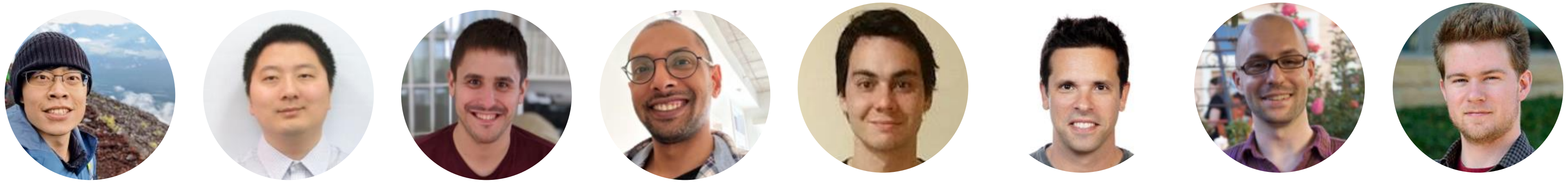
IEEE
WASPAA
2023

CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models

Hao-Wen Dong[1,2]* Xiaoyu Liu[1] Jordi Pons[1] Gautam Bhattacharya[1]
Santiago Pascual[1] Joan Serrà[1] Taylor Berg-Kirkpatrick[2] Julian McAuley[2]

[1] Dolby Laboratories [2] University of California San Diego

* Work done during an internship at Dolby



Dolby UC San Diego

Overview – Text-to-Audio Synthesis

More samples

salu133445.github.io/clipsonic

Prior Work – Text-to-Audio Synthesis

- Diffsound (Yang et al., 2023)
- AudioGen (Kreuk et al., 2023)
- AudioLDM (Liu et al., 2023)
- Make-An-Audio (Huang et al., 2023)
- Noise2Music (Huang et al., 2023)
- MusicLM (Agostinelli et al., 2023)

All rely on large amounts of **text-audio training pairs**

Can we learn text-to-audio synthesis *without* **using any text-audio pairs?**

Yang et al., "Diffsound: Discrete Diffusion Model for Text-to-sound Generation," *TASLP*, 2022.
Kreuk et al., "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.
Liu et al., "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," *ICML*, 2023.
Huang et al., "Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models," *ICML*, 2023.
Huang et al., "Noise2Music: Text-conditioned Music Generation with Diffusion Models," *arXiv preprint arXiv:2302.03917*, 2023.
Agostinelli et al., "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2302.03917*, 2023.

Why NOT Text-audio Pairs?

YouTube videos!

500 hours of videos uploaded per minute

5 billion
text-image pairs

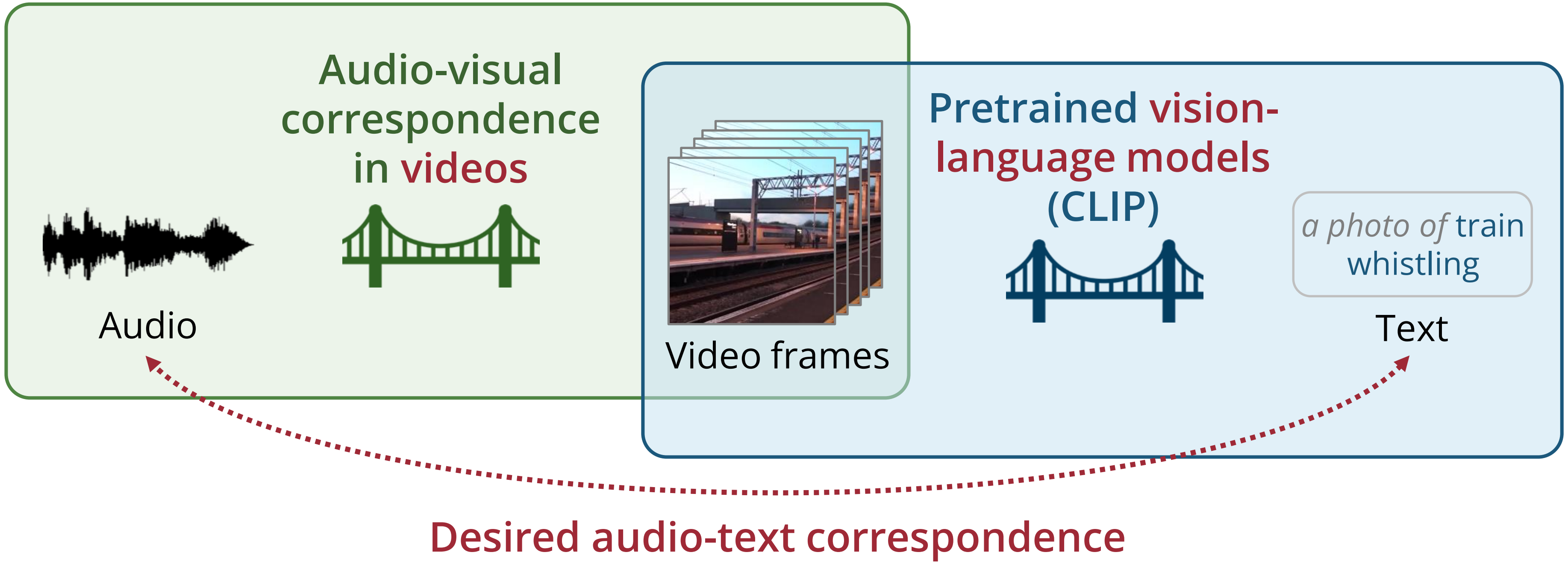
LAION-5B
(Schuhmann et al., 2023)

0.6 million
text-audio pairs

LAION-Audio-630K
(Wu et al., 2023)

Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," *NeurIPS, Datasets and Benchmarks Track*, 2023.
Wu et al., "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," *ICASSP*, 2023.

