ViSAGe: Video-to-Spatial Audio Generation

Jaeyeon Kim, Heeseung Yun, Gunhee Kim
Seoul National University
jaeyeonkim99@snu.ac.kr, heeseung.yun@vision.snu.ac.kr, gunhee@snu.ac.kr

ICLR
International Conference On
Learning Representations

Contents

Introduction

- Spatial Audio
  - Human perceive the world through both visual and auditory cues
  - Key to immersive experiences in visual scenes

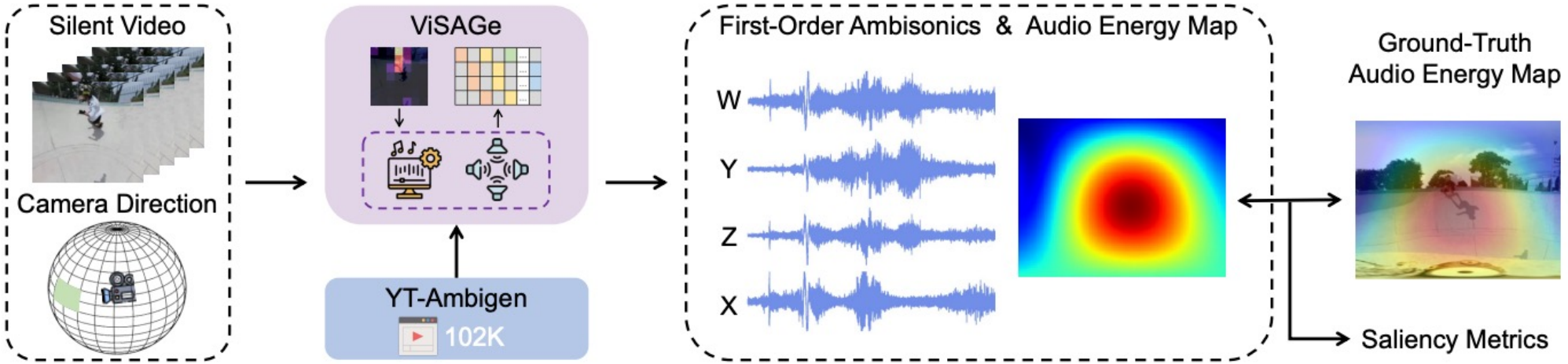
- Challenge in spatial audio production
  - Requires expensive equipment and expert skills
  - Sound effects in video are often recreated manually (i.e. Foley synthesis)

→ **Generating spatial audio for silent videos** has immediate and impactful applications

Introduction

- Previous Approaches
  - **Video-to-Audio Generation** → Generates only mono audio
  - **Audio Spatialization** → Requires reference mono audio
  - Combining the above may lead to inaccurate spatialization due to misalignment

→ Propose **new task, dataset, and method** to generate spatial audio from silent video



Video-to-Ambisonics Generation: Background

- **First-order Ambisonics (FOA)**
  - 3D surrounding sound format based on first-order spherical harmonic decomposition
  - Four channels **(W, X, Y, Z)**
    - W: omnidirectional microphone at center
    - (X, Y, Z): figure-of-eight microphone aligned with corresponding axis

Video-to-Ambisonics Generation: Task

- Generating **FOA** given **silent Field-of-view (FoV) video + Camera direction**
  - FoV: Wider application than panoramic video
  - However, lacks where visual event occur with in 3D surroundings

