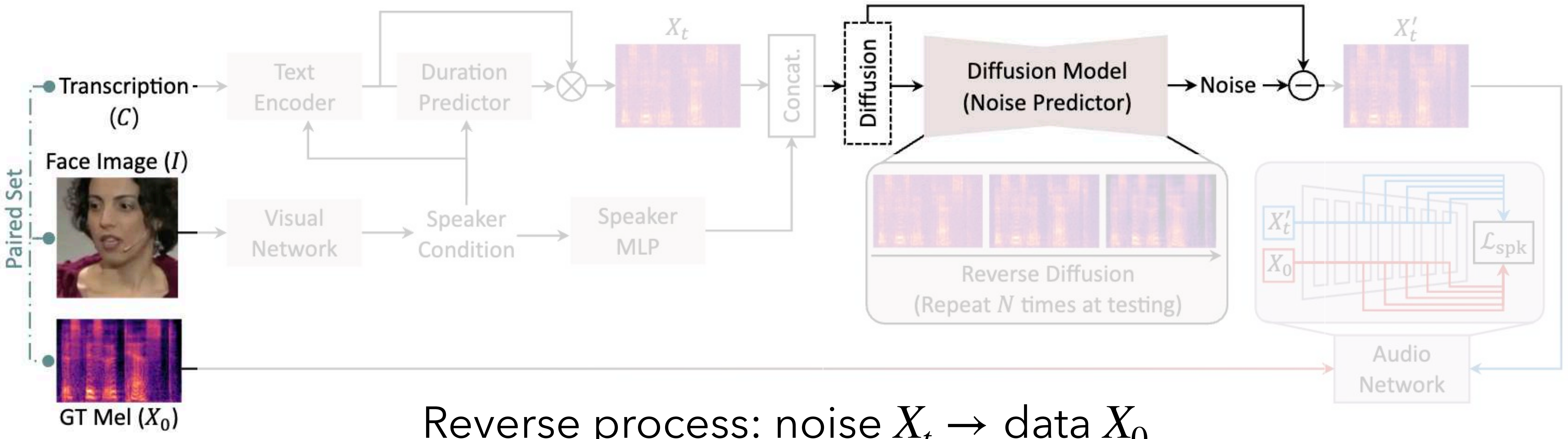
Predefined noise schedule $\beta_t = \beta_0 + (\beta_T - \beta_0)t$

Wiener process

Overall framework

Face-TTS (proposed method)

5. [**Training**] Predict noise by DDPM



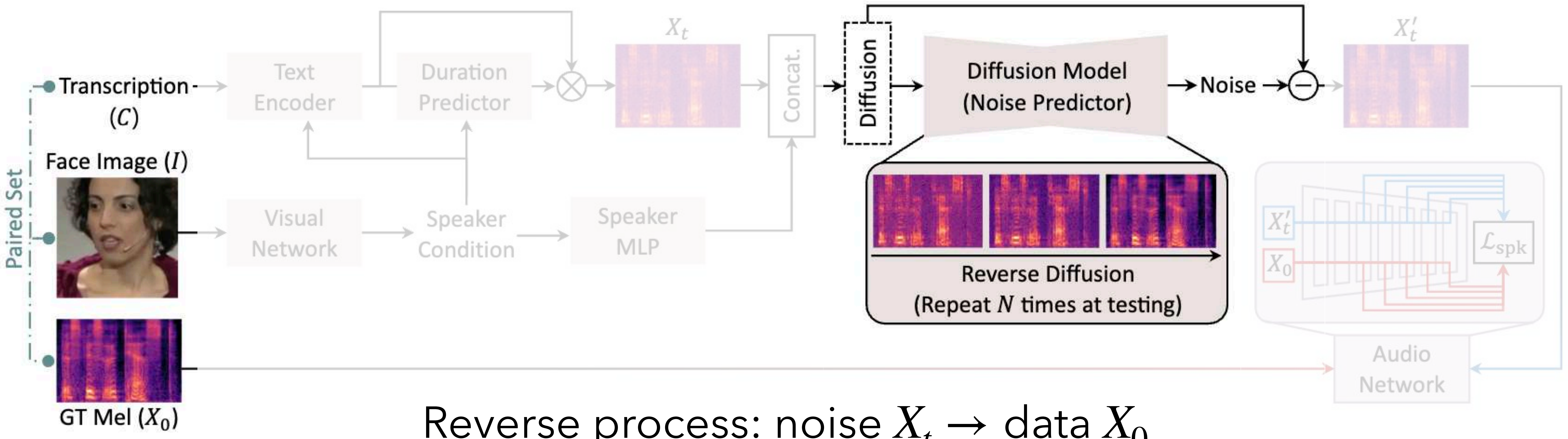
Reverse process: noise $X_t \rightarrow$ data X_0

$$dX_t = (-\frac{1}{2}X_t - \underbrace{\nabla_{X_t}\log P_t(X_t)}_{\text{Score, estimated by the network (UNet)}})\beta_t dt + \sqrt{\beta_t}d\widetilde{W}_t$$

Overall framework

Face-TTS (proposed method)

6. [**Inference**] Stochastically predict noise (reverse diffusion process)



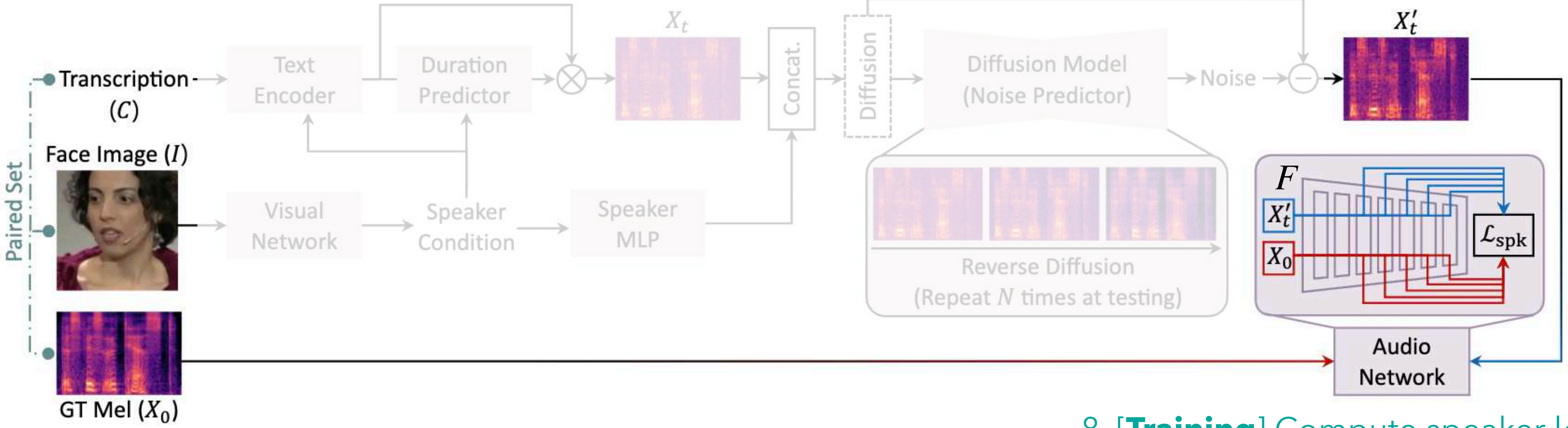
Reverse process: noise $X_t \rightarrow$ data X_0

$$dX_t = (-\frac{1}{2}X_t - \underbrace{\nabla_{X_t}\log P_t(X_t)}_{\text{Score, estimated by the network (UNet)}})\beta_t dt + \sqrt{\beta_t}d\widetilde{W}_t$$

Overall framework

Face-TTS (proposed method)

7. [**Training**] Encode audio feature from predicted Mel-spec.



8. [**Training**] Compute speaker loss

\+ diffusion loss, duration loss, prior loss

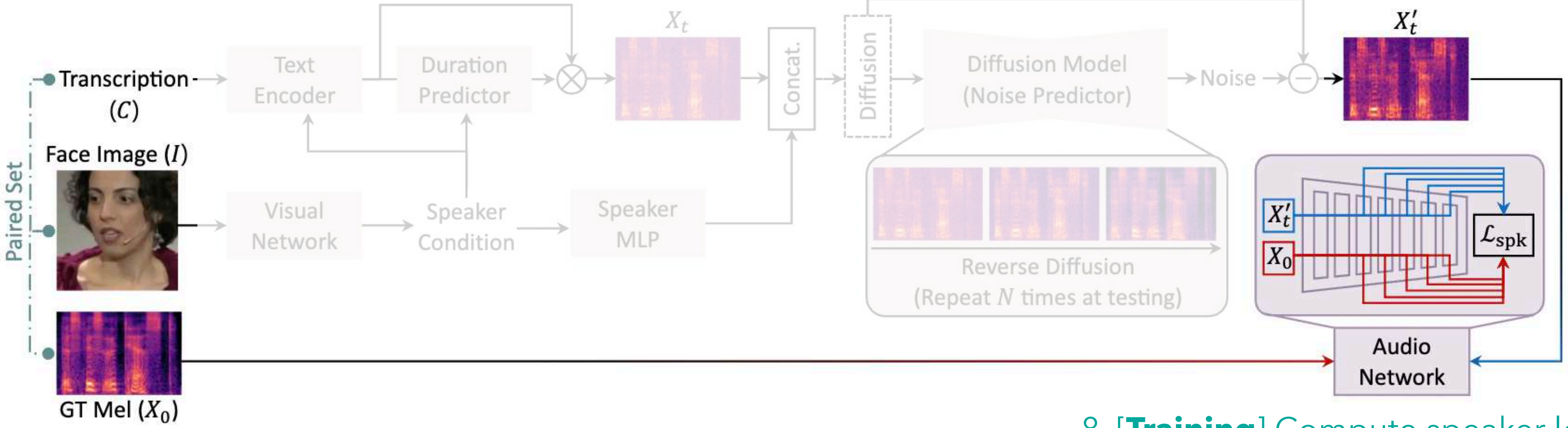
Speaker loss: $L = \sum_B |F_b(X_0) - F_b(X'_t)|$,

b indicates layers in audio network F

Overall framework

Face-TTS (proposed method)

7. [**Training**] Encode audio feature from predicted Mel-spec.