Leveraging the Visual Domain as a Bridge



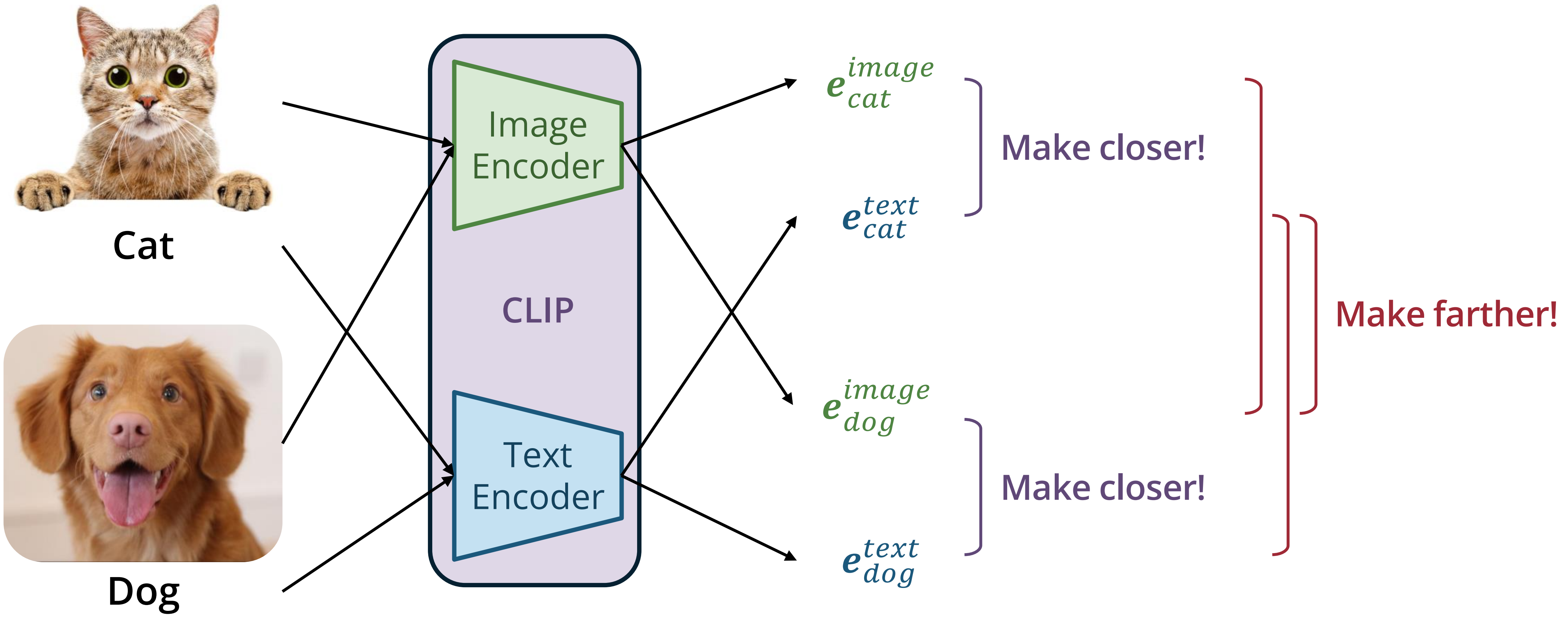
No text-audio pairs required!

Scalable to large video datasets!

Background

CLIP (Contrastive Language-Image Pretraining)

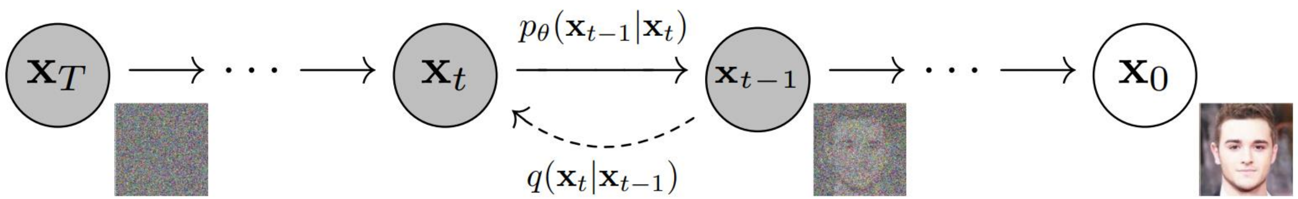
- Learn a shared embedding space for images and texts via *contrastive learning*



Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

Diffusion Model

Add noise gradually
(Forward diffusion process)



Remove noise gradually
(Backward diffusion process)

Input

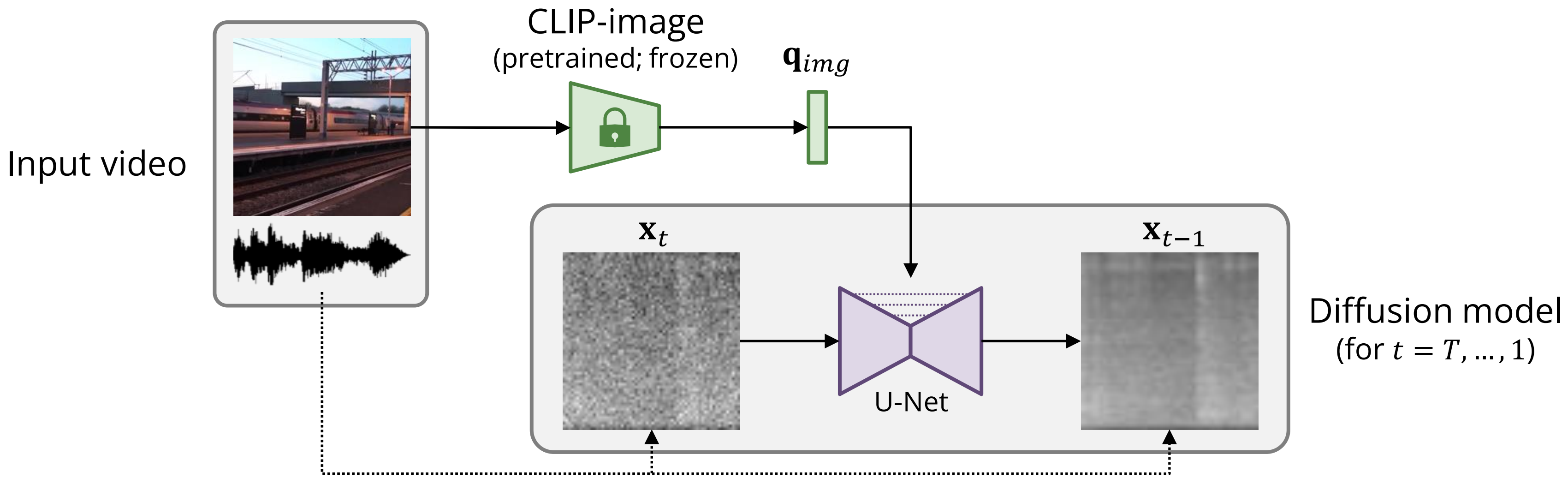
Output

Ho et al., "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

Method

Training – Image-queried

- We train an image-to-audio synthesis model using a diffusion model on mel spectrograms and a pretrained CLIP-image encoder



