ViSAGe

Video-to-Spatial Audio Generation

Gen4AVC

ICCV 2025 Workshop

ICLR 2025
(Extended version submitted to IJCV)

ICLR 2025

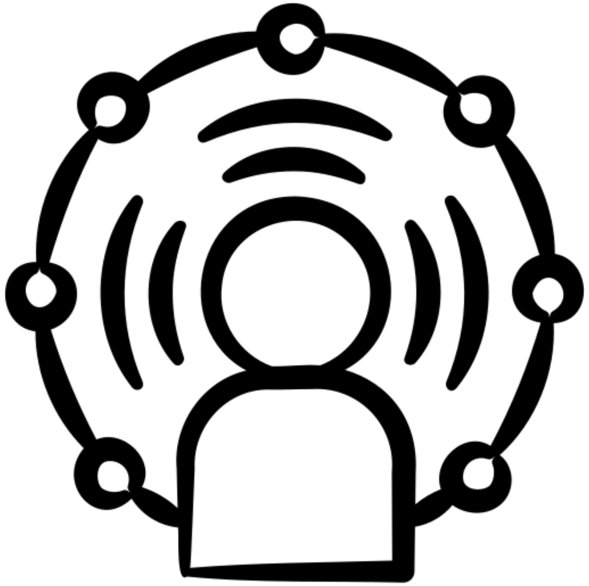
INTERNATIONAL JOURNAL OF COMPUTER VISION

Jaeyeon Kim, Heeseung Yun, Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

A Novel Task

- Can we generate spatial audio for videos?



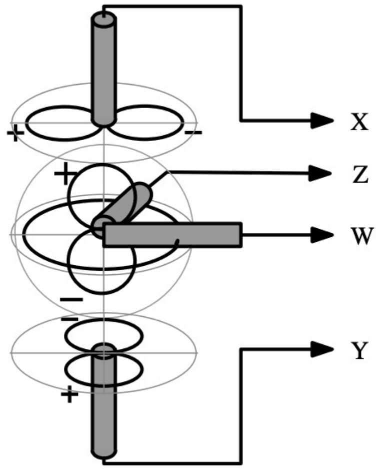
Spatial audio is essential for immersive audio-visual experience

Spatial audio production is expensive

Sound effects are often manually created (e.g., Foley synthesis)

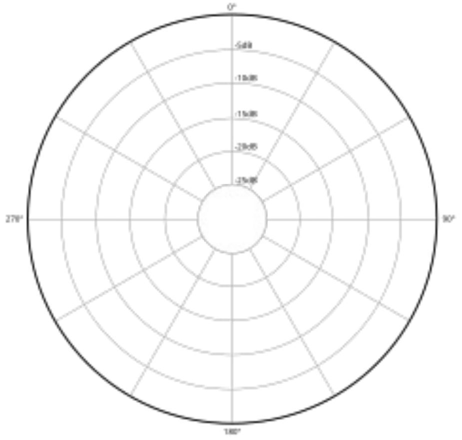
Spatial Audio → First-order Ambisonics (FOA)

- The most basic form of ambisonics (full-sphere surround sound format)
  - 3D surrounding sound format using 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



(a) Native 3D FOA recording

Microphone polar sensitivity



Omnidirectional

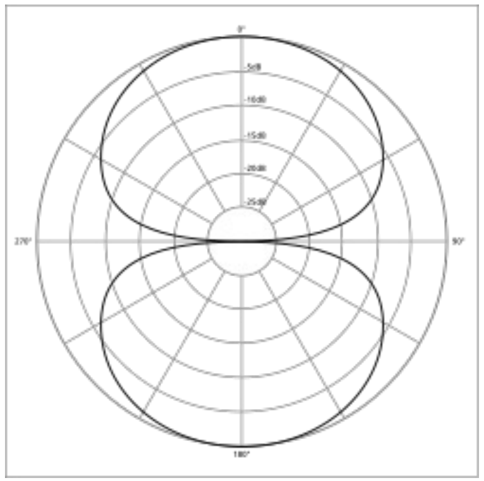
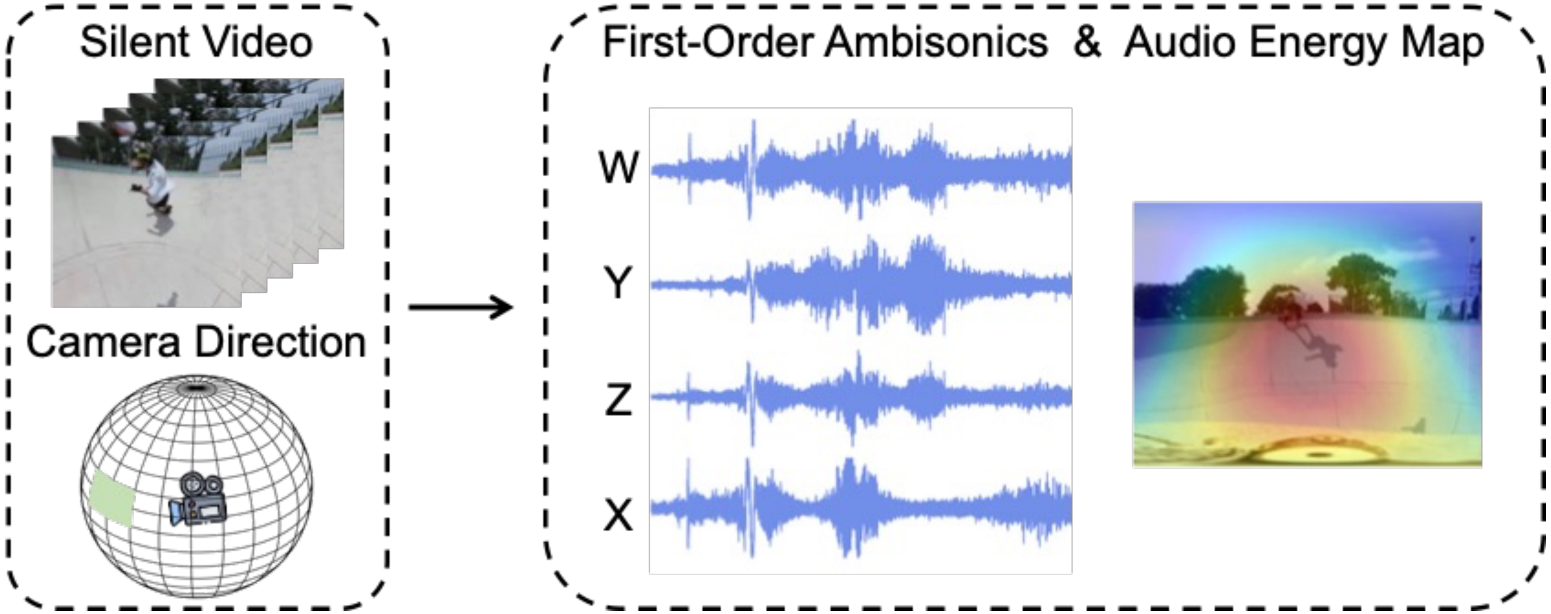


