JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Sight and Sound

CLIPSynth: Learning Text-to-audio Synthesis from Videos using CLIP and Diffusion Models

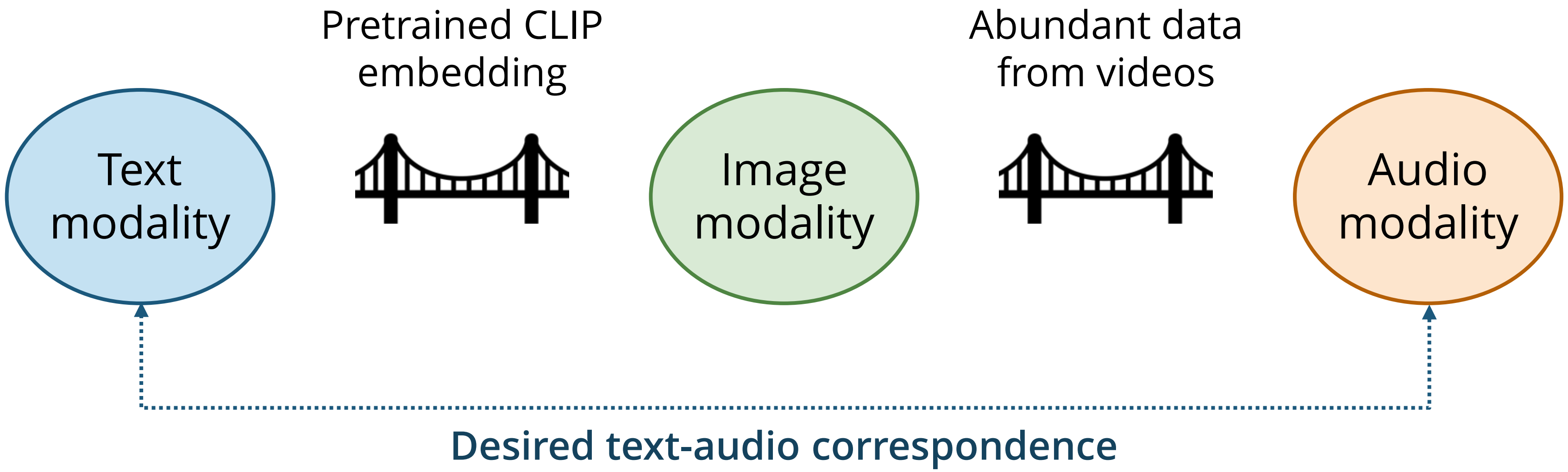
Hao-Wen Dong[1,2] * Gunnar A. Sigurdsson[1] Chenyang Tao[1] Jiun-Yu Kao[1] Yu-Hsiang Lin[1]
Anjali Narayan-Chen[1] Arpit Gupta[1] Tagyoung Chung[1] Jing Huang[1] Nanyun Peng[1,3] Wenbo Zhao[1]

[1]Amazon Alexa AI [2]University of California San Diego [3]University of California, Los Angeles