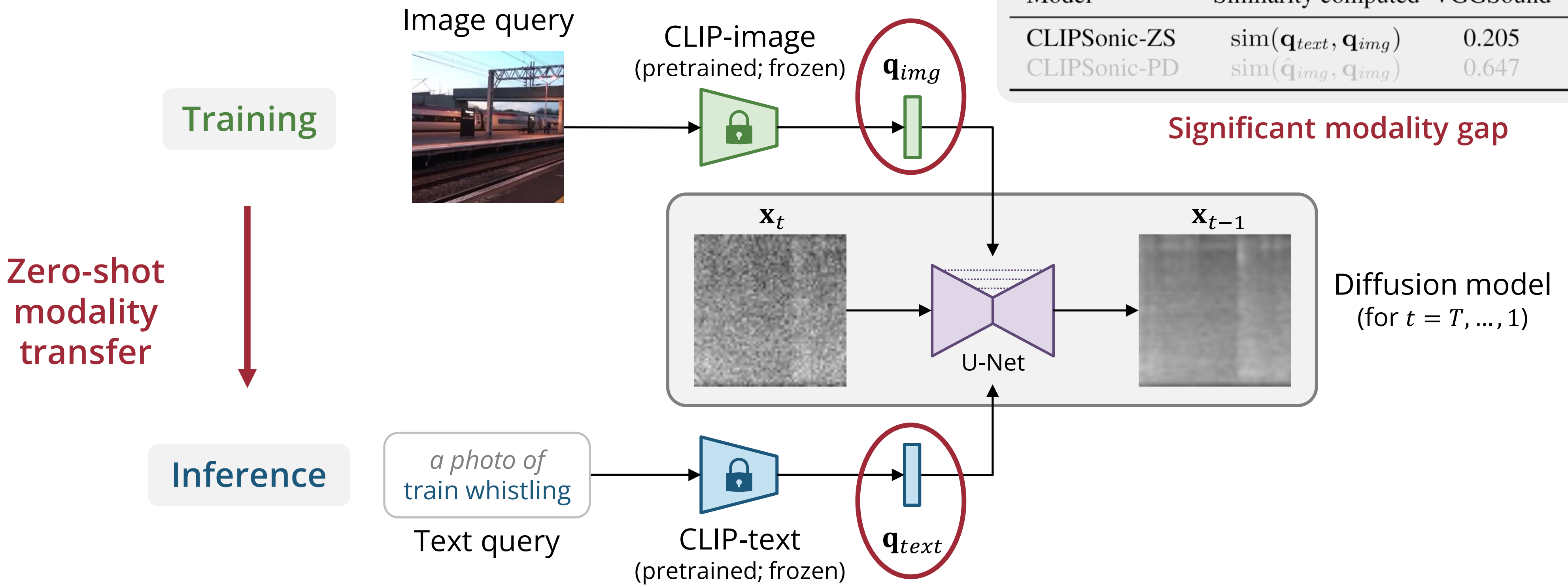
Inference – Zero-shot Modality Transfer (CLIPSonic-ZS)

- We first explore using a pretrained CLIP-text encoder directly

Table 2: Cosine similarities between various query embeddings.

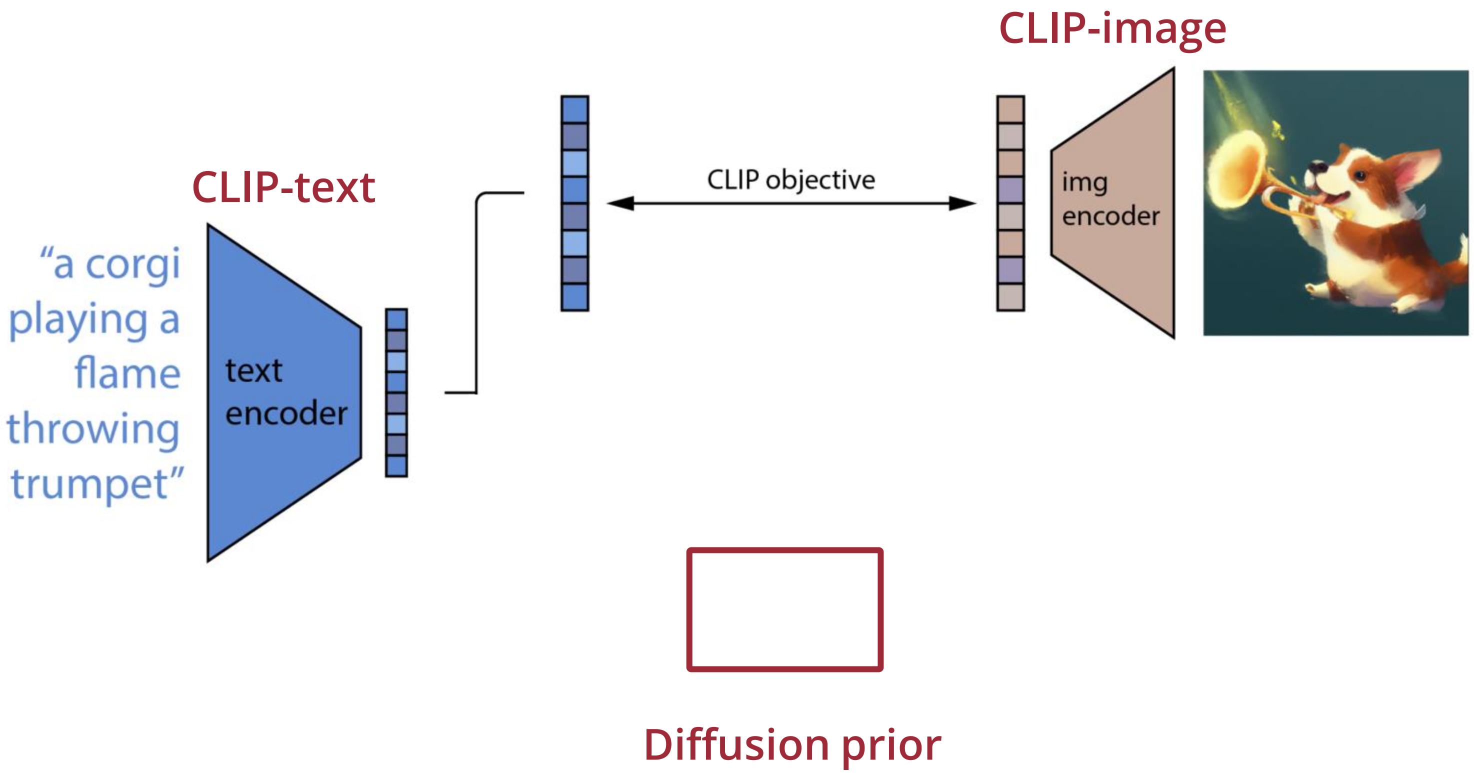
Model	Similarity computed	VGGSound	MUSIC
CLIPSonic-ZS	$\text{sim}(\mathbf{q}_{text}, \mathbf{q}_{img})$	0.205	0.245
CLIPSonic-PD	$\text{sim}(\hat{\mathbf{q}}_{img}, \mathbf{q}_{img})$	0.647	0.720