8. [**Training**] Compute speaker loss

+ diffusion loss, duration loss, prior loss

Prior loss: encoder output is regarded as normal distribution
Duration loss: MSE in logarithmic domain
Diffusion loss: the expectation of weighted losses associated with estimating gradients of log-density of noisy data at different times

Training

Data

- Conditions
  - Paired face image, audio, text descriptions
  - Enough amount to train the model
  - Open license

LRS3 (Afouras et al., 2018)

- Speaker ID
- Talking video (+audio)
- Text description
- ~150k clips

Training

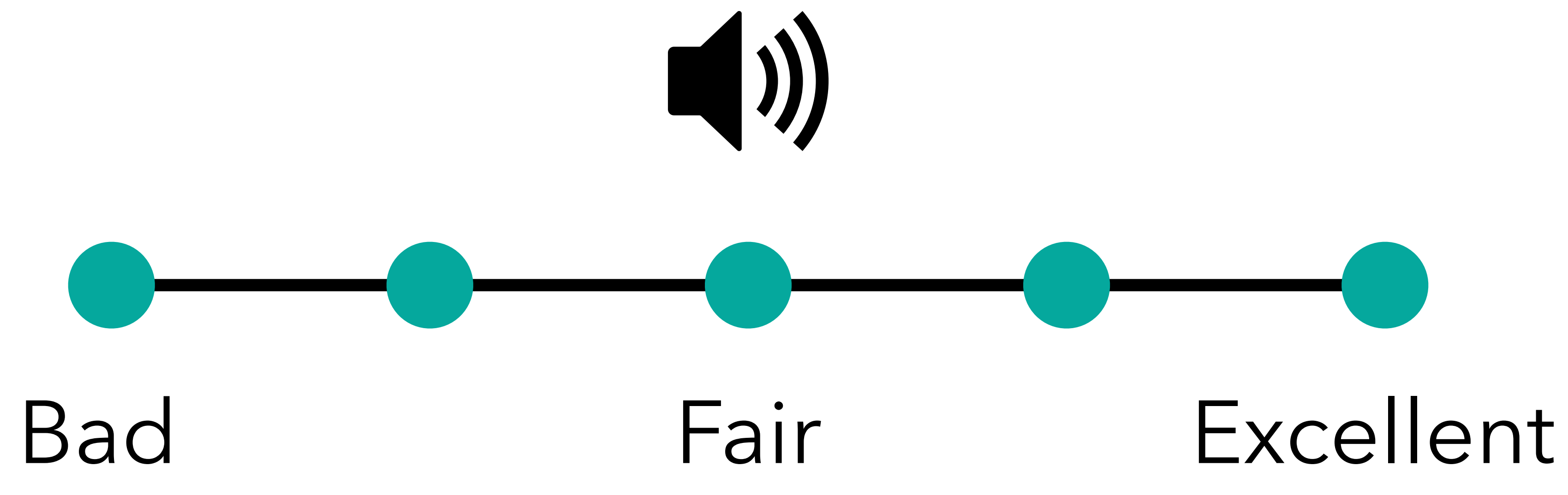
Traditional TTS datasets vs LRS3

	LJSpeech (Ito, 2017)	LibriTTS (Zen et al., 2019)	**LRS3** (Afouras et al., 2018)
Clean background?	O (Audiobook)	O (Audiobook)	X (In the wild)
With video?	X	X	O
# hours	24h	586h	407h
Multi speaker?	X	O	O
Length of clip	< 10s	10s <	Pretrain set: 12s Train-val: 3s

Evaluation

Mean opinion score (MOS) test

How about the quality?



Method	**Spk. ID**	**5-scale MOS**
Ground Truth	-	4.865±.001
Mel.+HiFi-GAN [19] (Upper bound)	-	4.653±.035
Grad-TTS [11]† (Seen)	Embed	3.718±.318
Face-TTS (Seen)	Audio	3.547±.331
Face-TTS (Seen)	Face	3.706±.154
Face-TTS (Unseen)	Audio	3.218±.249
Face-TTS (Unseen)	Face	3.282±.219

Evaluation

AB forced matching test

Which face is better to match?

Query

Answer :

A

B

Evaluation

ABX preference test

Which speech is better to match?

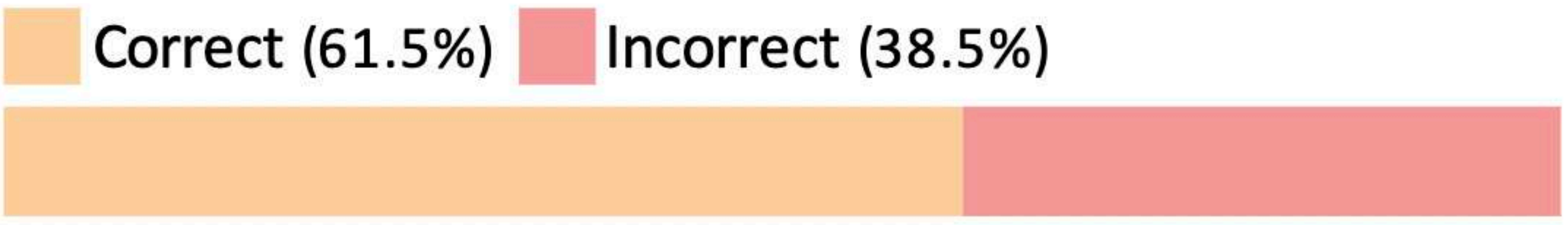
Query

Answer :

