ViSAGe: Video-to-Spatial Audio Generation

Jaeyeon Kim, Heeseung Yun, Gunhee Kim

SEOUL NATIONAL UNIV. VISION & LEARNING

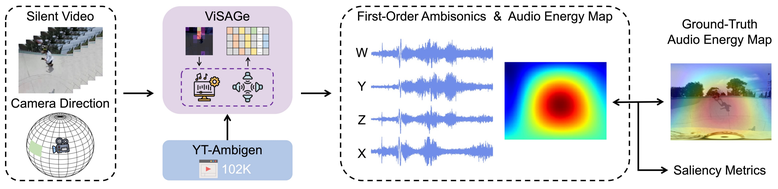
ICLR

[TL;DR] We propose YT-Ambigen, a novel dataset for video-to-ambisonics generation with 102k videos and ViSAGe framework for end-to-end spatial audio generation

Motivation

- **Spatial audio** is crucial for immersive experience of audio-visual scenes
- BUT production is expensive and challenging
- Existing works only partially tackles the problem
  - **Video-to-audio generation**: Generates mono audio
  - **Audio spatialization**: Requires reference mono audio
  - Combining above: May lead to suboptimal spatialization

Objective: Generate **first-order ambisonics (FOA)** from **silent FoV video & camera direction**



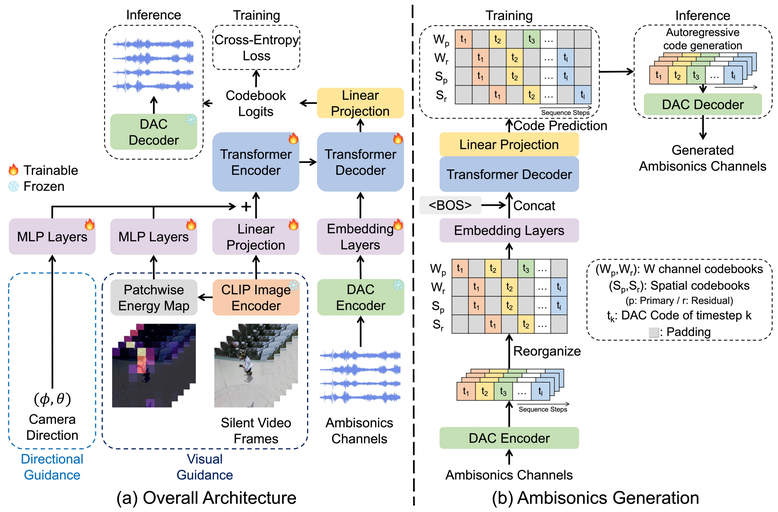
YT-Ambigen

- Introduce a large-scale dataset suitable for the task
- Curated by filtering audio events and audio-visual correspondence

Dataset	# Clips	Length	Audio Type	Audio Gen
VEGAS	28K	55h	Non-Spatial	✅
VAS	13K	24h	Non-Spatial	✅
VGGSound	200K	560h	Non-Spatial	✅
FairPlay	2K	5h	Binaural	❌
OAP	64K	26h	Binaural	❌
YT-360	89K	246h	FOA	❌
STARSS23	0.2K	7.5h	FOA	❌
YT- Ambigen	**102K**	**142h**	**FOA**	✅

ViSAGe

- E2E framework for **vi**deo-to-**s**patial **a**udio **ge**neration
  - Directional embedding and patchwise energy map
  - Efficient code generation pattern
  - Rotation augmentation
  - Directional and visual guidance



Experiments

- Comparison with two-stage baselines

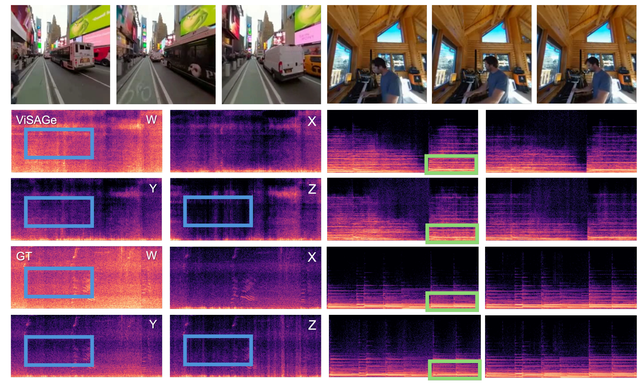
Model		Semantic Metrics		Spatial Metrics	
V2A	Spatialization	FAD_{dec} ↓	KLD_{dec} ↓	CC_{all} ↑	CC_{1fps} ↑
SpecVQGAN[1]	Ambi Enc.	5.94	2.56	0.349	0.337
	Audio Spatial.	6.40	2.43	0.619	0.587
Diff-foley[2]	Ambi Enc.	5.68	2.60	0.349	0.337
	Audio Spatial.	7.24	2.51	0.577	0.537
ViSAGe (Directional)		5.56	2.01	**0.721**	**0.671**
ViSAGe (Directional & Visual)		**3.86**	**1.71**	0.635	0.584

- Ablation on model components

Model				Semantic Metrics		Spatial Metrics	
PT	DIR	PE	RA	FAD_{dec} ↓	KLD_{dec} ↓	CC_{all} ↑	CC_{1fps} ↑
✅				**3.73**	1.76	0.430	0.398
✅	✅			3.74	1.77	0.524	0.482
	✅	✅		4.44	1.78	0.544	0.498
✅	✅	✅		4.01	1.78	0.531	0.486
✅	✅	✅	✅	3.86	**1.71**	**0.635**	**0.584**

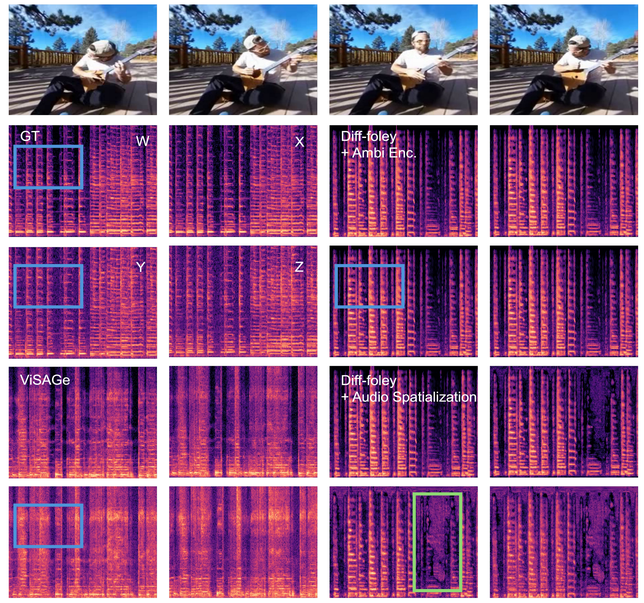
Qualitative Examples

- Generation result of ViSAGe



- Audio energy map visualization

- Comparison with two-stage baselines



References

[1] Iashin and Rahtu, Taming visually guided sound generation, In *BMVC*, 2021

[2] Luo et al., Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, In *NeurIPS*, 2023