Significant modality gap



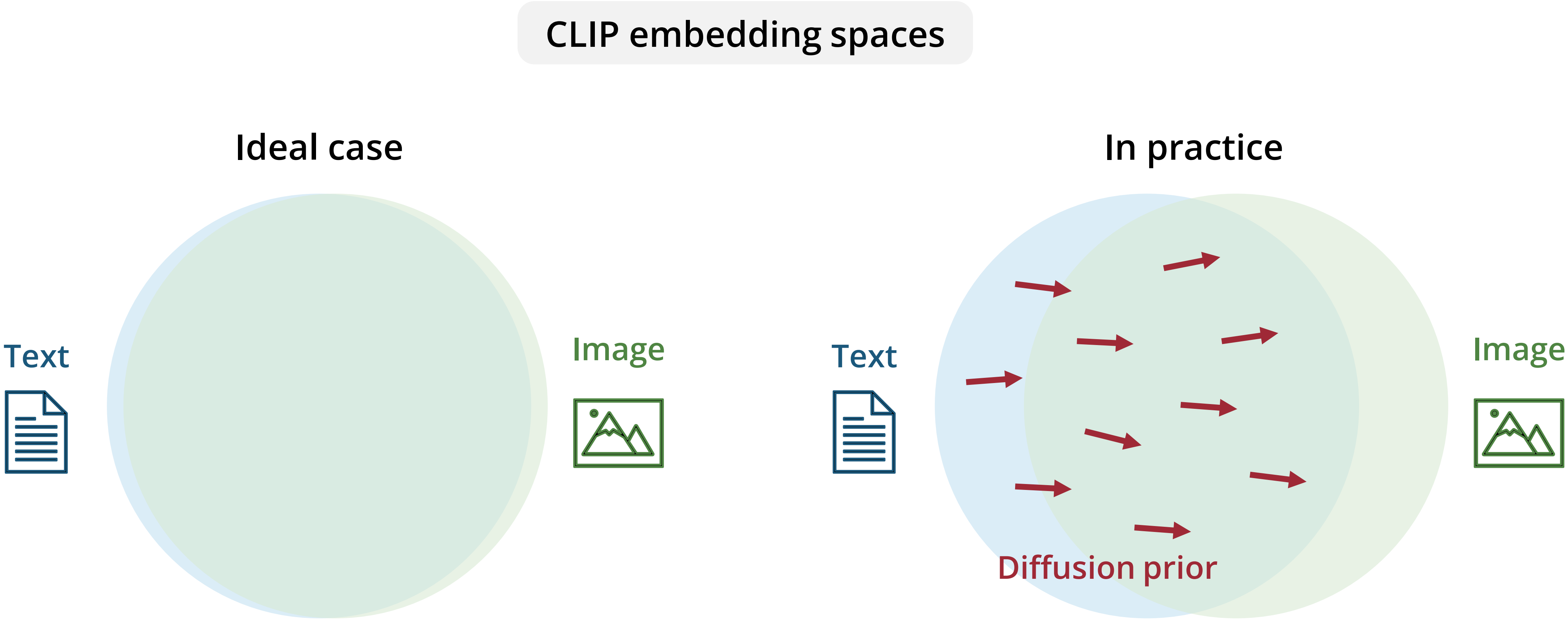
How to overcome this modality gap?

- We leverage a pretrained diffusion prior model (Ramesh et al., 2022)



Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.

Diffusion Prior (Ramesh et al., 2022)



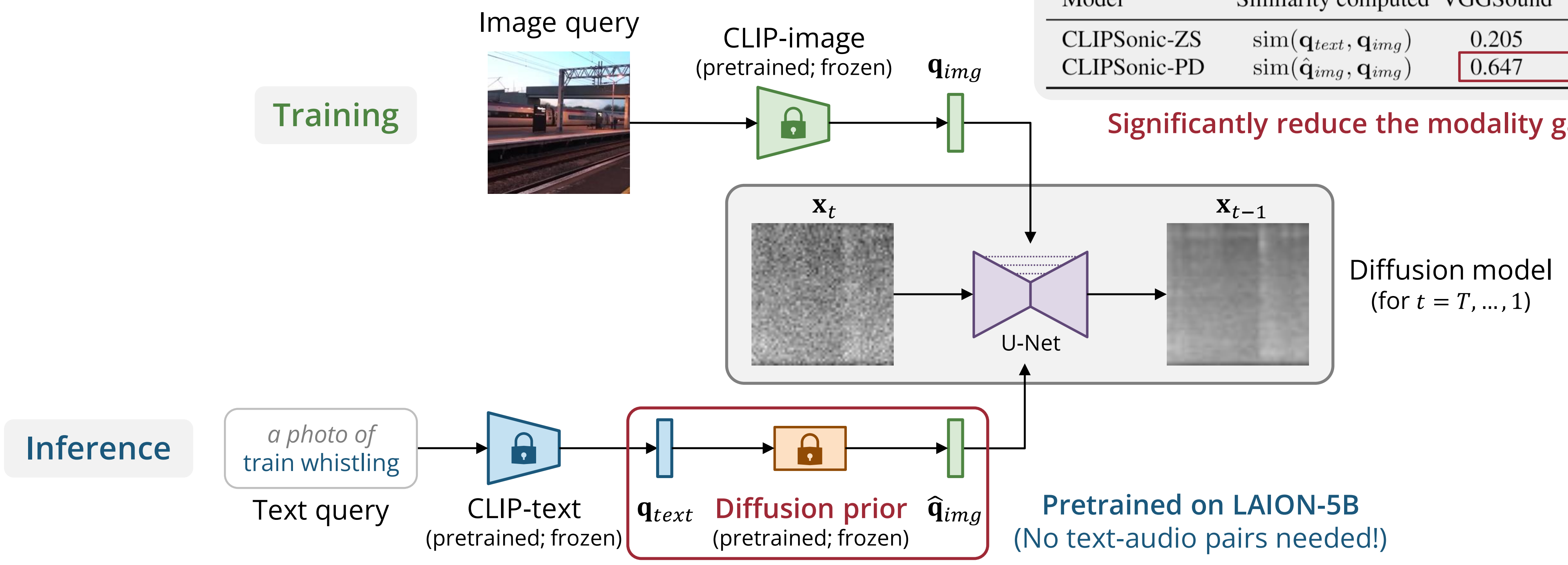
Inference – Pretrained Diffusion Prior (CLIPSonic-PD)

- We then explore using a pretrained diffusion prior model (Ramesh et al., 2022)

Table 2: Cosine similarities between various query embeddings.