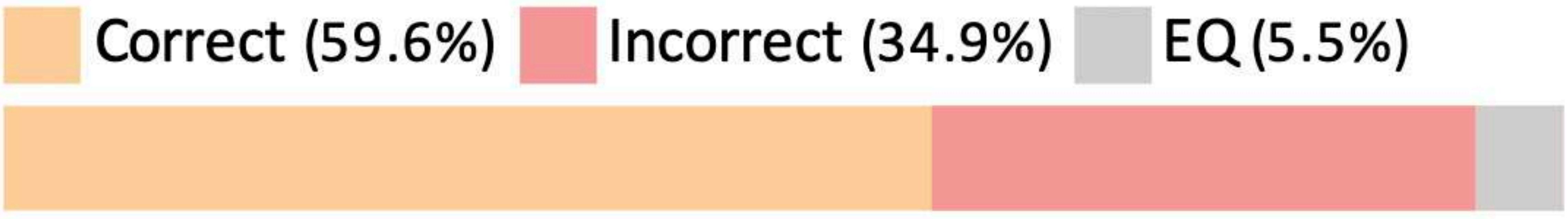
		I don't know
A	B	X

Evaluation

AB/ABX test



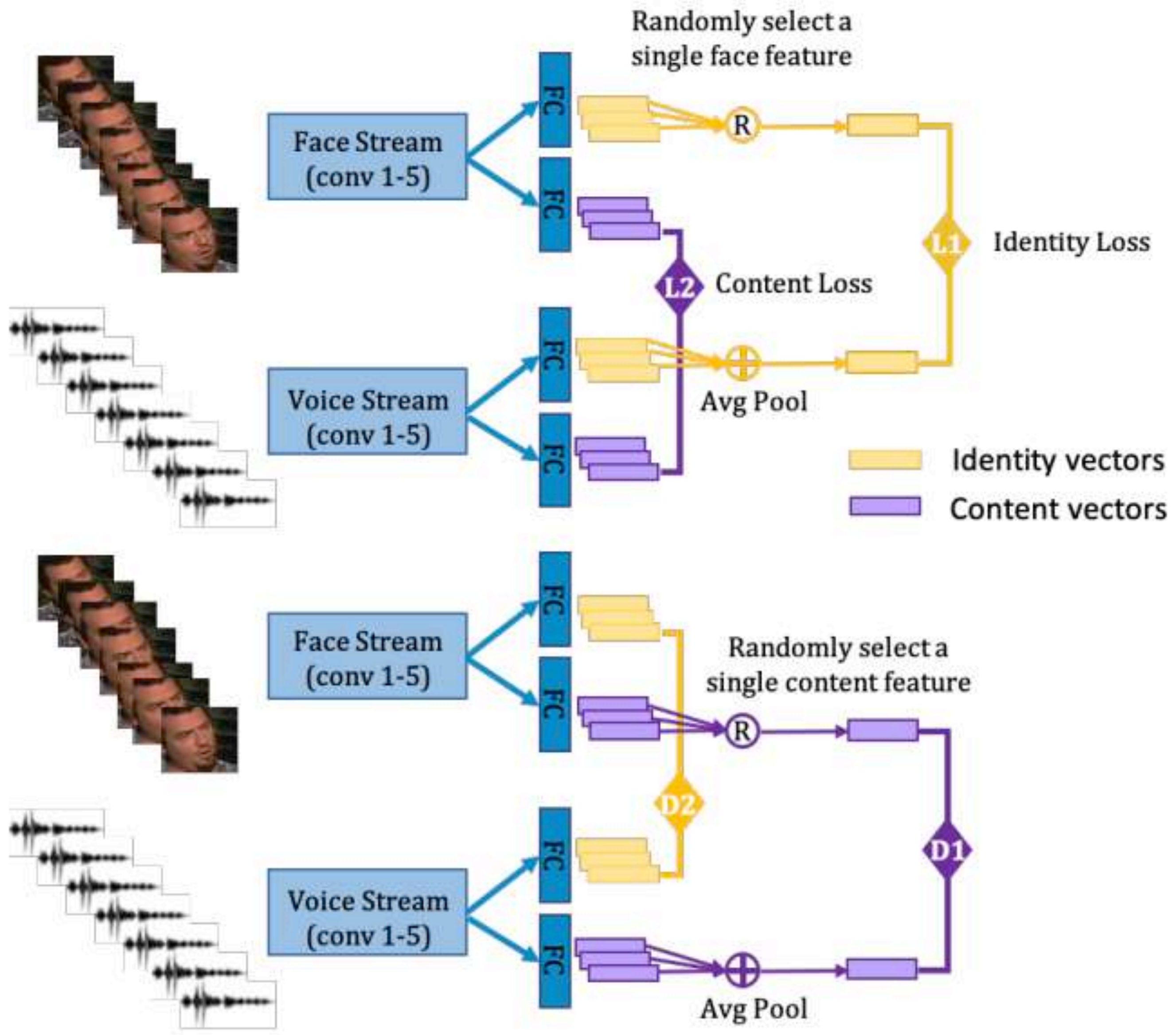
(a) AB test



(b) ABX test

Evaluation

Speaker identification matching accuracy



Cross-modal biometric embeddings
(Nagrani et al., 2020)
(Chung et al., 2020)

Method	Spk. ID	Acc. (%)
Mel.+HiFi-GAN [19] (Upper bound)	-	48.6
Grad-TTS [11]	Embed	19.4
FACE-TTS (w/o. $\mathcal{L}_{\text{spk}}$)	Face	35.4
FACE-TTS	Face	38.0

*Random: 20%

Demo

Unseen speakers (from face)

Text: '*The employees raced the elevators to the first floor. Givens saw Oswald standing at the gate on the fifth floor as the elevator went by.*'

Demo

Unseen speakers (from face)

Text: *'Four point eight to five point six seconds if the second shot missed.'*

Demo

Virtual speakers taken from LDM (Rombach et al., 2022)

More demo in https://facetts.github.io/

Virtual face

Text: '*The preference given to the Pentonville system destroyed all hopes of a complete reformation of Newgate.*'

Text: '*And one or two men were allowed to mend clothes and make shoes. The rules made by the Secretary of State were hung up in conspicuous parts of the prison.*'

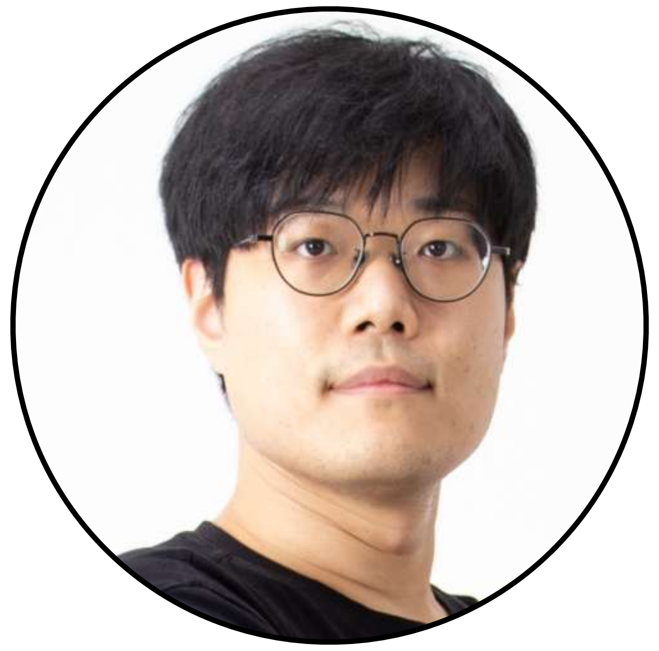
Audio Generation from Video and Text

Read, Watch, and Scream!
Sound Generation from Text and Video

NeurIPS 2024 Workshop on Video-Language Models


Yujin Jeong


Yunji Kim

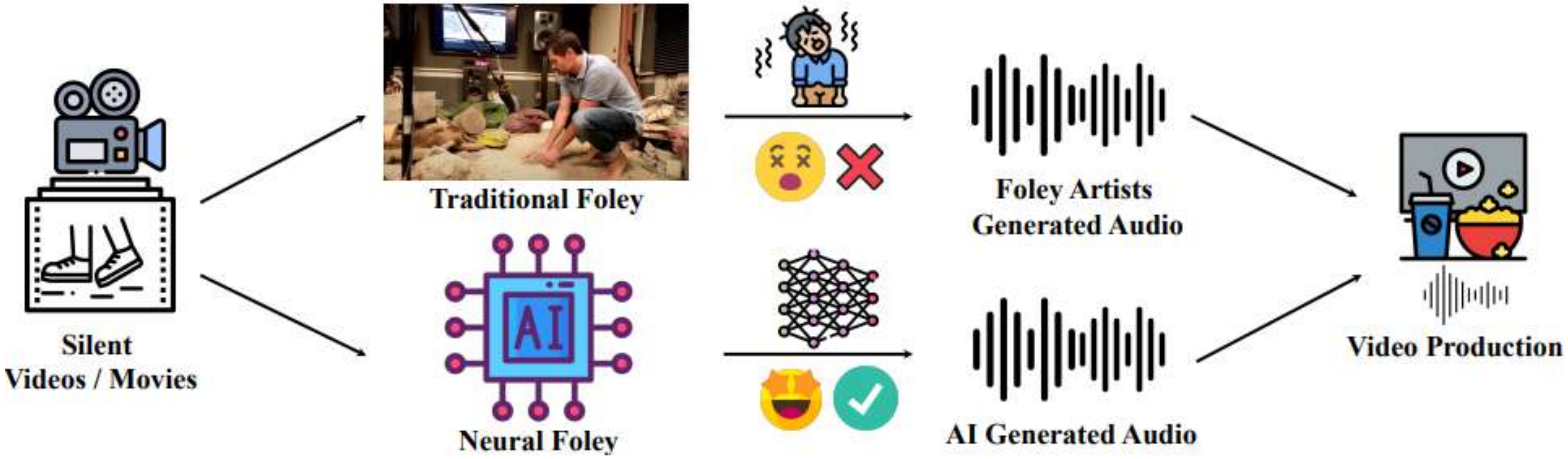

Sanghyuk Chun


Jiyoung Lee

https://naver-ai.github.io/rewas/

Sound generation from video

- Foley?
- Foley is the reproduction of sound effects that are added to films, and videos in post-production to enhance audio quality
- Existing video-to-sound generation methods have been roughly divided into two groups (foley or general sound generation)