  → **Camera direction** to represent where visual scene is taking place

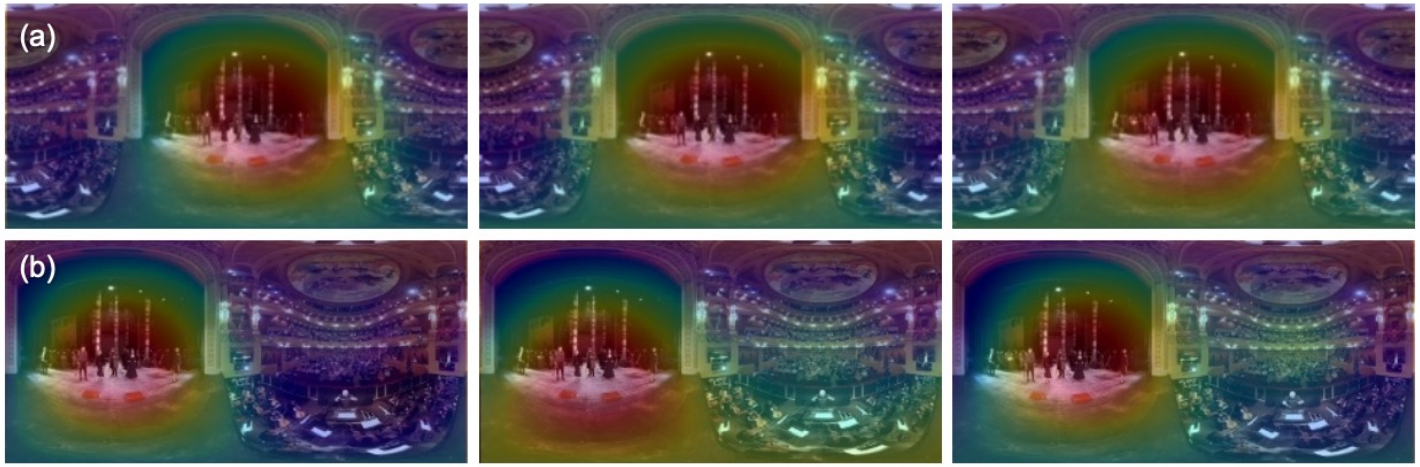


Figure 7: Qualitative example of the camera direction parameter. (a) Camera direction is set to front: $(0, 0)$. (b) Camera direction is set to front-left: $(\frac{\pi}{3}, 0)$.

Video-to-Ambisonics Generation: Evaluation Metrics

- Semantic Metrics
  - Adopt widely used FAD, KLD metrics for decoded FOA (**FAD_{dec}, KLD_{dec}**)
  - Evaluate fidelity of each channel using FAD (**FAD_{avg}**)

- Spatial Metrics
  - Generated audio cannot be directly compared to ground-truth audio
  - Adopt **visual saliency metric** between **generated and ground-truth audio energy map**

CC: 0.632, AUC: 0.833

Dataset: YT-Ambigen

- Introduce **YT-Ambigen**
  - Existing datasets: No spatial audio or not suitable for audio generation
  - Curate large-scale (102K clips) dataset tailored for video-to-ambisonics generation