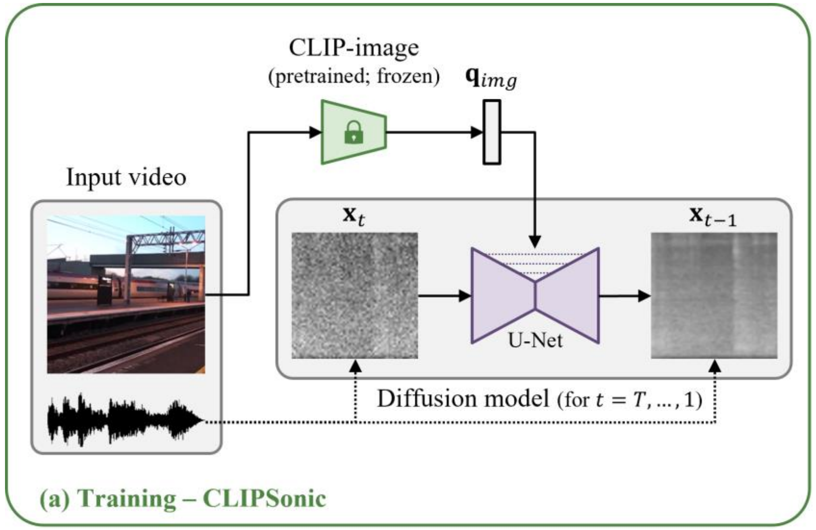
Model	Similarity computed	VGGSound	MUSIC
CLIPSonic-ZS	$\mathrm{sim}(\mathbf{q}_{text}, \mathbf{q}_{img})$	0.205	0.245
CLIPSonic-PD	$\mathrm{sim}(\hat{\mathbf{q}}_{img}, \mathbf{q}_{img})$	0.647	0.720

Significantly reduce the modality gap



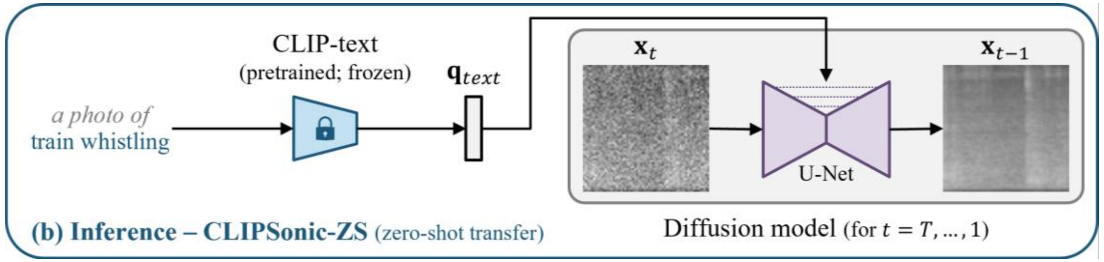
Recap

Training

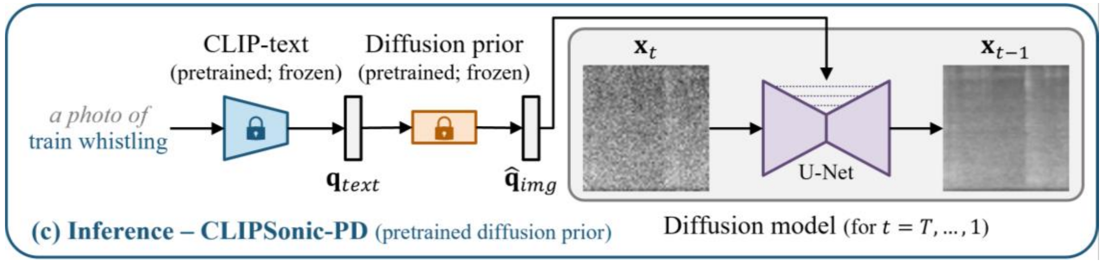


CLIPSonic-IQ
(Image-queried)

Inference



CLIPSonic-ZS
(Zero-shot transfer)



CLIPSonic-PD
(Pretrained diffusion prior)

Experiments

Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Noisy videos with diverse sounds

Zhao et al., "The Sound of Pixels," *ECCV*, 2018.
Chen et al., "VGGSound: A Large-Scale Audio-Visual Dataset," *ICASSP*, 2020.

Implementation Details

Mel spectrogram configuration

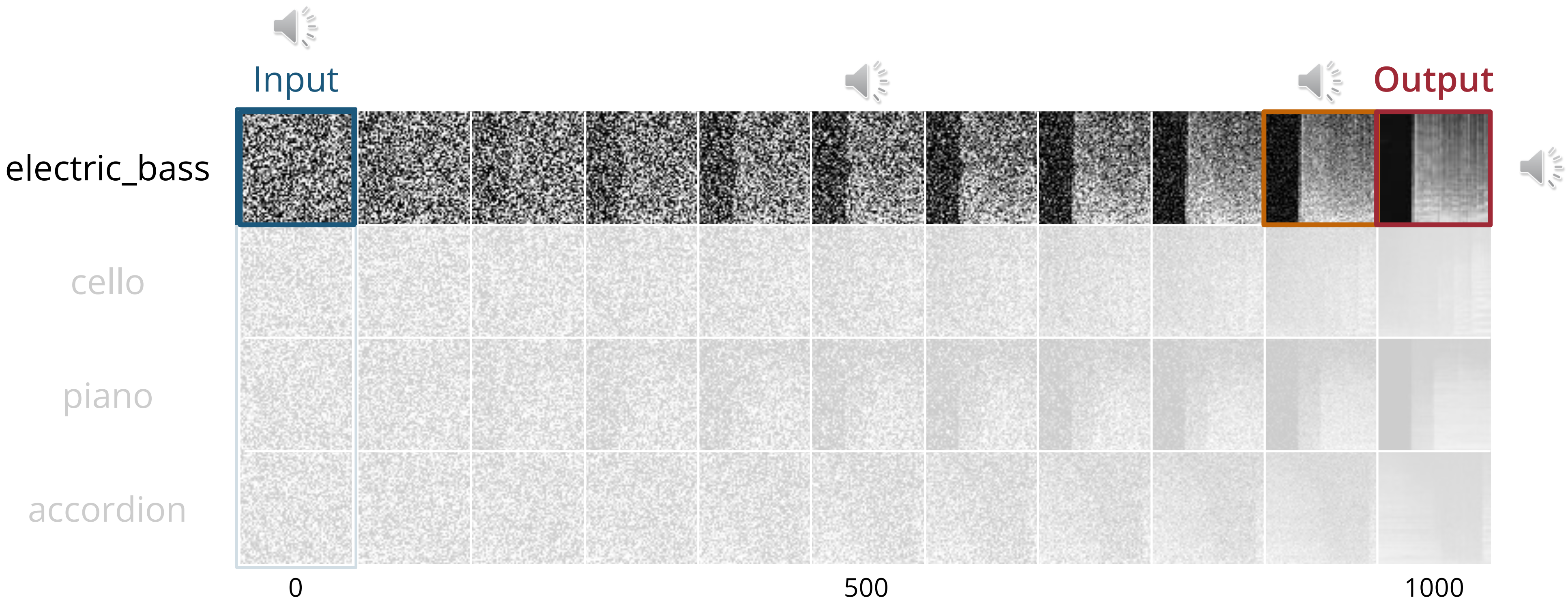
- Sampling rate: 16 kHz
- Hop size: 512
- FFT filter size: 2048
- 64 mel bands
- Inverted back to waveforms using BigVGAN (Lee et al., 2023)

Diffusion model

- Based on Improved DDPM (Nichol and Dhariwal, 2019)
- Diffusion steps:
  - Training: 4000
  - Inference: 1000
- Training iterations
  - MUSIC: 200K (1 day on 2 RTX 2080 Tis)
  - VGGSound: 500K (2 days on 2 RTX 2080 Tis)

Nichol and Dhariwal, "Improved Denoising Diffusion Probabilistic Models," *ICML*, 2019.
Lee et al., "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," *ICLR*, 2023.