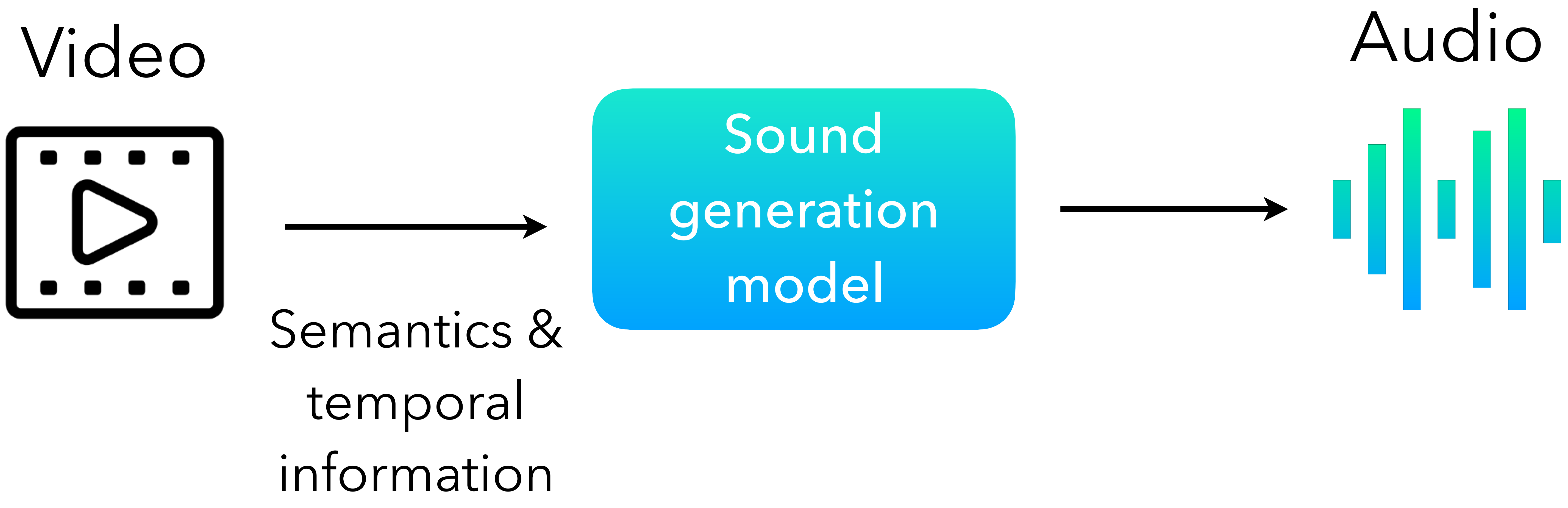


“Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models” NeurIPS (2023)

Foley artist
(Credit: Wikipedia)

Sound generation from video

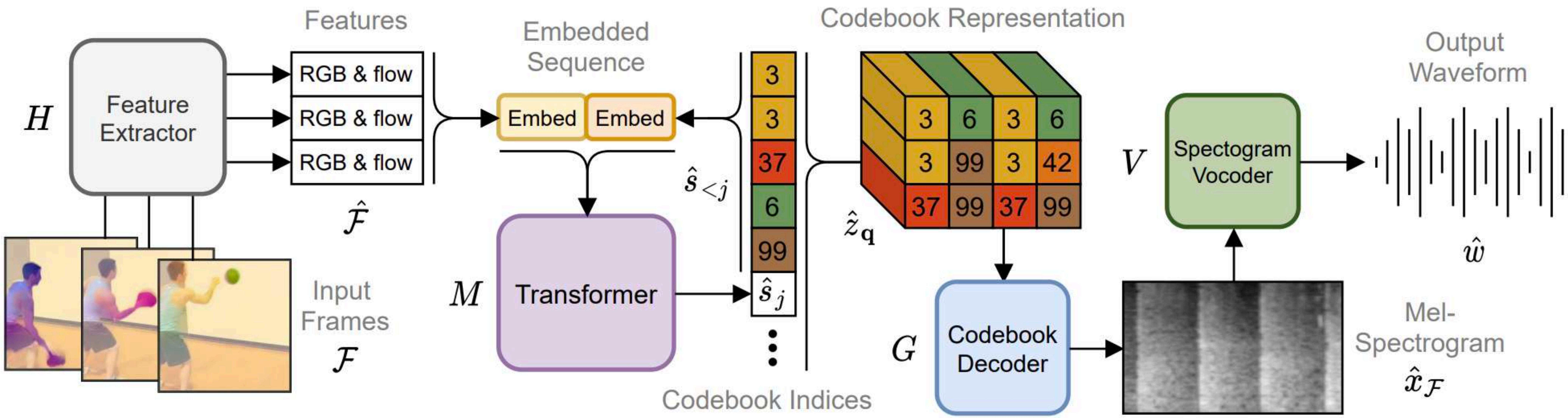
,also called video-to-sound generation



Previous work

SpecVQGAN

- SpecVQGAN (BMVC'21) takes RGB and optical flow of videos, and uses a transformer to generate indices of a spectrogram VQVAE (vector quantized variational autoencoder) autoregressively
- Proposes a set of metrics for automatic evaluation of visually-guided spectrogram synthesis



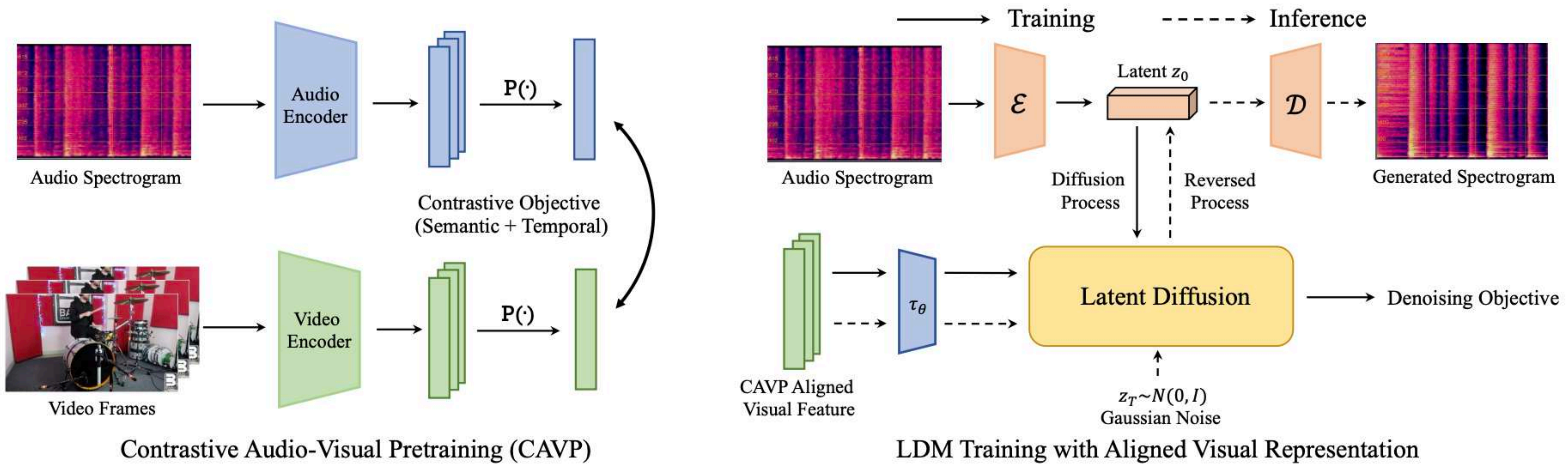
"Taming Visually Guided Sound Generation" BMVC (2021)

Previous work

Diff-Foley

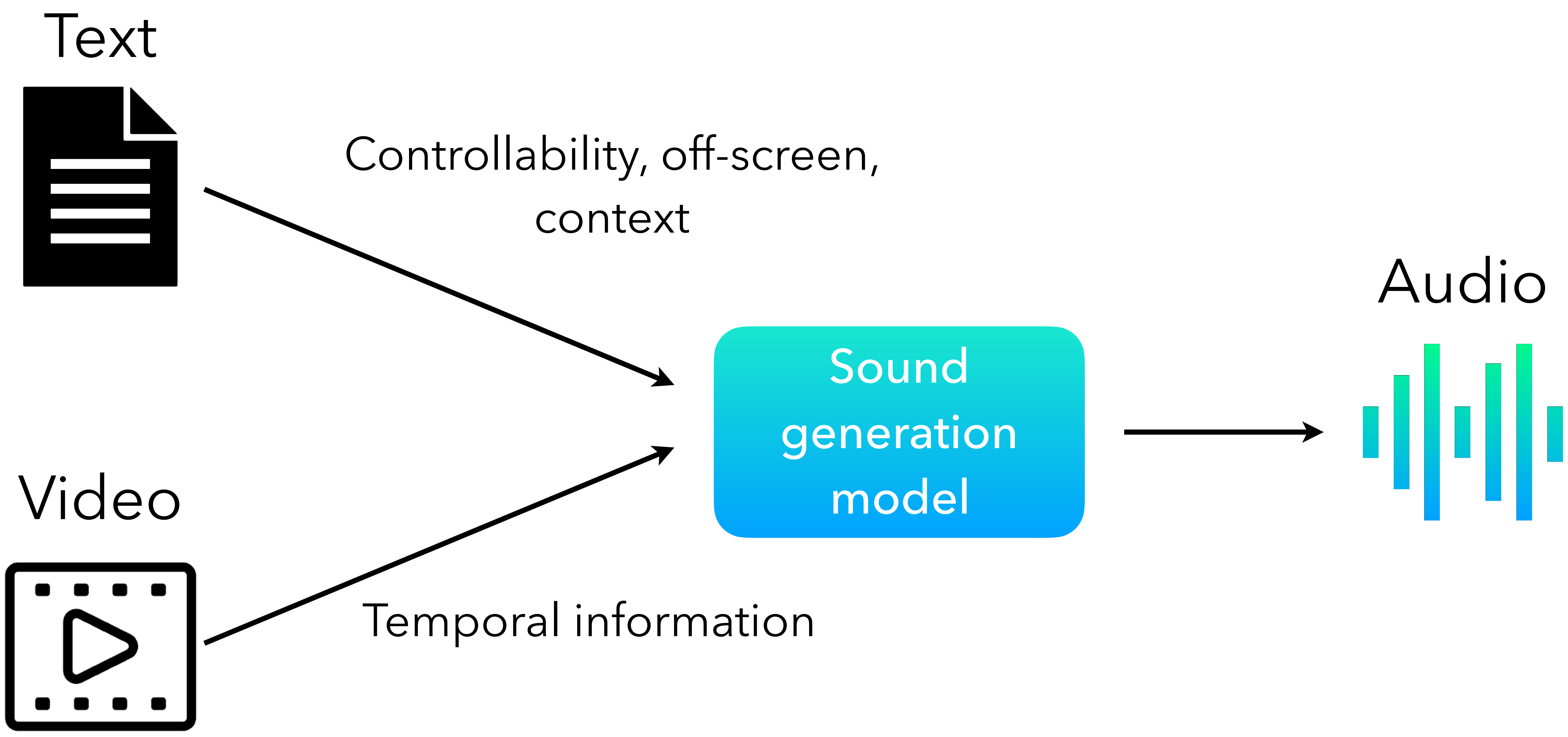
- Stage 1: contrastive audio-visual pretraining
- Stage 2: LDM training with an aligned visual representation