Figure-8 (or Bidirectional)

F. Zotter & M. Frank, Ambisonics: A practical 3D audio theory for recording, studio, production, ..., Springer 2019

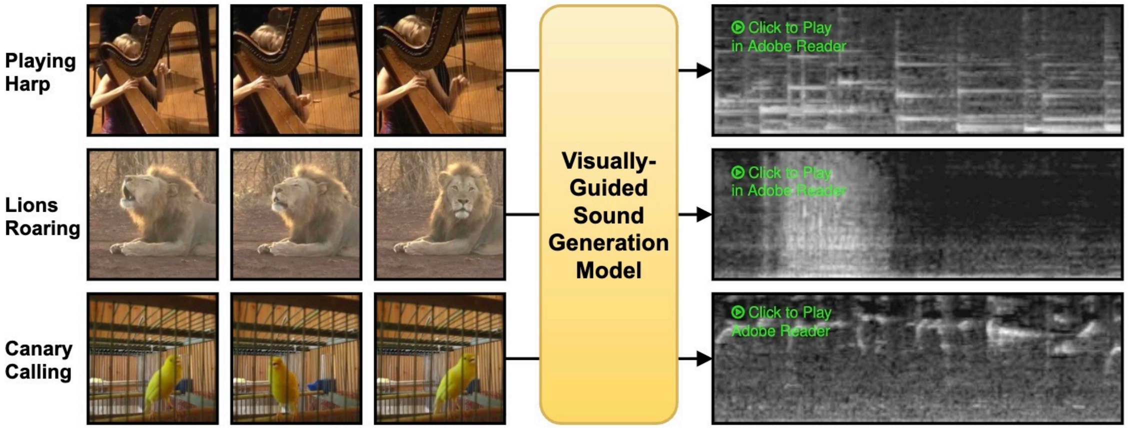
A Novel Task : Video-to-Ambisonics Generation

- Generate FOA given silent field-of-view (FoV) video + camera direction
  - FoV: wider application than panoramic videos
  - However, FoV alone lacks where visual event occur in 3D surroundings
  - → **Camera direction** to represent where visual scene is taking place (i.e., audio should be coherently perceived by a person viewing from a specific direction)

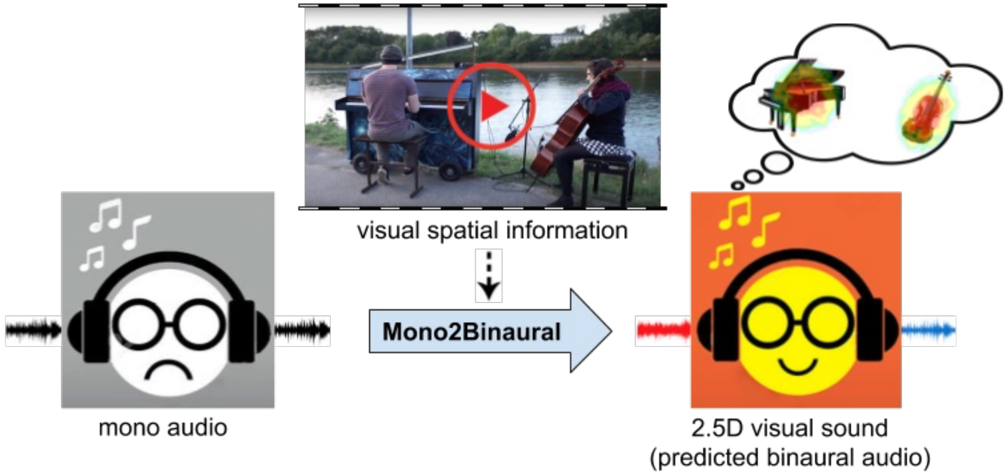


Prior Work

- Can we generate spatial audio with **existing approaches**?



Video-to-Audio Generation
→ generates only mono audio



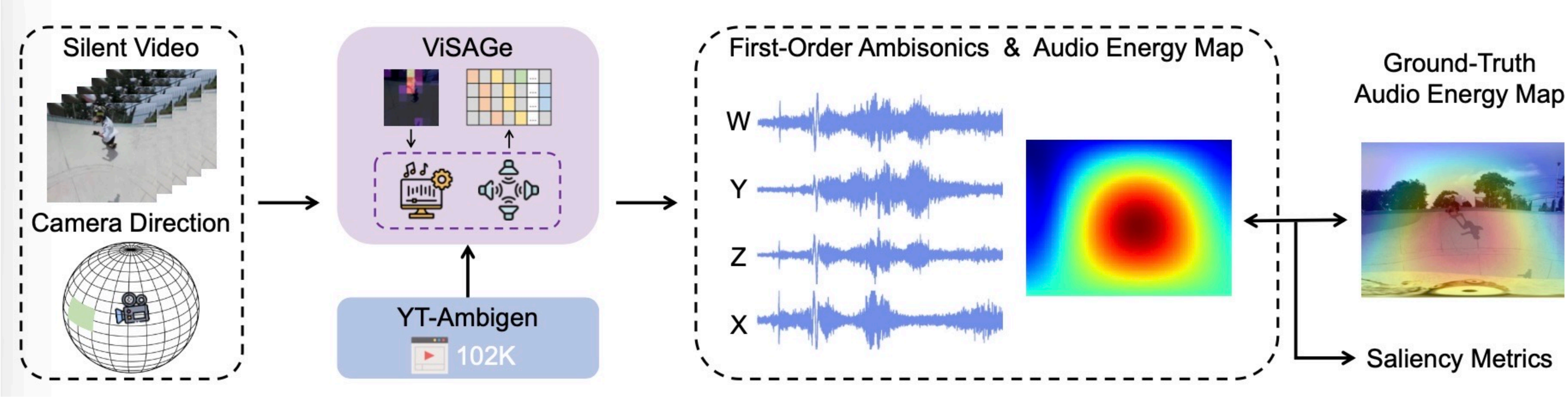
Audio Spatialization
→ require reference mono audio

Combining both approaches may lead to
inaccurate spatialization due to misalignment

V. Iashin et al., PAM:Taming Visually Guided Sound Generation, BMVC 2021
R. Gao et al., 2.5D Visual Sound, CVPR 2019

Contributions

- A novel task: Video-to-ambisonics generation
- A novel dataset: YT-Ambigen
- A novel model: ViSAGe
- A novel evaluation: a mix of semantic/spatial metrics