Inference – Examples



Text-to-Audio Synthesis – Demo

Rapping

Sea waves

Thunder

Smoke detector beeping

Playing table tennis

Playing violin fiddle

Text-to-Audio Synthesis – Demo

	Rapping	Sea waves	Smoke detector beeping
CLIPSonic-ZS (zero-shot modality transfer)			
CLIPSonic-PD (pretrained diffusion prior)			

	Playing table tennis	Thunder	Playing violin fiddle
CLIPSonic-ZS (zero-shot modality transfer)			
CLIPSonic-PD (pretrained diffusion prior)			

Text-to-Audio Synthesis – Listening Test

Table 3: Listening test results for text-to-audio synthesis (MOS).

Model	VGGSound		MUSIC	
	Fidelity	Relevance	Fidelity	Relevance
CLIPSonic-ZS	2.55 ± 0.22	2.01 ± 0.27	2.98 ± 0.23	3.87 ± 0.24
CLIPSonic-PD	$\mathbf{3.04} \pm \mathbf{0.20}$	2.86 ± 0.25	$\mathbf{3.67} \pm \mathbf{0.18}$	3.91 ± 0.24
Ground truth	3.78 ± 0.19	3.54 ± 0.29	3.90 ± 0.17	4.34 ± 0.18

Significant performance improvement against the baseline!

Image-to-Audio Synthesis – Demo (Out-of-distribution)

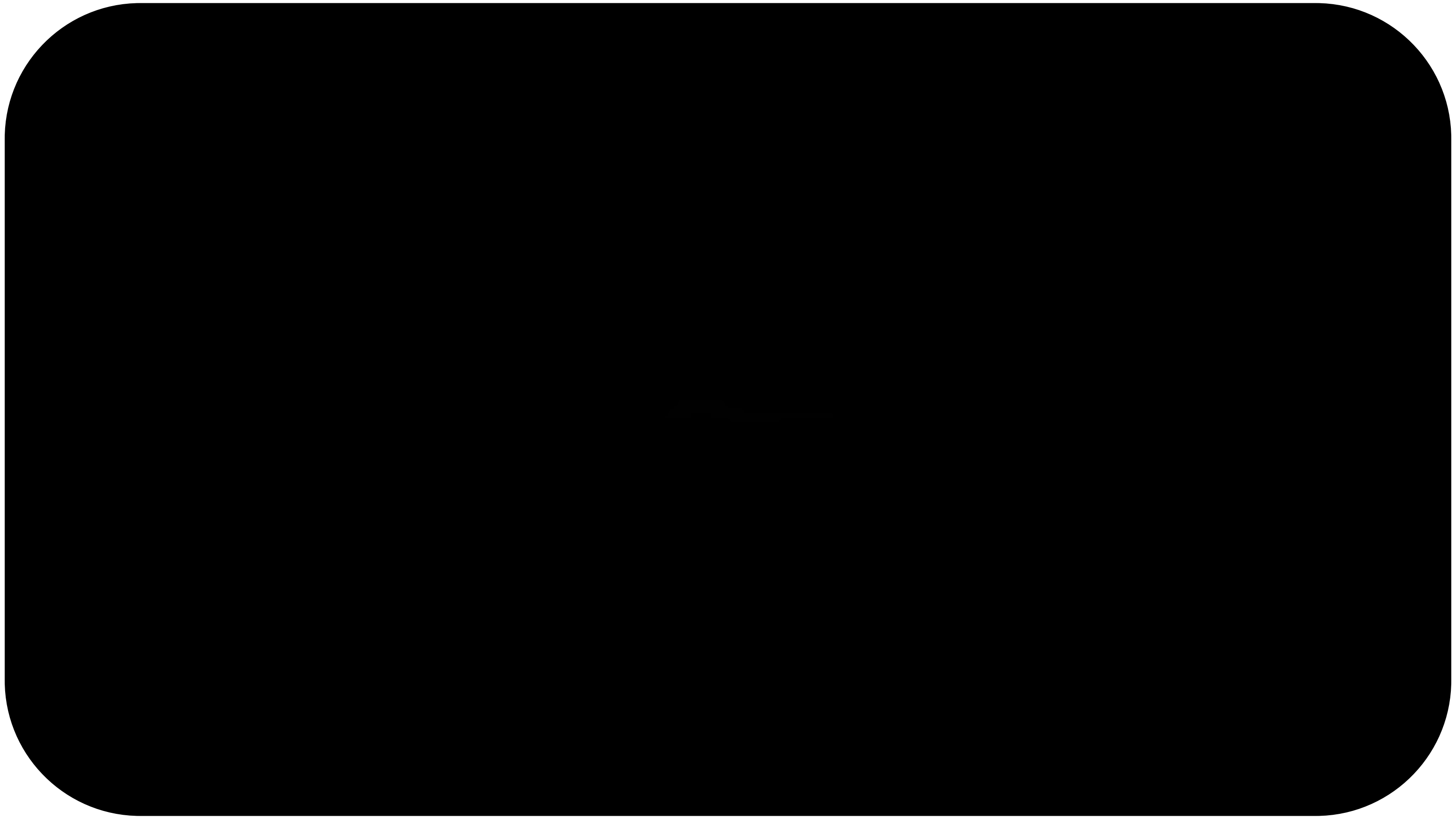


Image-to-Audio Synthesis – Demo (Out-of-distribution)

CLIPSonic-IQ (ours)

Im2wav (Sheffer & Adi, 2023)

SpecVQGAN (Iashin & Rahtu, 2021)

Sheffer and Adi, "I Hear Your True Colors: Image Guided Audio Generation," *ICASSP*, 2023.
Iashin and Rahtu, "Taming Visually Guided Sound Generation," *BMVC*, 2021.

Image-to-Audio Synthesis – Listening Test

Table 4: Listening test results for image-to-audio synthesis (MOS).

Model	Fidelity	Relevance
CLIPSonic-IQ (image-queried)	**3.29 ± 0.16**	3.80 ± 0.19
SpecVQGAN [20]	2.15 ± 0.17	2.54 ± 0.23
im2wav [21]	2.19 ± 0.15	**3.90 ± 0.22**

State-of-the-art image-to-audio performance!

Sheffer and Adi, "I Hear Your True Colors: Image Guided Audio Generation," *ICASSP*, 2023.
Iashin and Rahtu, "Taming Visually Guided Sound Generation," *BMVC*, 2021.