😭 Require a large training cost and time for stages 1 and 2



"Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models" NeurIPS (2023)

ReWaS (our work)



“Read, watch, and scream! Sound generation from text and video” NeurIPS Workshop (2024)

ReWaS (our work)

example

Text

“In an ornate, historical hall, a massive tidal wave peaks and begins to crash. Two surfers, seizing the moment, skillfully navigate the face of the wave”

Video

Sound generation model

Audio aligned to video-and-text

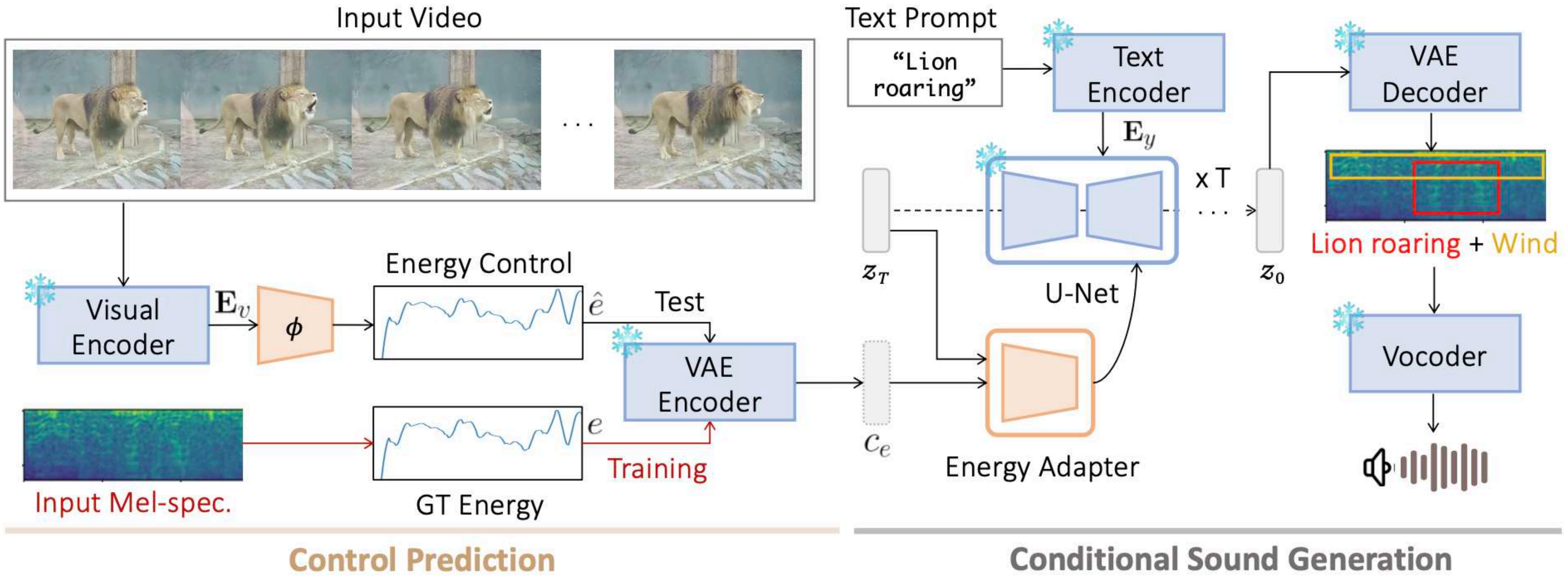
“Read, watch, and scream! Sound generation from text and video” NeurIPS Workshop (2024), video from Kling

Framework

ReWaS

😁 Reduce training cost data by leveraging a pretrained audio generation model (i.e., AudioLDM)

😁 Energy control gives an intermediate bridge to map visual content into the audio model

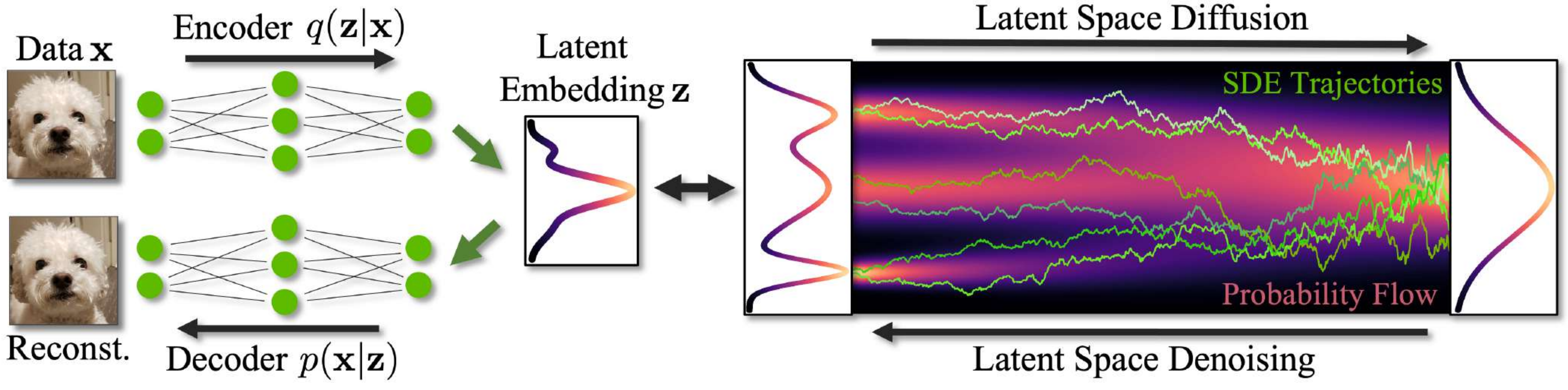


"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024)