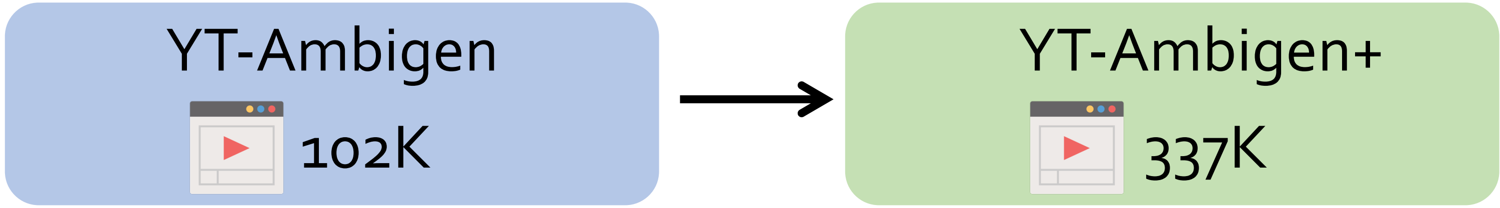
YT-Ambigen Dataset

- 102,364 5-sec clips with FOA and camera direction (ϕ, θ)
  - Sourced from 5.2K panoramic videos with FOA collected from YouTube
  - Existing datasets: No spatial audio or not suitable for audio generation

	Dataset	# Clips	Length	Audio Type	Audio Gen
Audio gen but no spatial	VEGAS	28K	55h	Non-Spatial	✅
	VAS	13K	24h	Non-Spatial	✅
	VGGSound	200K	560h	Non-Spatial	✅
Spatial but no audio gen	FairPlay	2K	5h	Binaural	❌
	OAP	64K	26h	Binaural	❌
	YT-360	89K	246h	FOA	❌
	STARSS23	0.2K	7.5h	FOA	❌
	YT- Ambigen	**102K**	**142h**	**FOA**	✅

YT-Ambigen+ Dataset (IJCV Extension)

- Harvest all accessible accessible panoramic videos with ambisonics
  - Filter invalid videos and apply similar post-processing to obtain 5s clips
  - Total **336584** clips, **19265** human validated test split
  - PaSST classification: top-1 labels cover 393/527 AudioSet labels and 930/1,000 ImageNet labels



Dataset	# of Clips	Video Length	Video Type	Audio Type	All Spatial Channels	Open Domain	Audio Generation
YT-Ambigen	102K	142h	FoV	FOA	✓	✓	✓
YT-Ambigen+	337K	468h	FoV	FOA	✓	✓	✓

Evaluation Metrics

- How can we evaluate the generated spatial audio?
  - Semantic metrics: the generated audio should convey the content
  - Spatial metrics: all channels should create a spatially coherent sound field

- Semantic metrics
  - Adopt two widely used Fréchet Audio Distance (FAD) and Kullback-Leibler Divergence (KLD)
  - FAD: perceptual quality and fidelity of the generated audio
  - KLD: KLD btw. the class distributions of the generated and reference audio (how effectively the generated audio captures the intended audio concepts)
  - Computed $\mathbf{FAD_{dec}}$, $\mathbf{KLD_{dec}}$ for decoded mono audio (from FOA)
  - Compute the average of $\mathbf{FAD_{avg}}$ from each channel

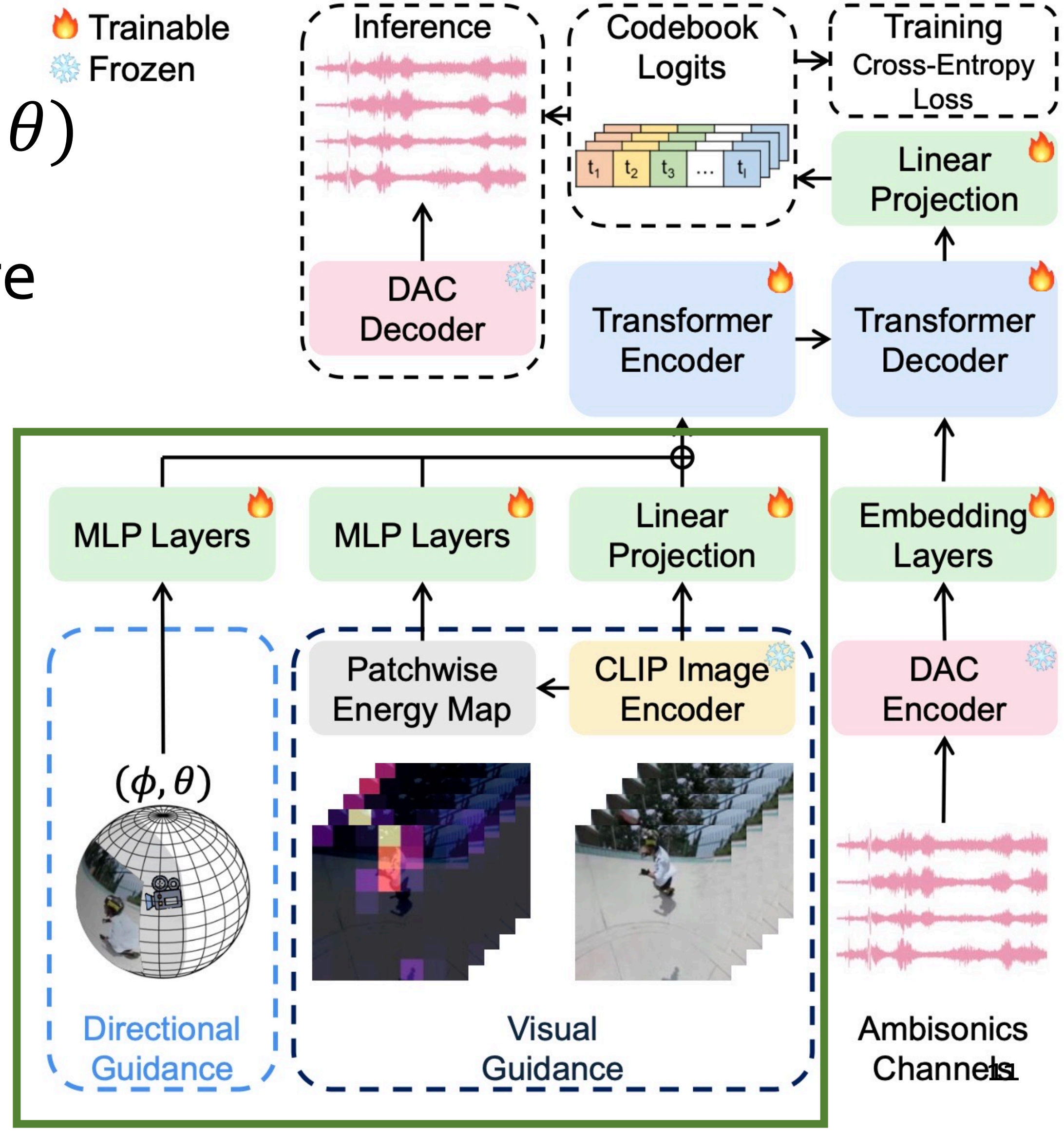
Evaluation Metrics

- Spatial metrics
  - FOA can be used to generate an audio energy map over the sphere
  - Then compute visual saliency metric between generated and ground-truth audio energy map
  - Correlation Coefficient (CC) and Area Under Curve (AUC)