Objective Evaluation Metrics

- Evaluated with Fréchet audio distance (FAD) and CLAP score

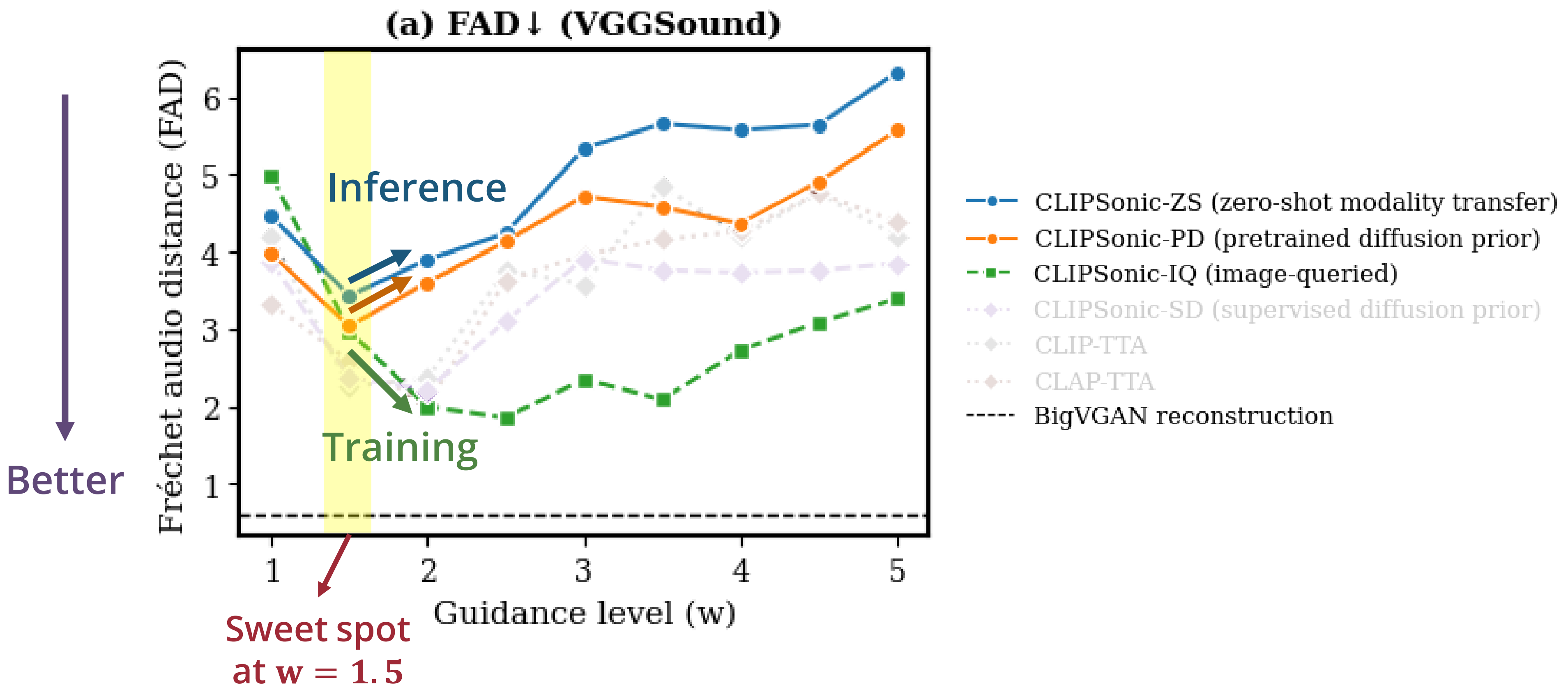
Table 1: Evaluation results on VGGSound and MUSIC datasets, evaluated at $w = 1.5$.

Model	Without text-audio pairs	Query modality		VGGSound		MUSIC	
		Training	Inference	FAD ↓	CLAP [illegible]	[illegible]	[illegible]
CLIPSonic-IQ (image-queried)	-	Image	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
CLIPSonic-ZS (zero-shot modality transfer)	✓	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.284
CLIPSonic-PD (pretrained diffusion [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	13.51	0.254
CLIPS[illegible]	[illegible]	[illegible]	Text	2.37	0.234	12.13	0.299
[illegible]	✓	Text	Text	2.26	0.292	9.39	0.298
[illegible]	✗	Text	Text	2.58	0.296	10.92	0.303
BigVGAN mel spectrogram reconstruction	-	-	-	0.60	0.204	6.21	0.272

Check out our paper for more results!