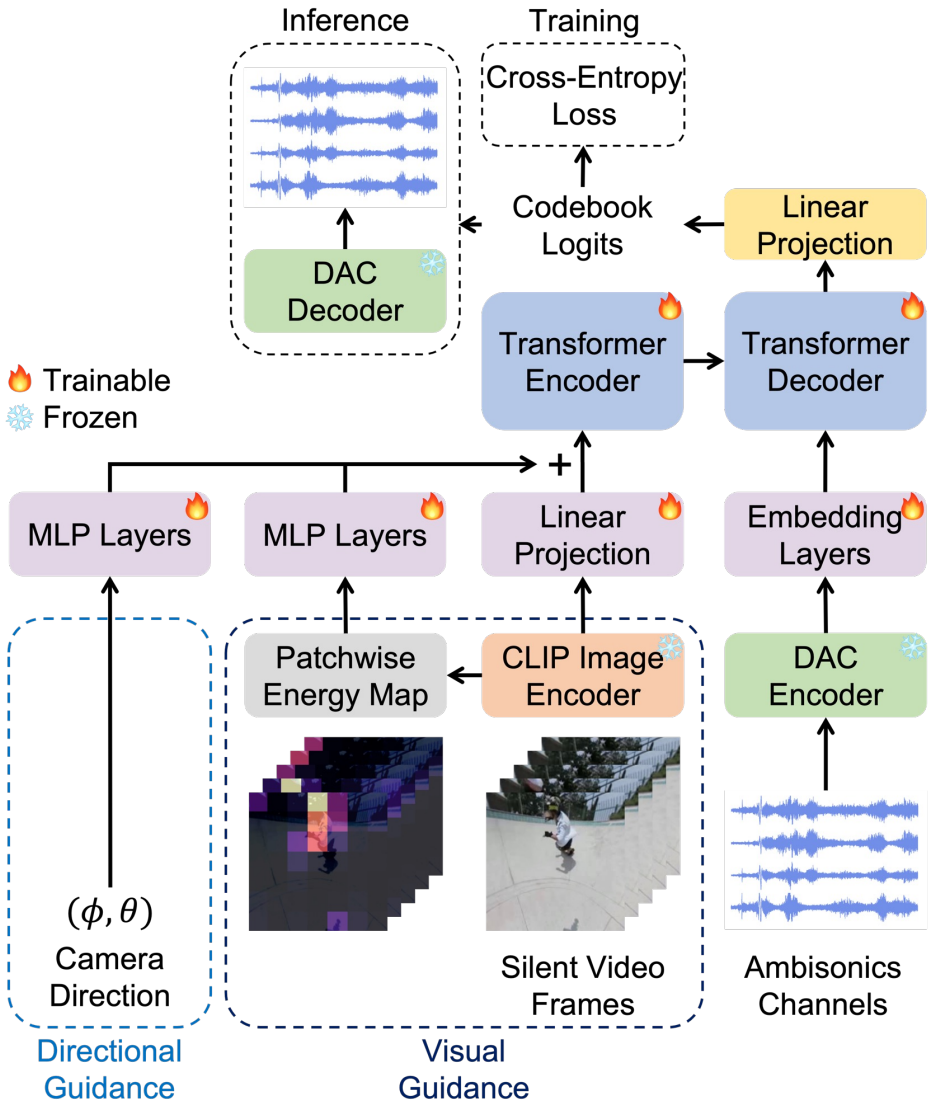
Table 1: Comparison of YT-Ambigen with existing datasets. FoV and 360° respectively denote field-of-view videos and panoramic videos. NS, B, and FOA refer to non-spatial audio, binaural audio, and first-order ambisonics, respectively. (*Number of audio classes < 15)

Dataset	# of Clips	Video Length	Video Type	Audio Type	All Spatial Channels	Open Domain	Audio Generation
Greatest Hits (Owens et al., 2016)	1K	11h	FoV	NS	-	✗	✓
VEGAS (Zhou et al., 2018)	28K	55h	FoV	NS	-	✓*	✓
VAS (Chen et al., 2020b)	13K	24h	FoV	NS	-	✓*	✓
VGGSound (Chen et al., 2020a)	200K	560h	FoV	NS	-	✓	✓
FairPlay (Gao & Grauman, 2019)	2K	5h	FoV	B	✓	✗	✗
OAP (Vasudevan et al., 2020)	64K	26h	360°	B	✓	✗	✗
REC-STEEET (Morgado et al., 2018)	123K	3.5h	360°	FOA	✓	✗	✗
YT-ALL (Morgado et al., 2018)	3976K	113h	360°	FOA	✗	✓	✗
YT-360 (Morgado et al., 2020)	89K	246h	360°	FOA	✗	✓	✗
STARSS23 (Shimada et al., 2024)	0.2K	7.5h	360°	FOA	✓	✗	✗
YT-Ambigen	102K	142h	FoV	FOA	✓	✓	✓

Approach: ViSAGe

- Propose **ViSAGe**
  - Encoder-decoder transformer
  - Encoder: Visual features
  - Decoder: Neural audio codecs