* Work done during an internship at Amazon

amazon UC San Diego UCLA

Bridging Text-audio Correspondence with Image Modality



Pretrained CLIP embedding

Abundant data from videos

Text modality

Image modality

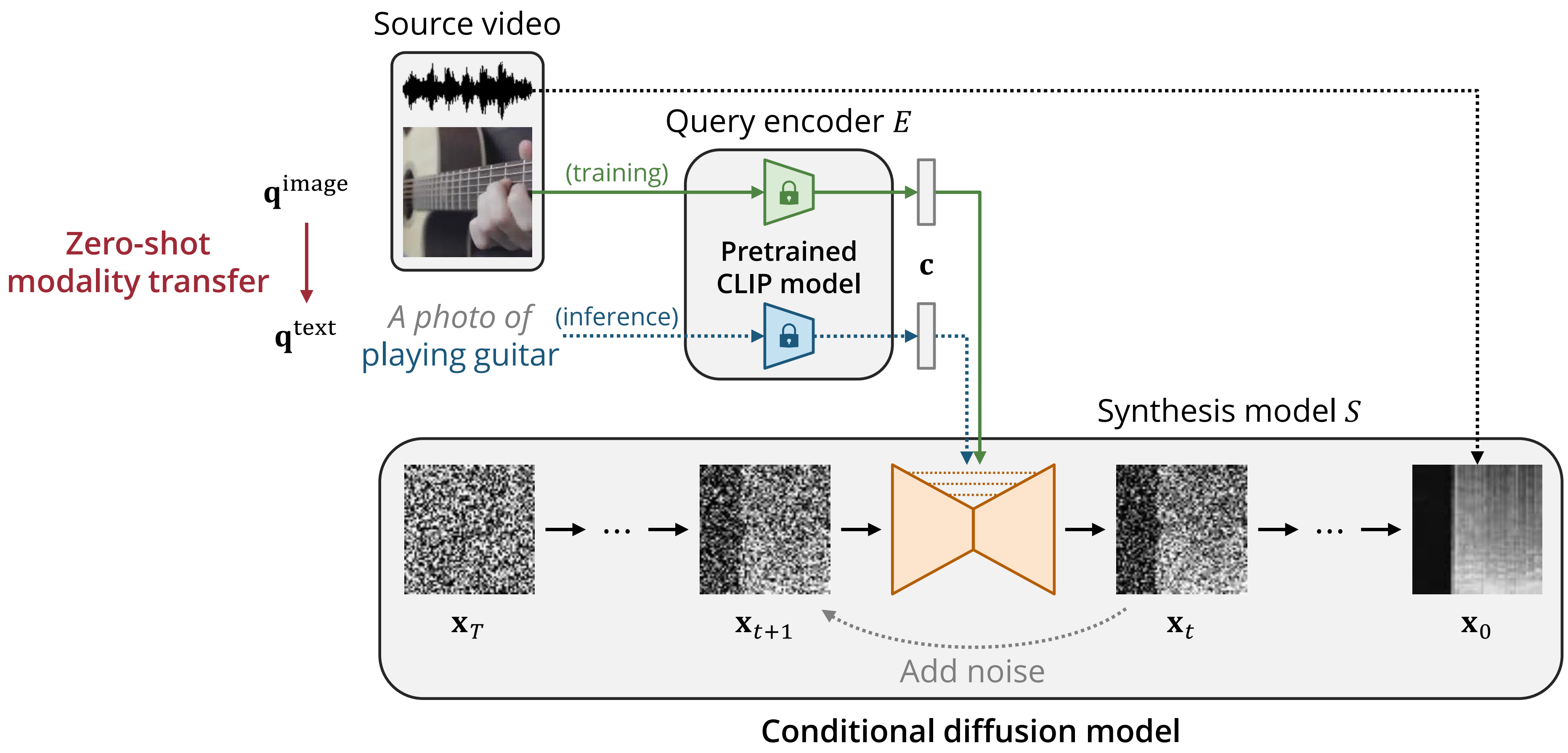
Audio modality

Desired text-audio correspondence

No text-audio pairs required!

Scalable to large video datasets!

CLIPSynth


Source video
Query encoder E
(training)
$\mathbf{q}^{\text{image}}$
Zero-shot modality transfer
$\mathbf{q}^{\text{text}}$
A photo of playing guitar
(inference)
Pretrained CLIP model
$\mathbf{c}$
Synthesis model S
$\mathbf{x}_T$
…
$\mathbf{x}_{t+1}$
$\mathbf{x}_t$
…
$\mathbf{x}_0$
Add noise
Conditional diffusion model

Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

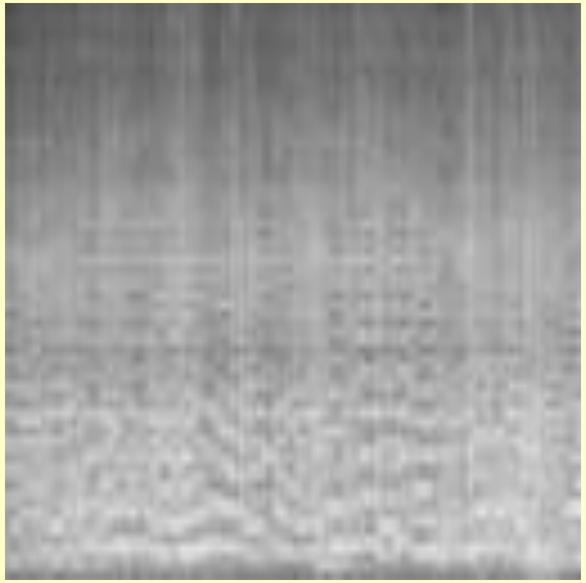
Bird chirping, tweeting

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

Text-to-Audio Synthesis Demo on MUSIC

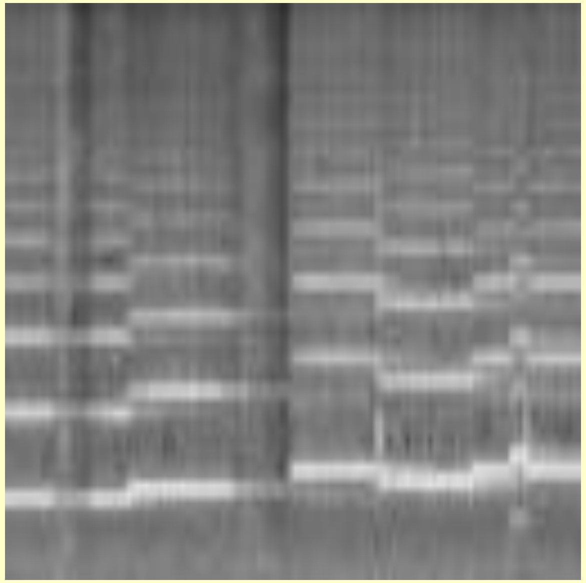
Demo

Cello



Acoustic Guitar

Flute

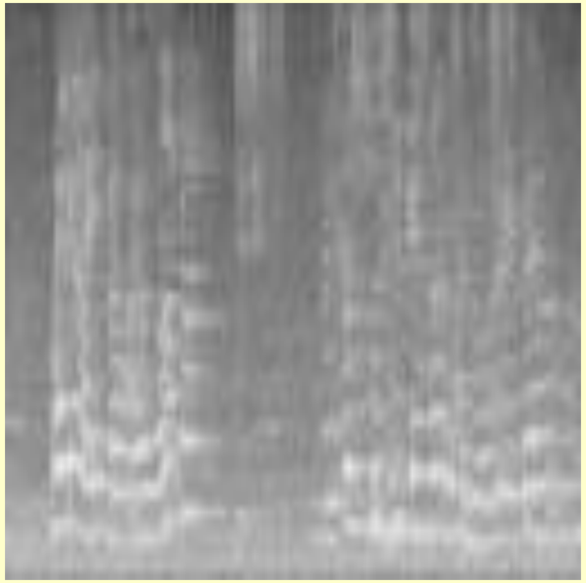


Text-to-Audio Synthesis Demo on VGGSound

Demo

People Crowd

Child Speech, child speaking



Ambulance Siren

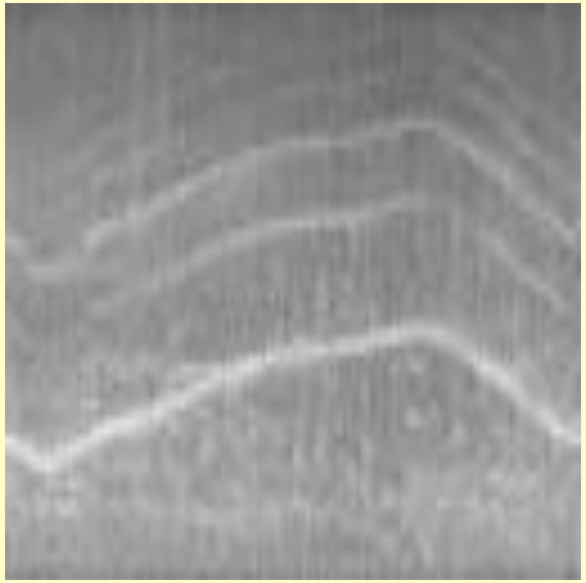


Image-to-Audio Synthesis Demo on VGGsound

Demo

More Results in the Paper!