CC: 0.632, AUC: 0.833

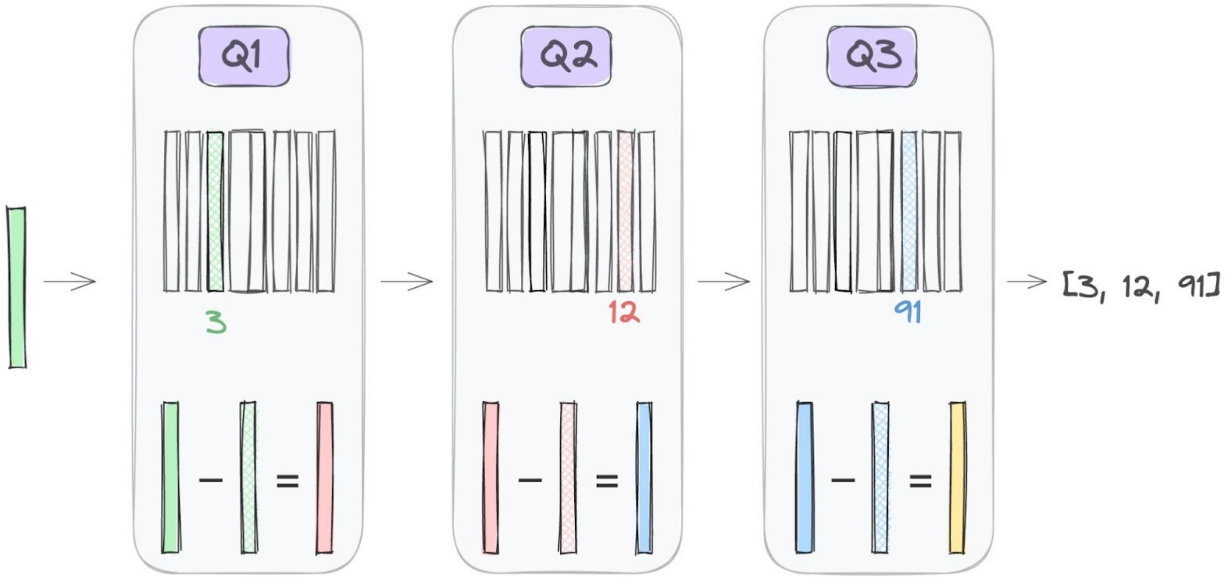
ViSAGe Approach – Encoding

- Generate FOA $A = (W, X, Y, Z)$ for video frames V and a camera direction $D = (\phi, \theta)$
- Encoder-decoder Transformer architecture
  - 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



ViSAGe Approach – Decoding

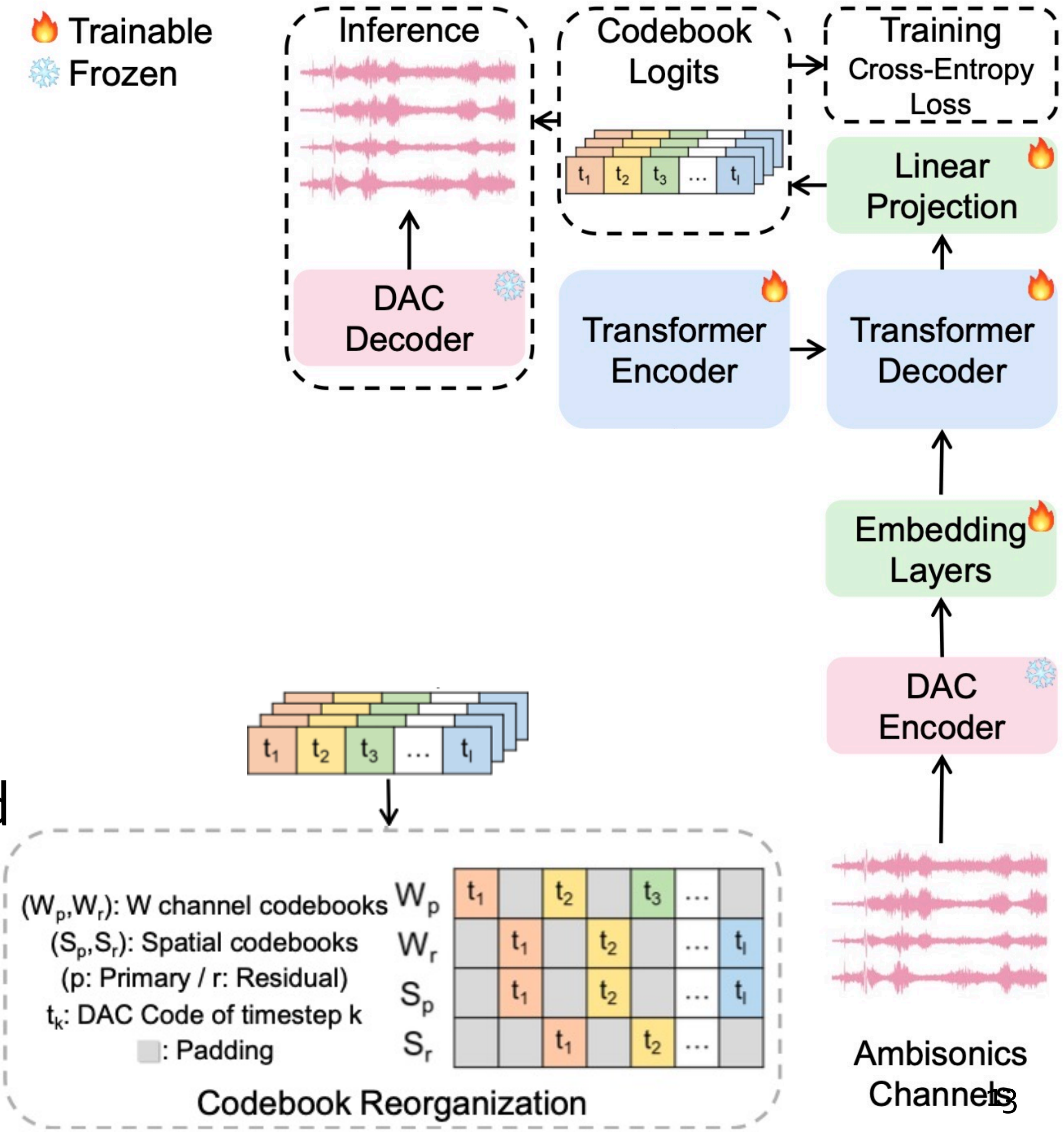
- Descript Audio Codec (DAC) encoder
  - A SOTA neural codec for open-domain audio via residual vector quantization (RVQ)
- RVQ
  - Progressively finer approximation to high-dim vectors by a cascade of codebooks
  - The primary codebook: first-order quantization of the input vector
  - The *residuals*: further quantized using a secondary codebook