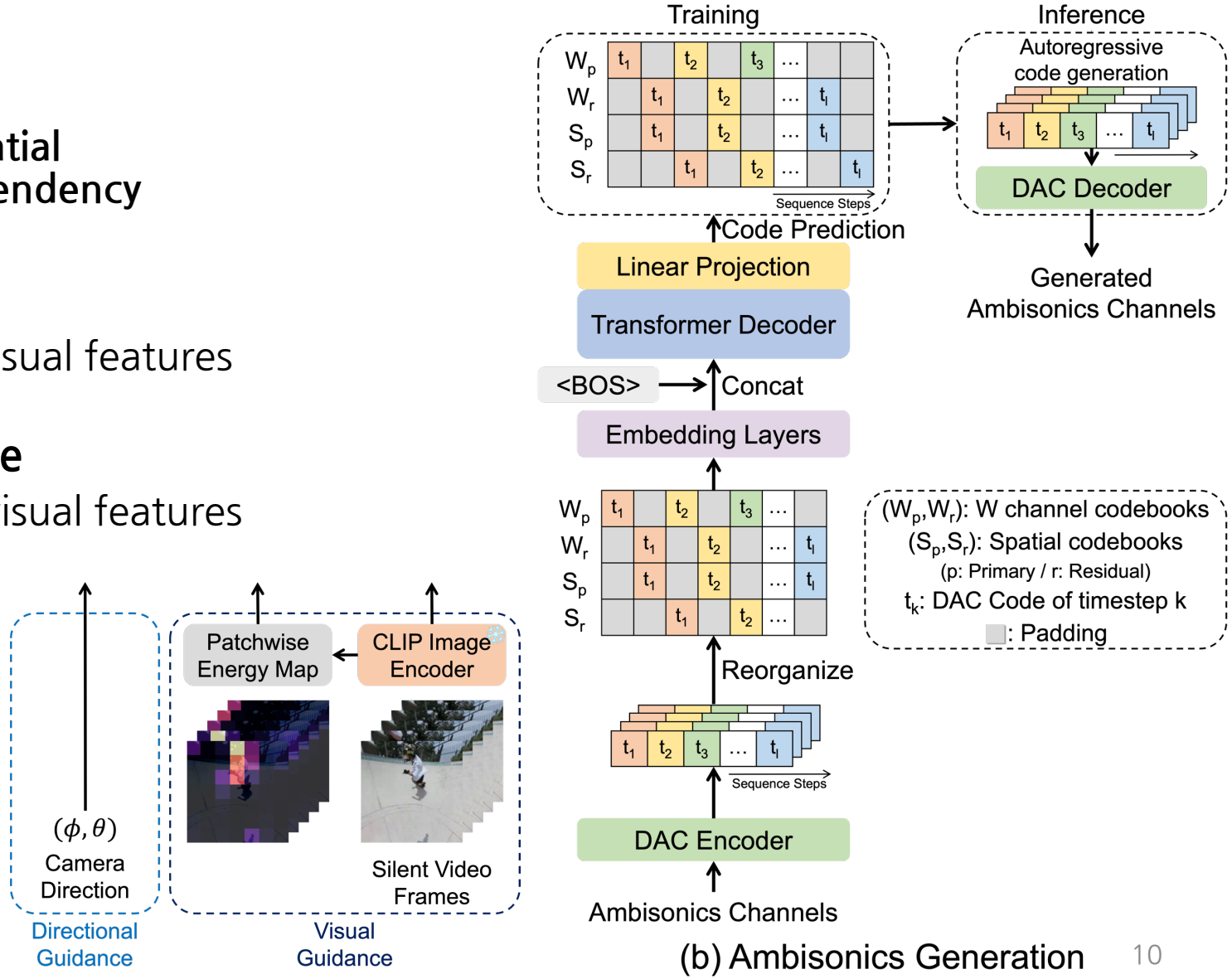
- Conditional Encoding
  - **CLIP features**: Capture semantic content
  - **Patchwise energy map**: Capture fine-grained spatial cues
  - **Direction Embedding:** Control overall spatial directivity



(a) Overall Architecture

Approach: ViSAGe

- Ambisonics Generation
  - **Code Generation Pattern**
    - Effectively model both the **spatial dependency** and **residual dependency**
  - **Rotation Augmentation**
    - Rotate the azimuth by 90
    - Guide model to disentangle visual features from viewing direction
  - **Directional and Visual Guidance**
    - CFG on both directional and visual features to enhance the generation