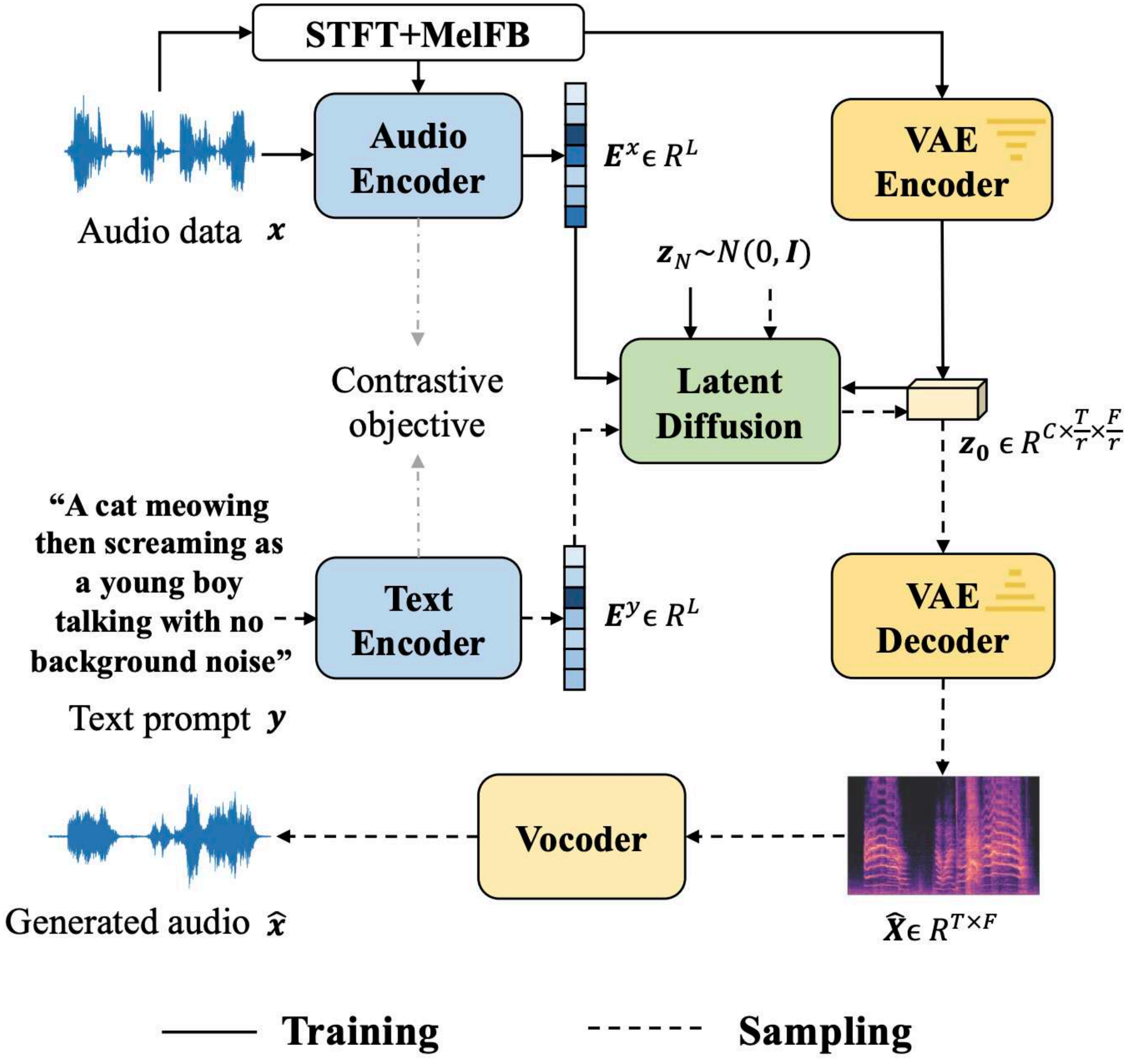
Base audio generation model

AudioLDM

- Latent diffusion model for audio generation
- Utilize CLAP (contrastive language-audio learning, similar to CLIP [2]) embeddings to enable text-to-audio generation without using language-audio pairs to train LDMs



LDM (latent diffusion model)
(Credit: Karsten et al., NeurIPS'2023 tutorial)

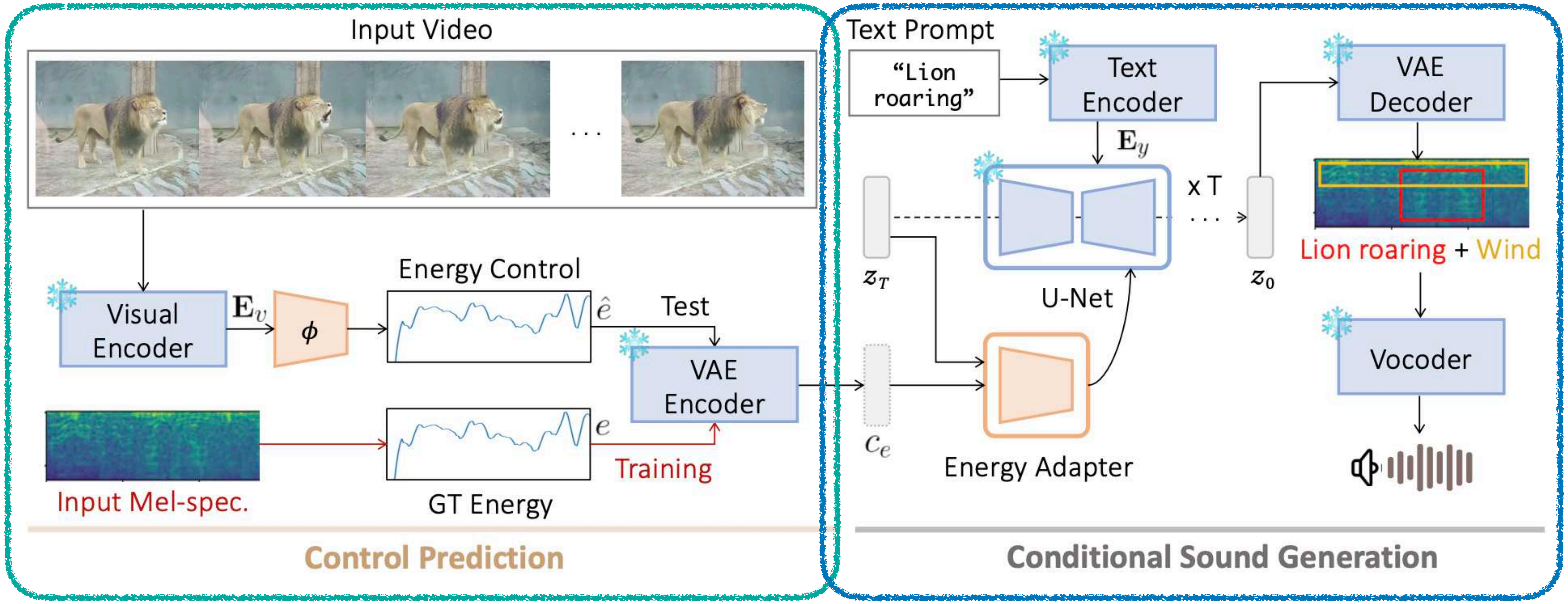


[1] "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models" ICML (2023)
[2] "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation" ICASSP (2023)

Training for ReWaS

ReWaS

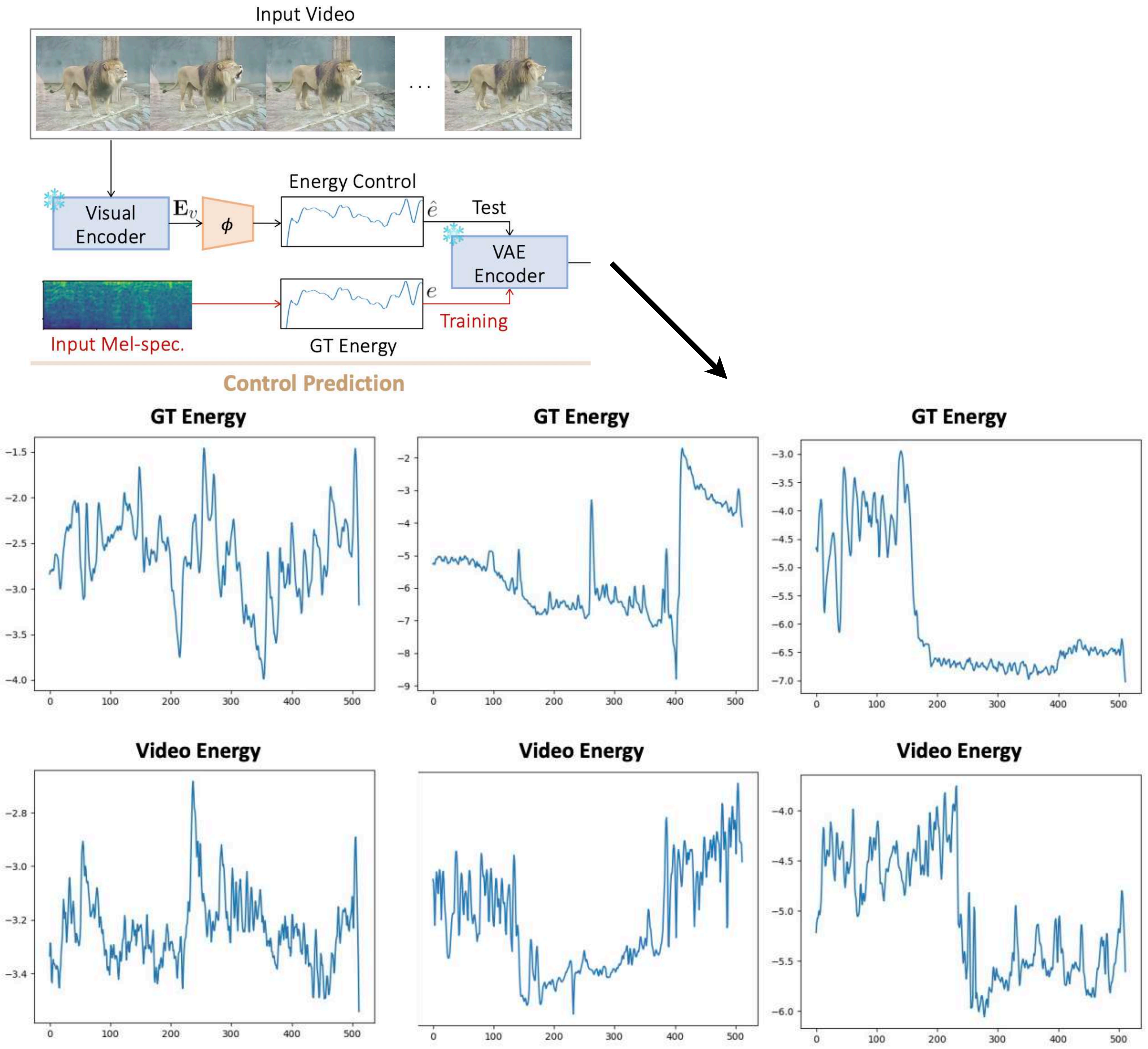
- Part A: Video-to-energy training, with video prediction network (ϕ)
- Part B: Energy adapter training with audio energy



"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024)