R. Kumar et al., High-Fidelity Audio Compression with Improved RVQGAN, NeurIPS 2023
https://www.assemblyai.com/blog/what-is-residual-vector-quantization

ViSAGe Approach – Decoding

- DAC encoder
  - Each FOA channel → an audio code matrix
- Code generation patterns
  - Better model both the spatial dependency and residual dependency (as later codebooks depend on earlier ones by RVQ)
- Training
  - The cross-entropy loss between the predicted codes and the GTs

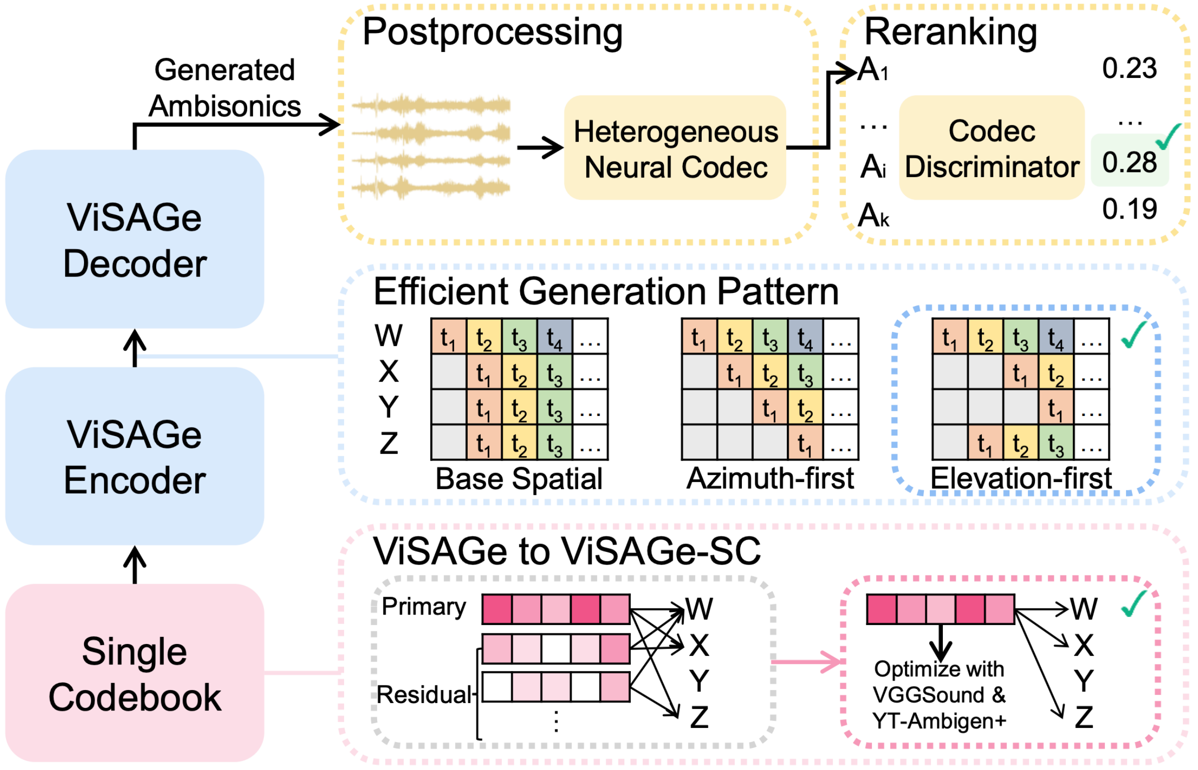


Efficient Generation with ViSAGe-SC (IJCV Extension)

- Core bottleneck of ViSAGe: RVQ codebooks of DAC
  - N *4 codes per timestep (N : # codebooks ~ 9)
  - High computation complexity and susceptible to artifacts

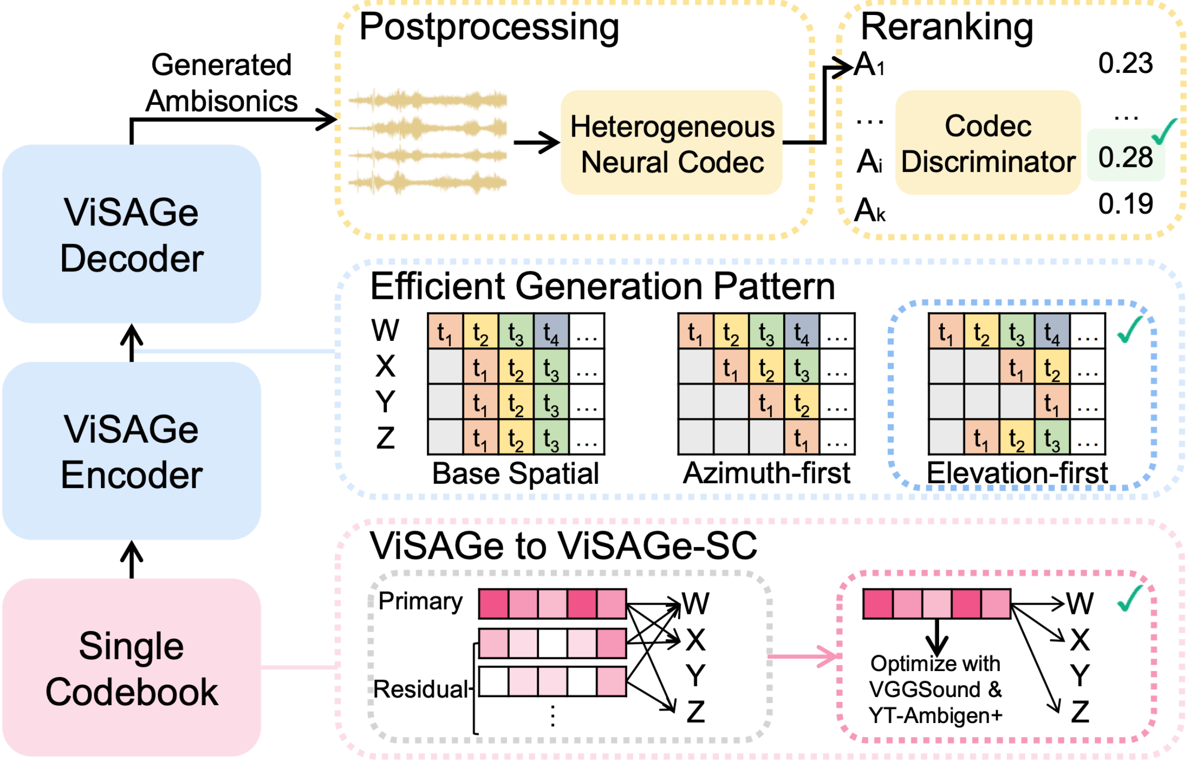
- ViSAGe-SC (Single codebook)
  - Let's use one codebook per channel!
  - 4 codes per timestep, but limitation in audio fidelity and quality