Objective evaluation

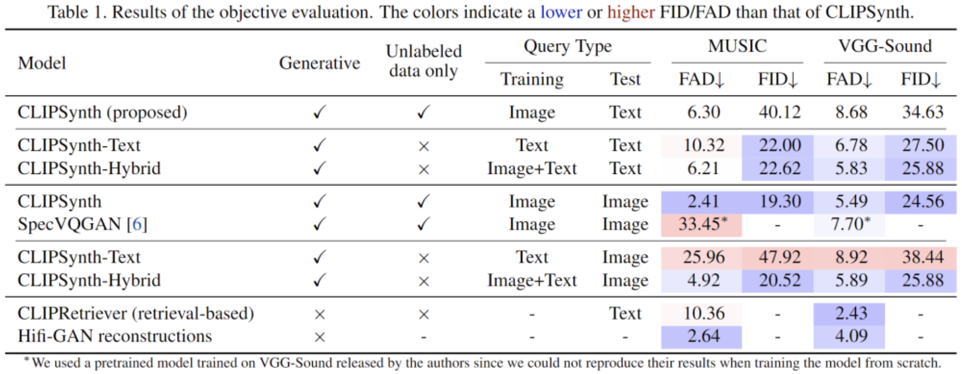


Table 1. Results of the objective evaluation. The colors indicate a lower or higher FID/FAD than that of CLIPSynth.

Model	Generative	Unlabeled data only	Query Type		MUSIC		VGG-Sound	
			Training	Test	FAD↓	FID↓	FAD↓	FID↓
CLIPSynth (proposed)	✓	✓	Image	Text	6.30	40.12	8.68	34.63
CLIPSynth-Text	✓	×	Text	Text	10.32	22.00	6.78	27.50
CLIPSynth-Hybrid	✓	×	Image+Text	Text	6.21	22.62	5.83	25.88
CLIPSynth	✓	✓	Image	Image	2.41	19.30	5.49	24.56
SpecVQGAN [6]	✓	✓	Image	Image	33.45*	-	7.70*	-
CLIPSynth-Text	✓	×	Text	Image	25.96	47.92	8.92	38.44
CLIPSynth-Hybrid	✓	×	Image+Text	Image	4.92	20.52	5.89	25.88
CLIPRetriever (retrieval-based)	×	×	-	Text	10.36	-	2.43	-
Hifi-GAN reconstructions	×	-	-	-	2.64	-	4.09	-

* We used a pretrained model trained on VGG-Sound released by the authors since we could not reproduce their results when training the model from scratch.

Subjective listening test

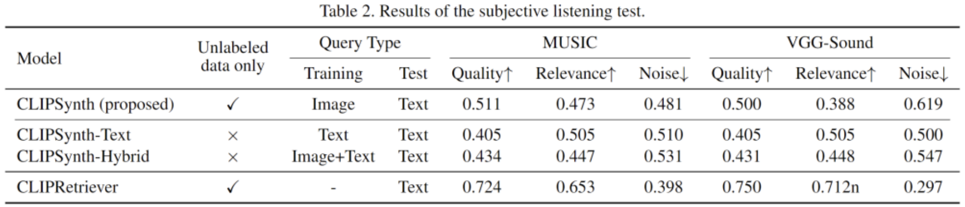


Table 2. Results of the subjective listening test.