Video-to-energy prediction

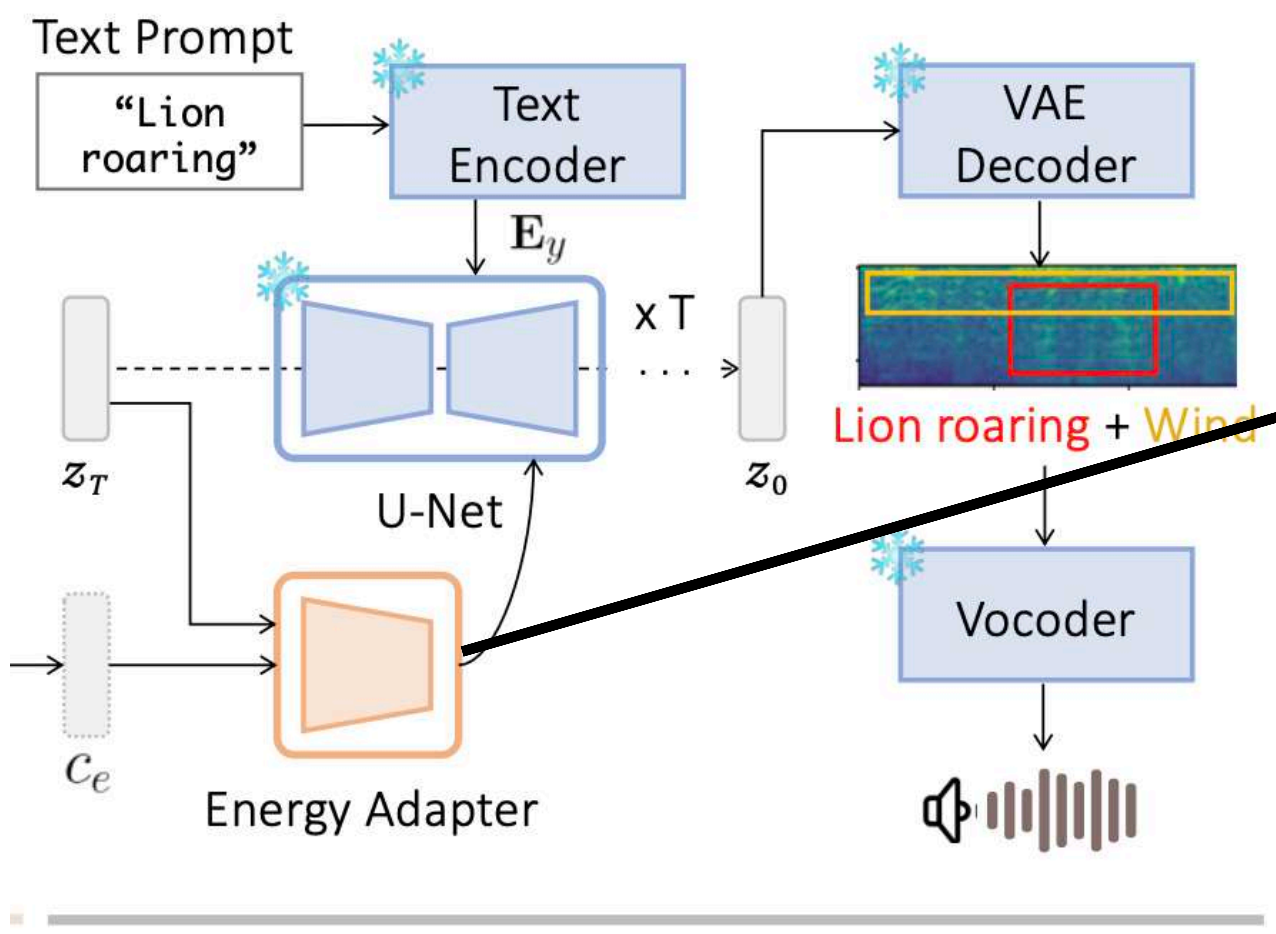
- Video implicitly represents the power of the audio spectrum [1,2]
- For example, the distance or size of objects relates to the volume of sound.
- Our energy (mean of spectrogram on the frequency axis) prediction operates as a **time-varying structured control** to complement the sound according to the **dynamics of the given video**.
- We train the energy prediction model with mean squared error (MSE) loss



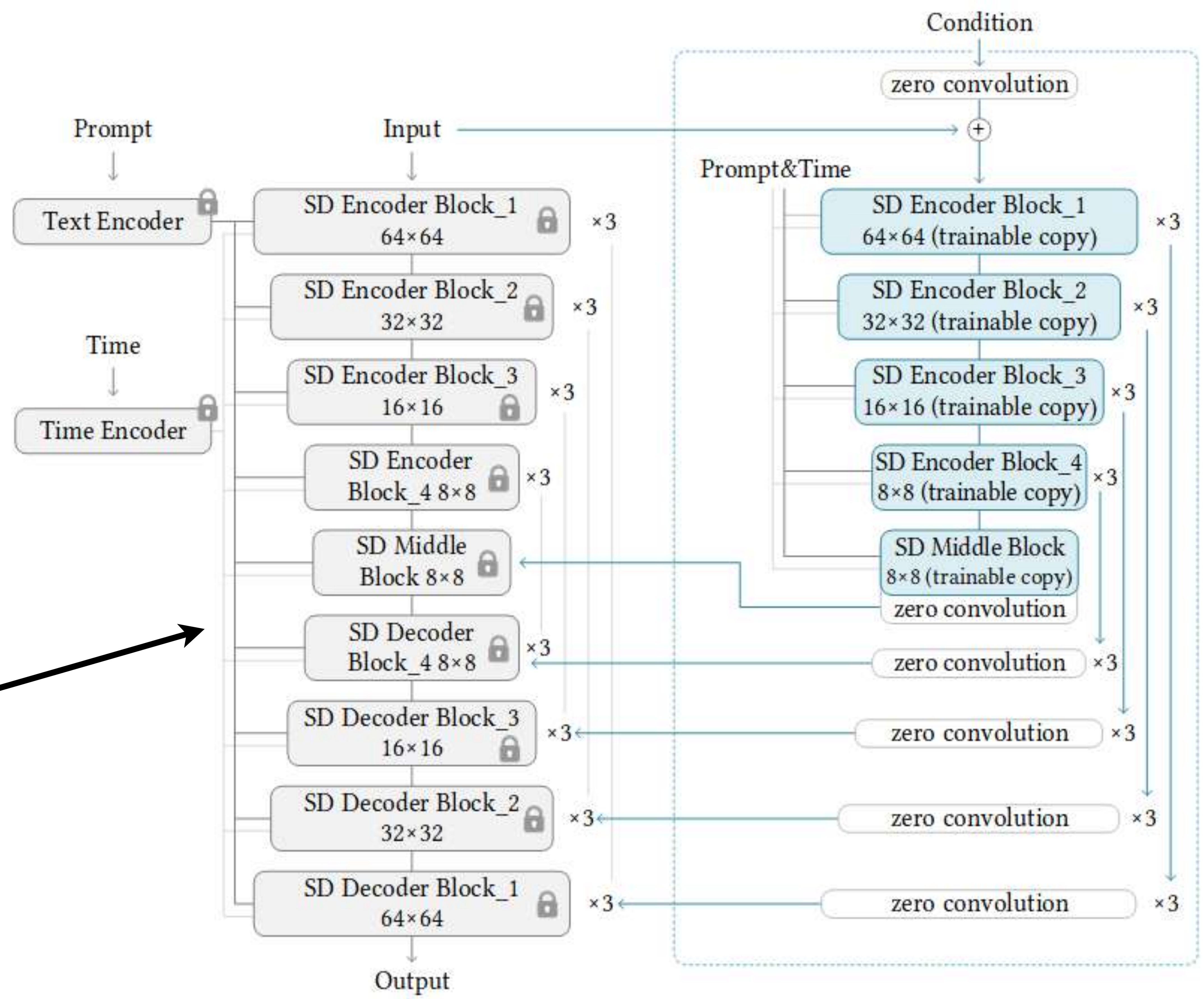
[1] "The Power of Sound (TPoS): Audio Reactive Video Generation with Stable Diffusion" ICCV (2023)
[2] "Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment" CVPR (2023)

Energy adapter

- Motivated by ControlNet, we design an energy adapter to condition energy for AudioLDM
- **Training only additional parameters** (relatively small than original) while freezing pretrained parameters of AudioLDM



Conditional Sound Generation



(a) Stable Diffusion (b) ControlNet

"Adding Conditional Control to Text-to-Image Diffusion Models" ICCV (2023)

Quantitative comparison with SoTAs

- **ReWaS** generates **high-quality** audio (low FD, FAD) and **highly visual-related** audio (high AV-align)
- A human study (73 participants) shows the robustness of ReWaS with a larger margin than SoTAs

Table 2: Performance comparison on VGGSound (Chen et al. 2020) with reproduced five seconds audio samples. "Energy" and "TP" denote energy MAE and number of the trainable parameters.

Model	FD↓	FAD↓	MKL↓	CLAP↑	MAE↓	AV-align↑	# TP↓
SpecVQGAN	26.63	5.57	3.30	0.1336	0.1422	0.2851	379M
Im2wav	16.87	5.94	2.53	0.4001	0.1310	0.2763	365M
Diff-Foley	21.96	6.46	3.15	0.4010	0.1571	0.2059	859M
Seeing&Hearing	20.72	6.58	**2.34**	**0.5805**	0.1668	0.1858	-
ReWaS (Ours)	**15.24**	**2.16**	2.78	0.4353	**0.1149**	**0.3008**	**204M**

Table 4: Human evaluation of V2A methods on audio quality, audiovisual relevance, and temporal alignment with 5-scale MOS.