- How can we improve audio quality while using less codebooks?
  - 9× fewer codes per timestep, 4× faster training and 3-5× faster inference



Efficient Generation with ViSAGe-SC

- 1. Better single codebook codec
  - Optimize UniCodec on both VGGSound and YT-Ambigen+
- 2. No more residual dependency! Let's generate codes efficiently
  - Reduce the generation length by half
- 3. Improvement after generation
  - Chaining heterogeneous codec
  - Candidate reranking

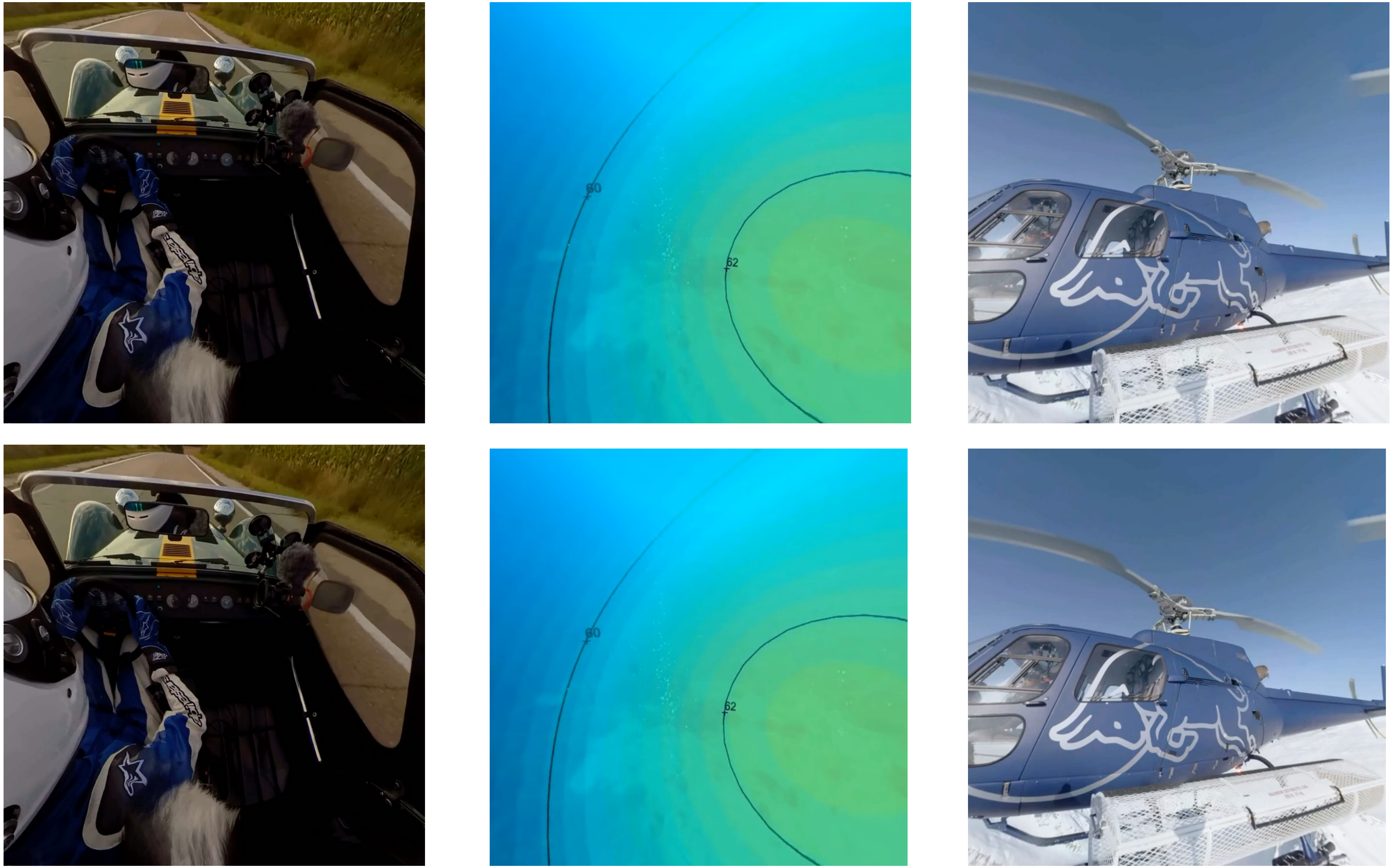


Y. Jiang et al., UniCodec: Unified Audio Codec with Single Domain-Adaptive Codebook, ACL 2025

Examples of Generated Sound

GT

ViSAGe



https://jaeyeonkim99.github.io/visage/

Experiments

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

Model		Semantic Metrics			Spatial Metrics					
					CC↑			AUC↑		
V2A	Spatialization	$\text{FAD}_{\text{dec}}\downarrow$	$\text{KLD}_{\text{dec}}\downarrow$	$\text{FAD}_{\text{avg}}\downarrow$	All	1fps	5fps	All	1fps	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

Experiments

- ViSAGe-SC significantly outperforms ViSAGe
  - Single codebook < Residual codebook
  - Performance improves + (1) finetuning + (2) postprocessing + (3) augmentation
  - 5× faster in training and 3–5× faster in inference

Model	Semantic Metrics ↓				Spatial Metrics ↑					
					CC			AUC		
	FAD_{dec}	KLD_{dec}	FAD_{avg}	E-L1	All	1fps	5fps	All	1fps	5fps
(a) Ablation from ViSAGe to ViSAGe-SC										
ViSAGe	5.72	2.56	5.35	-	0.418	0.378	0.335	0.731	0.711	0.687
w/ Single Codebook (2025)	5.83	2.43	10.66	-	0.492	0.438	0.369	0.786	0.757	0.719
w/ Single Codebook (Ours)	4.81	**2.39**	6.84	-	0.491	0.436	0.368	0.782	0.753	0.716
+Postprocessing	4.25	2.47	**4.83**	-	0.488	0.435	0.368	0.781	0.753	0.717
+Rotation Augmentation	**4.23**	2.48	**4.83**	-	**0.535**	**0.474**	**0.402**	**0.815**	**0.782**	**0.741**