Model	Unlabeled data only	Query Type		MUSIC			VGG-Sound		
		Training	Test	Quality↑	Relevance↑	Noise↓	Quality↑	Relevance↑	Noise↓
CLIPSynth (proposed)	✓	Image	Text	0.511	0.473	0.481	0.500	0.388	0.619
CLIPSynth-Text	×	Text	Text	0.405	0.505	0.510	0.405	0.505	0.500
CLIPSynth-Hybrid	×	Image+Text	Text	0.434	0.447	0.531	0.431	0.448	0.547
CLIPRetriever	✓	-	Text	0.724	0.653	0.398	0.750	0.712n	0.297

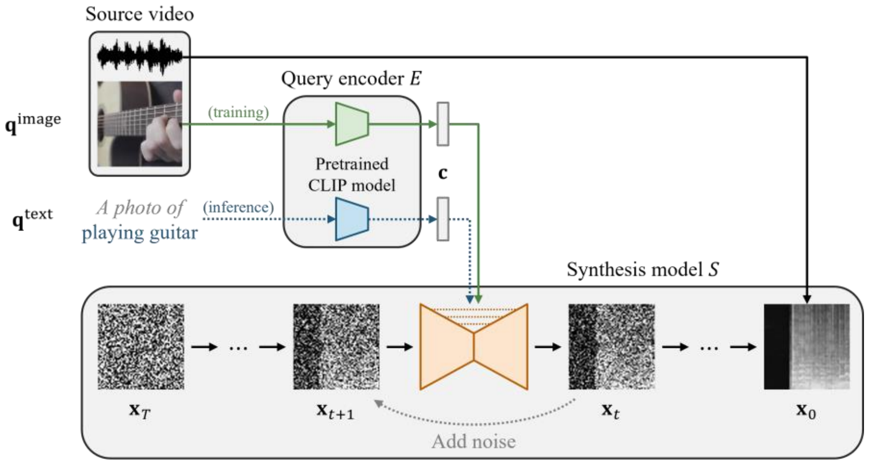
Limitations & Future Work

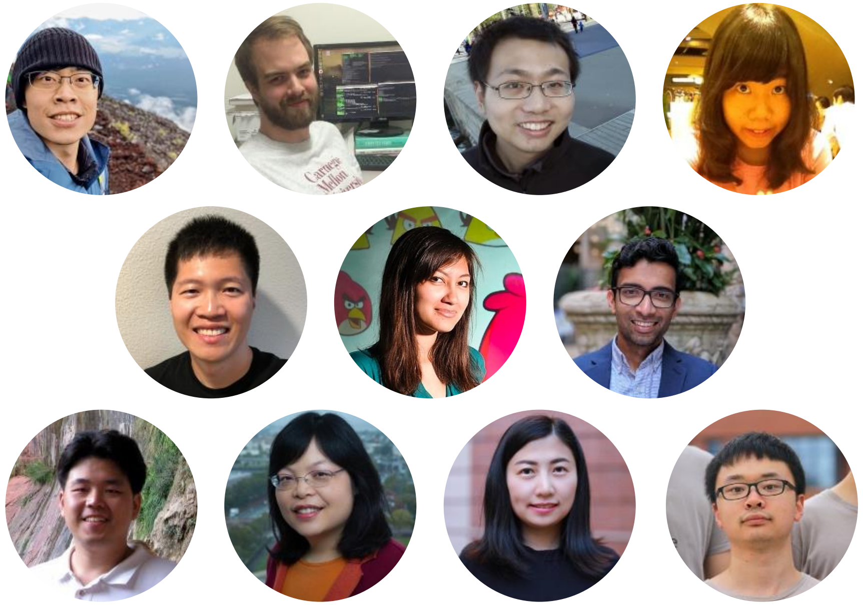
- Off-screen sounds occur frequently in videos
  - Increases undesired zero-shot modality transfer gap
- Cannot handle purely audio-specific queries
  - Because they have little meaning in the visual domain
  - For example, “loud,” “quiet,” “high-pitched” and “low-pitched”
- How to enable combinatory prompts?
  - For example, “piano + guitar”
- Scale up to larger video datasets!

Thank you!

CLIPSynth

A new text-to-audio synthesis model that can be trained **using only unlabeled videos**





amazon UC San Diego UCLA