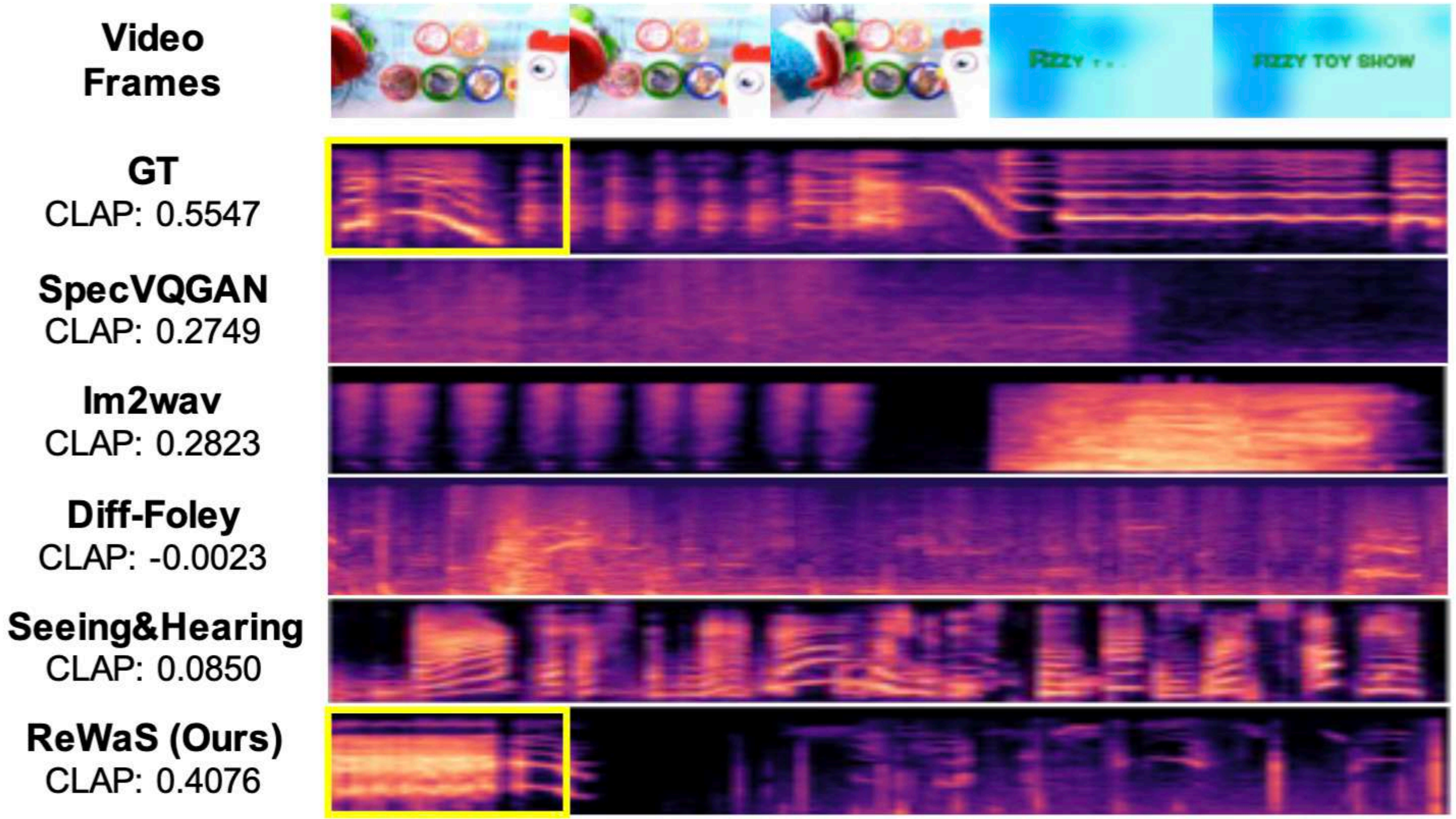
Model	Audio Quality ↑	Relevance ↑	Temporal Alignment ↑
SpecVQGAN	2.76	2.50	2.64
Im2wav	2.97	3.18	3.01
Diff-Foley	2.89	2.97	2.98
ReWaS (Ours)	**3.70**	**4.04**	**3.68**

"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024)

Qualitative comparison with SoTAs

Why do we need text prompts for video-to-audio generation?

- Videos in the real world are so **noisy**!
    - Sometimes, it is hard to distinguish the semantic and redundant frames
    - Existing video-to-audio methods often fail to generate the audio of main subjects.
- Text prompts help to concentrate on the main subjects' sound



Text prompt: "chicken clucking"

"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024)

More demo

VGGSound dataset

Prompt: car engine

"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024)

More demo

VGGSound dataset

`Prompt: cat growling`

"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024)

More demo

Generated video

`Prompt: A rally car swiftly navigates a turn on the racetrack`

"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024), video from Kling

More demo

Generated video

Prompt: A chef is cutting onions in a kitchen, preparing for the dish

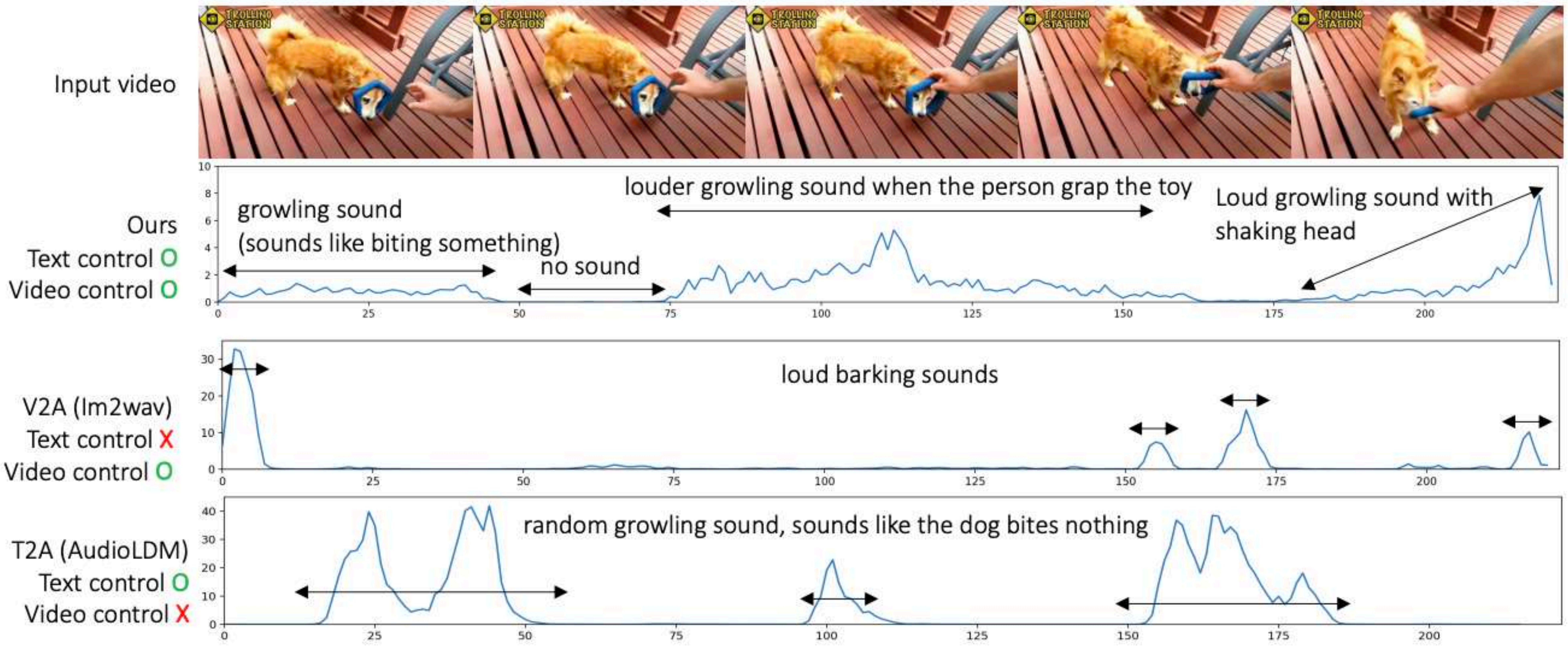
"Read, watch, and scream! Sound generation from text and video" NeurIPS Workshop (2024), video from Kling

Summary

Existing text-to-audio: 😁 strong generalization 😭 cannot imply temporal alignment

Existing video-to-audio: 😭 weak generalization 😁 strong temporal alignment for visual content

ReWaS: 😁 strong generalization 😁 strong temporal alignment for visual content 😁 efficient training



“Read, watch, and scream! Sound generation from text and video” NeurIPS Workshop (2024)

Thank you!
If you have a question,
please send me a message :)
lee.j@navercorp.com