Experiments

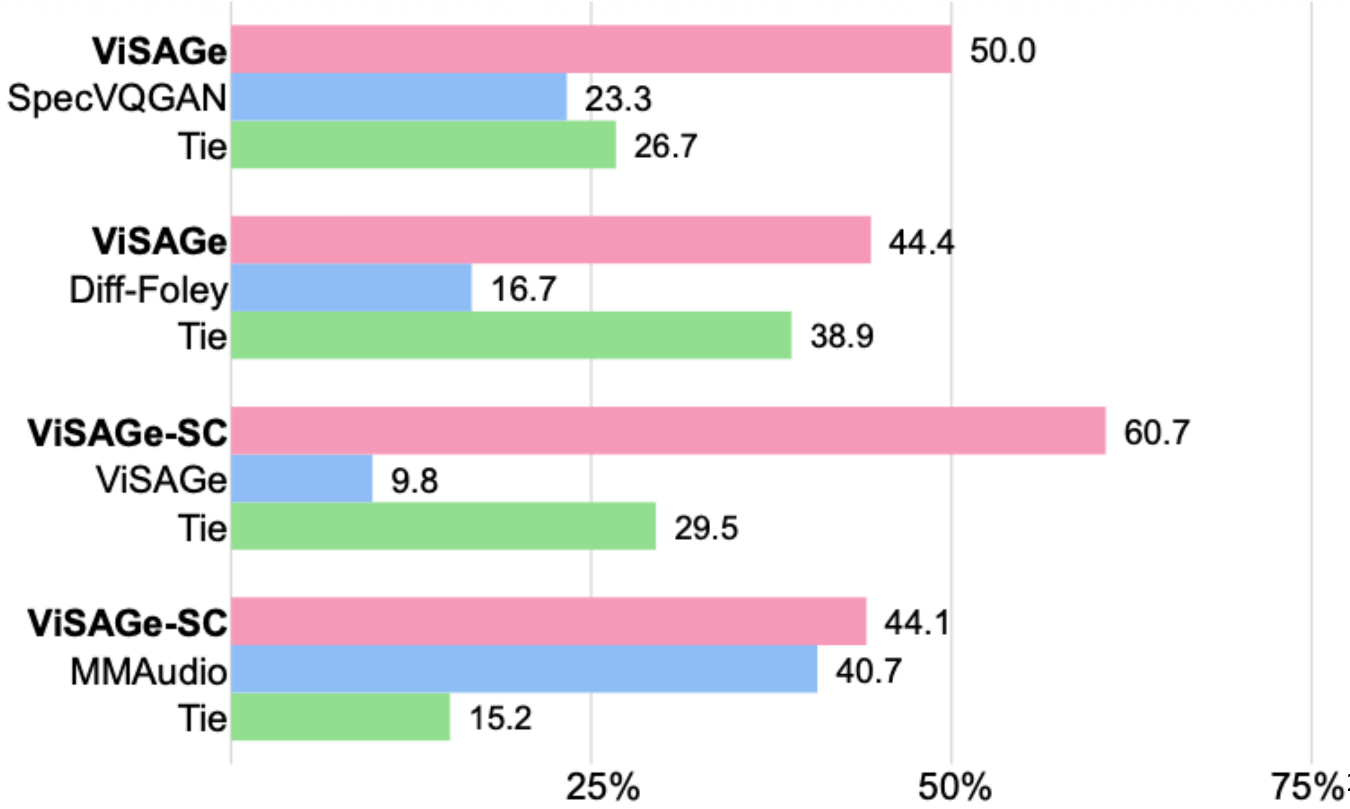
- ViSAGe-SC outperforms all other models
  - Fall short in semantic metrics compared to MMAudio
  - → Promising future direction on enhancing semantic modeling for spatial audio generation

Model		Semantic Metrics ↓			Spatial Metrics ↑						
						CC			AUC		
		FAD_{dec}	KLD_{dec}	FAD_{avg}	E-L1	All	1fps	5fps	All	1fps	5fps
(d) Comparison to Baseline Models											
SpecVQGAN	Ambi Enc.	6.41	3.21	5.08	0.128	0.280	0.266	0.251	0.653	0.645	0.637
	Audio Spatial.	6.23	2.99	7.46	**0.103**	0.499	0.480	0.446	0.784	0.772	0.750
Diff-Foley	Ambi Enc.	5.71	2.92	5.46	0.108	0.280	0.266	0.251	0.653	0.646	0.637
	Audio Spatial.	6.61	2.85	8.82	0.104	0.469	0.447	0.411	0.763	0.749	0.727
MMAudio*	Ambi Enc.	5.21	2.36	**2.99**	0.197	0.280	0.266	0.251	0.652	0.645	0.637
	Audio Spatial.	**3.26**	**2.29**	3.95	0.109	0.503	0.485	0.452	0.788	0.776	0.754
ViSAGe-SC		4.23	2.48	4.83	0.106	0.535	0.474	0.402	0.815	0.782	0.741
ViSAGe-SC (Directional Guidance Only)		5.31	2.73	5.71	**0.103**	**0.643**	**0.584**	**0.506**	**0.874**	**0.843**	**0.801**