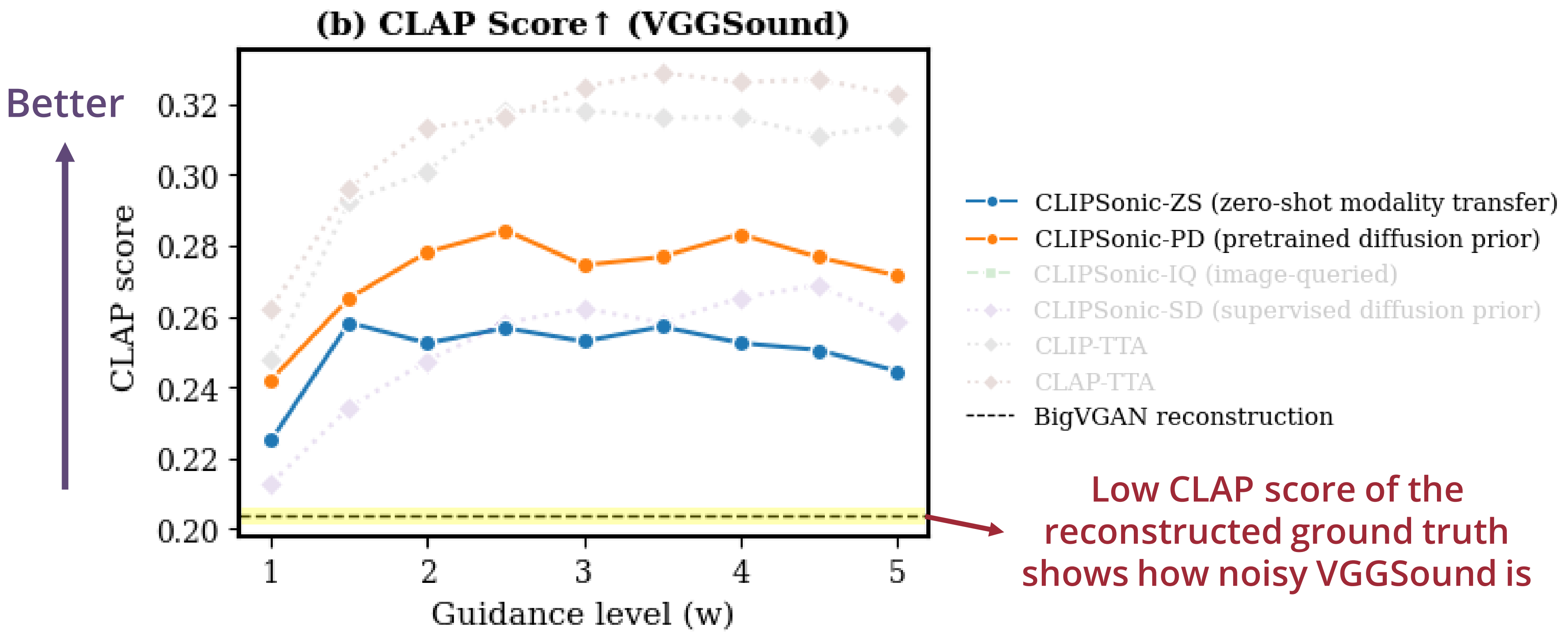
Effects of Classifier-free Guidance

- A guidance level of $w = 1.5$ leads to the lowest FADs for inference



Effects of Classifier-free Guidance

- Larger guidance level leads to stronger adherence to input text query

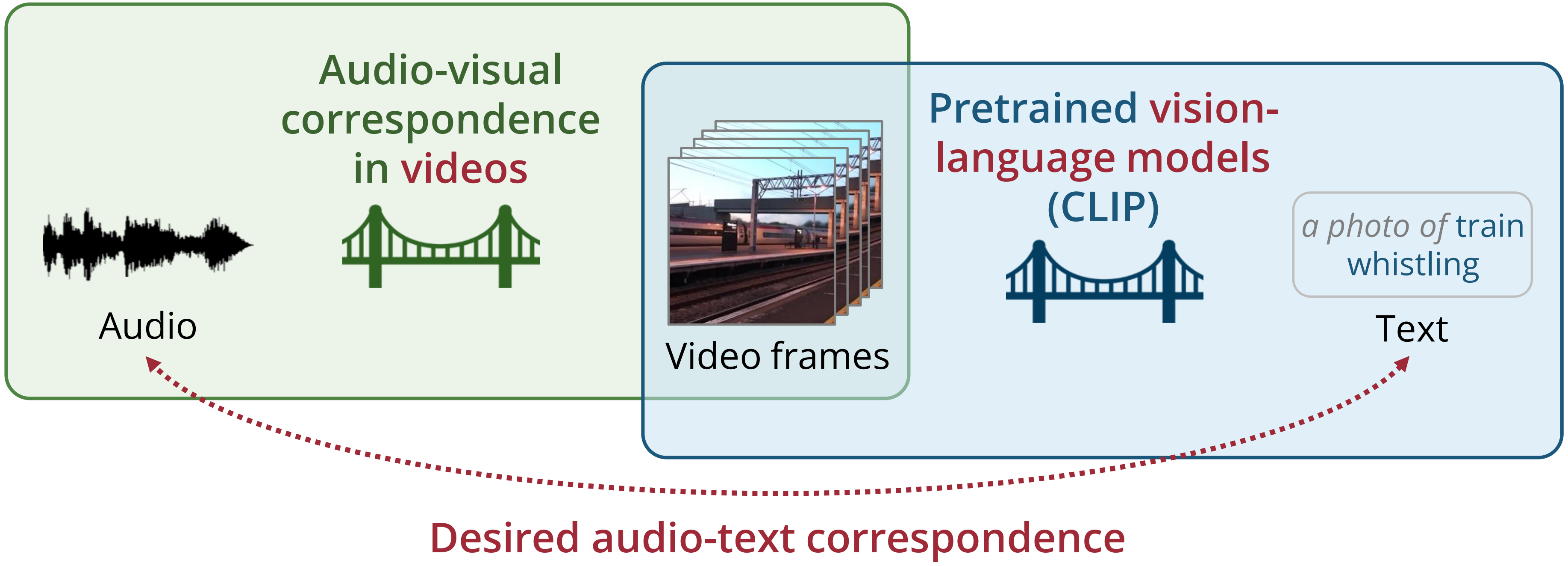


Limitations & Future Work

- Off-screen sounds occur frequently in videos
- Cannot handle purely audio-specific queries
- Can we enable compositional prompts?
- Scale up to larger video datasets!

Conclusion

Leveraging the Visual Domain as a Bridge

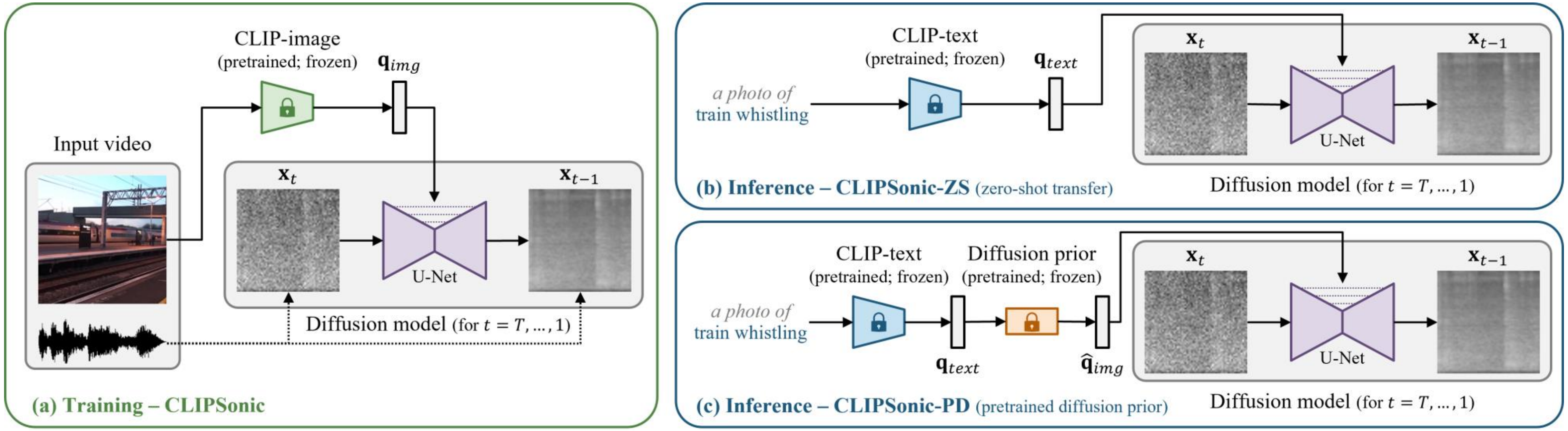


No text-audio pairs required!

Scalable to large video datasets!

Summary

- Proposed a text-to-audio synthesis model that requires *no* text-audio pairs
- Achieved strong text-to-audio synthesis performance
- Achieved state-of-the-art performance in image-to-audio synthesis



Thank you!

Paper: arxiv.org/abs/2306.09635
Demo: salu133445.github.io/clipsonic