(b) Ambisonics Generation

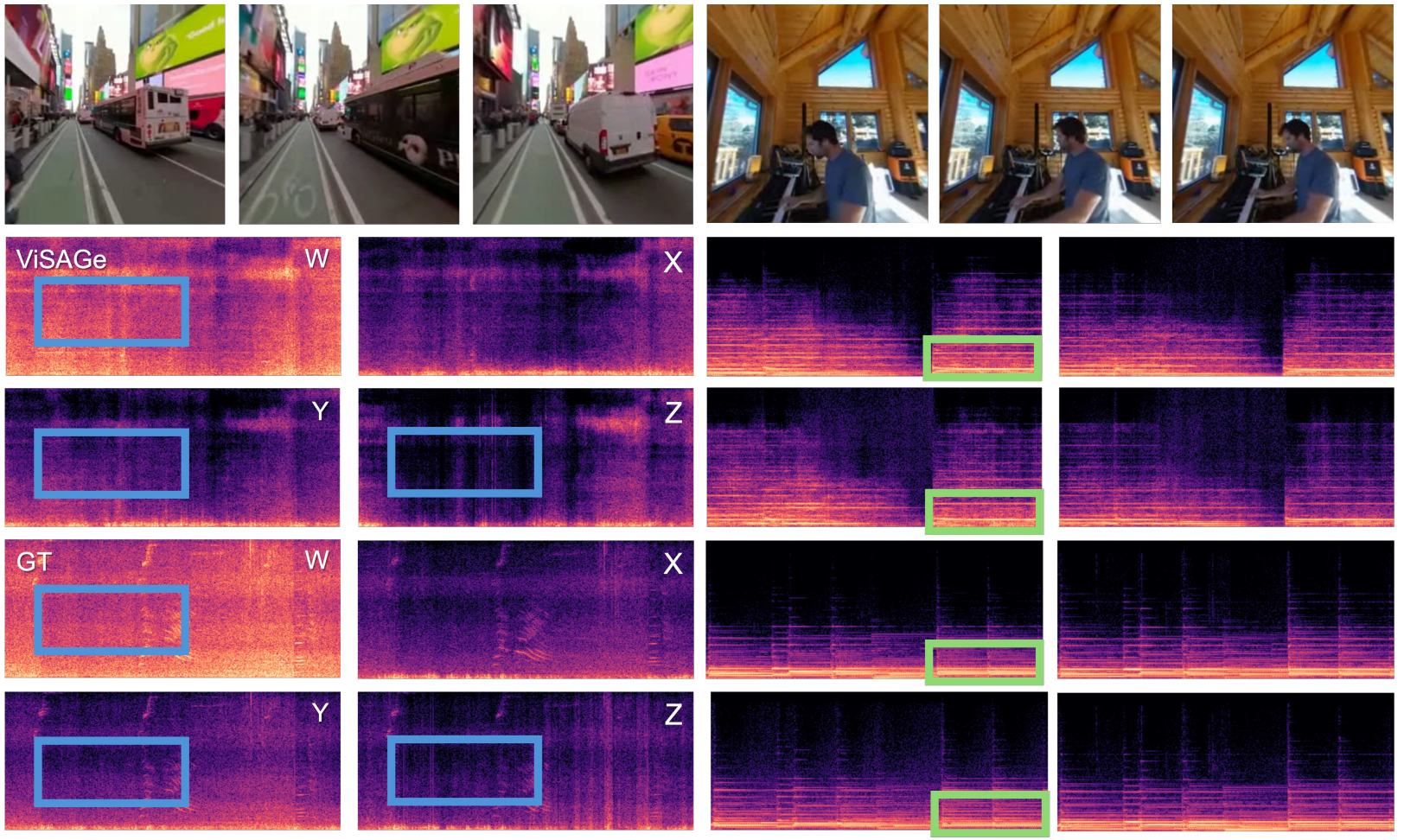
Experiment

- **ViSAGe outperform two-stage baselines** in both semantic and spatial metrics
  - Two-stage baseline: Video-to-audio generation + Audio spatialization

Model		Semantic Metrics			Spatial Metrics					
V2A	Spatialization	$FAD_{dec}\downarrow$	$KLD_{dec}\downarrow$	$FAD_{avg}\downarrow$	CC↑ All	CC↑ 1fps	CC↑ 5fps	AUC↑ All	AUC↑ 1fps	AUC↑ 5fps
Comparison to baseline models										
SpecVQGAN	Ambi Enc.	5.94	2.56	5.62	0.349	0.337	0.322	0.687	0.680	0.670
	Audio Spatial.	6.40	2.43	7.90	0.619	0.587	0.547	0.848	0.828	0.802
Diff-foley	Ambi Enc.	5.68	2.60	5.53	0.349	0.337	0.322	0.687	0.680	0.670
	Audio Spatial.	7.24	2.51	8.76	0.577	0.537	0.494	0.826	0.803	0.777
ViSAGe (Directional)		5.56	2.01	4.76	**0.721**	**0.671**	**0.624**	**0.890**	**0.864**	**0.839**
ViSAGe (Directional & Visual)		**3.86**	**1.71**	**4.20**	0.635	0.584	0.531	0.846	0.819	0.790