H. K. Cheng et al., MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis, CVPR 2025

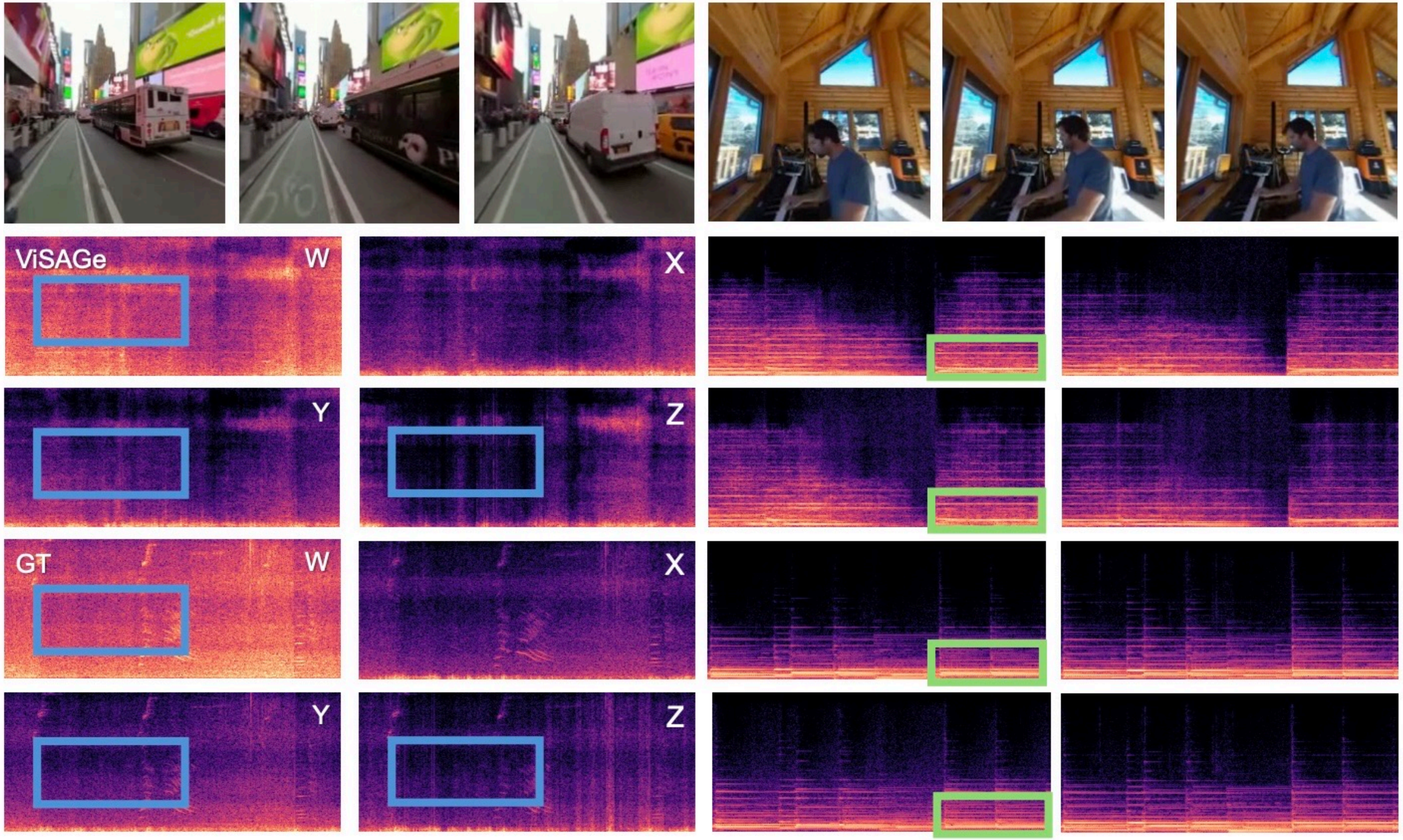
Qualitative Results

- Human evaluation on two-sample hypothesis testing
  - A total of 16 participants, each evaluating up to 30 randomly selected videos
  - The overall preference of two different samples (naturalness, relevance to the video, and the perceived spatial effects)



Qualitative Results

- Linear spectrograms



Blue boxes: ViSAGe can capture differences btw. spatial channels

Green boxes: ViSAGe generates semantically plausible events

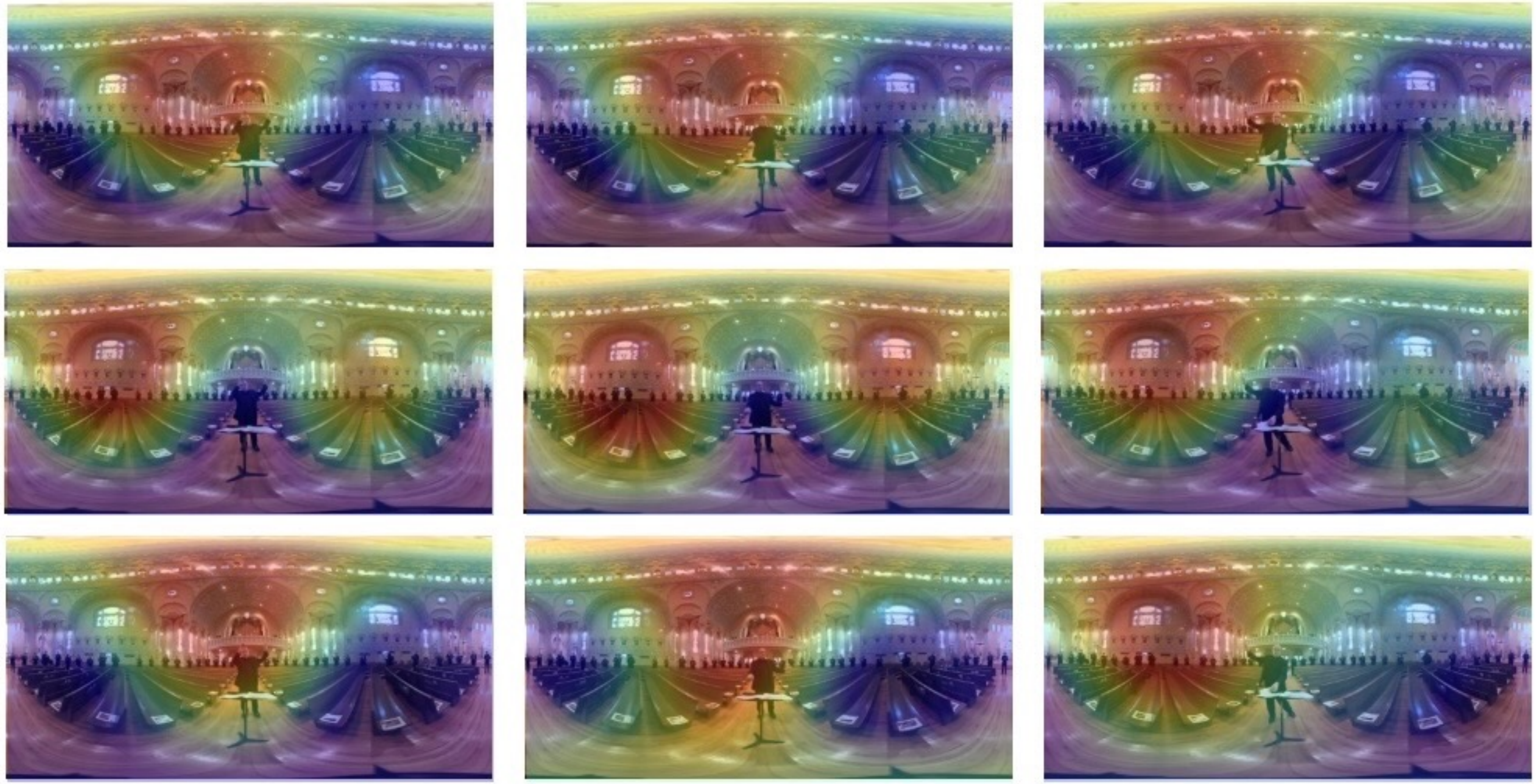
Qualitative Results

- Audio energy visualization

GT

Baseline

ViSAGe



Takeaways

- Propose a new task and datasets of video-to-ambisonics generation
  - Introduce YT-Ambigen and evaluation metrics to support the task
  - Introduce YT-Ambigen+ with 3x larger in scale with a manually verified test set
- Propose a new method ViSAGe and ViSAGe-SC
  - ViSAGe outperforms two-stage methods in both spatial and semantic metrics
  - ViSAGe-SC achieves better performance, 4× faster training, and 3–5× faster inference compared to ViSAGe
  - ViSAGe to ViSAGe-SC: (1) better Single Codebook Codec, (2) efficient generation pattern, (c) postprocessing and Reranking

Project Page