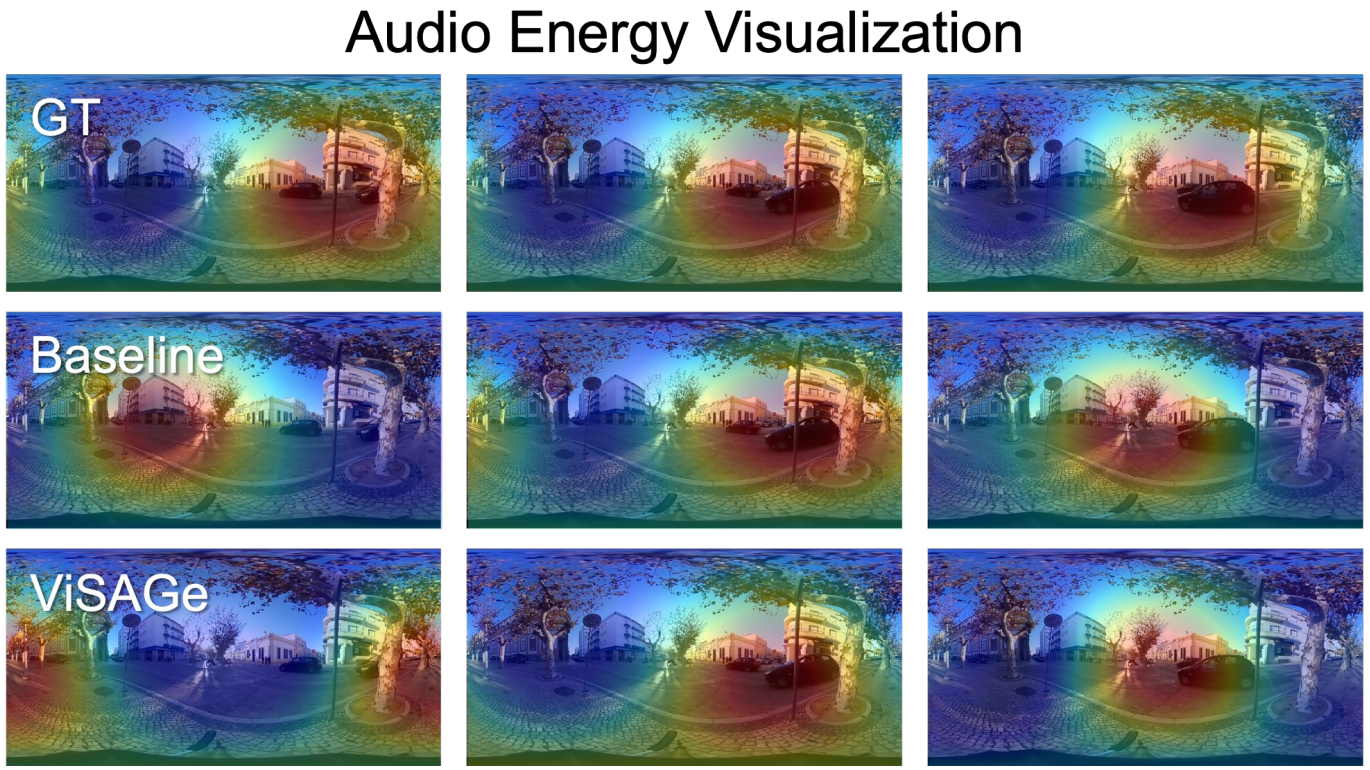
Experiment

- Qualitative Examples



Experiment

- Qualitative Examples



Conclusion

- Propose new task of **video-to-ambisonics generation**
- Introduce **evaluation metrics** and **YT-Ambigen** to support the task
- Proposed method **ViSAGe** outperforms two-stage methods in both spatial and semantic metrics

Thank you!!

Project Page