Diff-SAGe: End-to-End Spatial Audio Generation using Diffusion Models

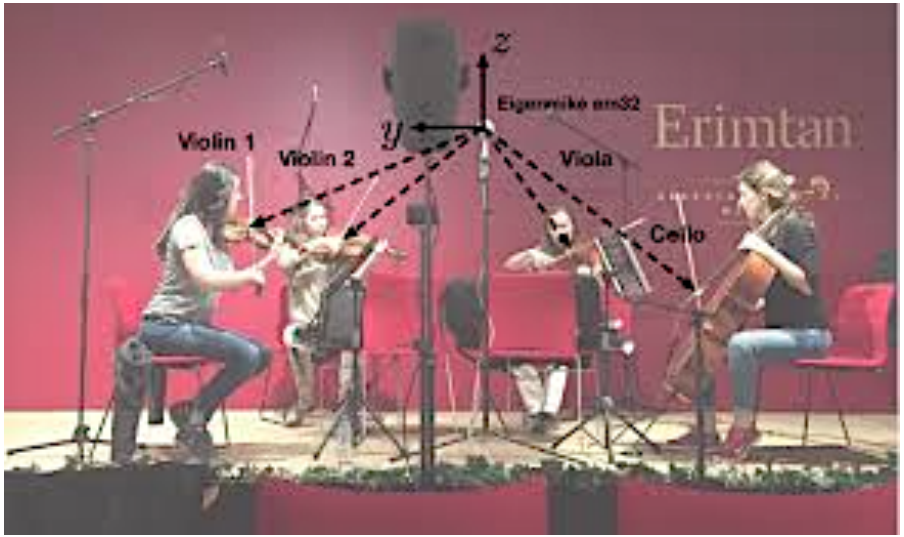
Saksham Singh Kushwaha[1,2,*], Jianbo Ma[2], Mark Thomas[2], Yapeng Tian[1], Avery Bruni[2]

[1]University of Texas at Dallas, [2]Dolby Laboratories

*(*Work done during internship at Dolby)*

Spatial Audio is important

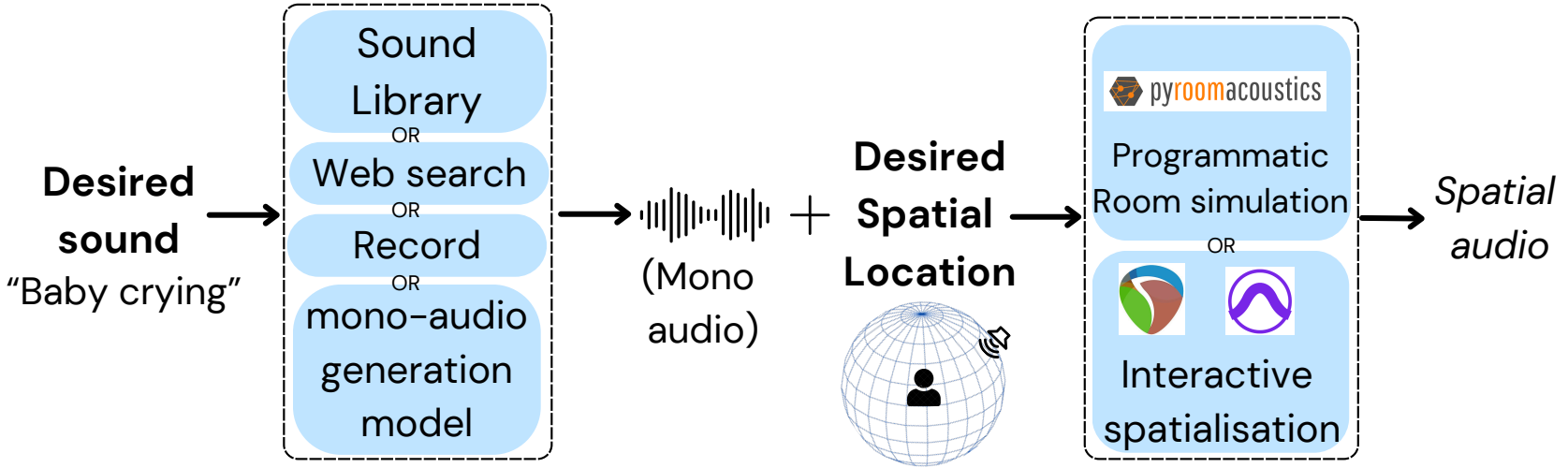
Spatial Audio. Generated using ChatGPT-4o [1]



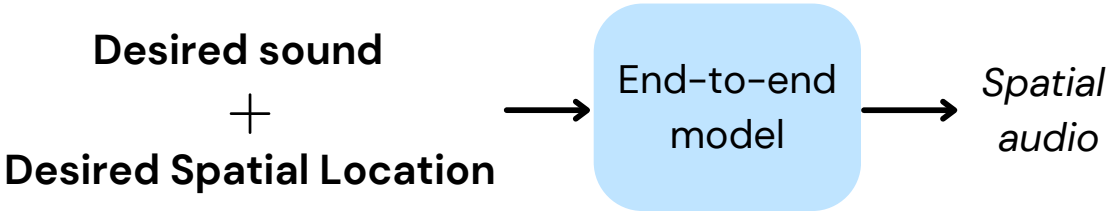
Setup for the recording of the classical quartet [2]

[1] OpenAI, “GPT-4 Technical Report”
[2] Coteli et. al, “Multiple sound source localization with steered response power density and hierarchical grid refinement”

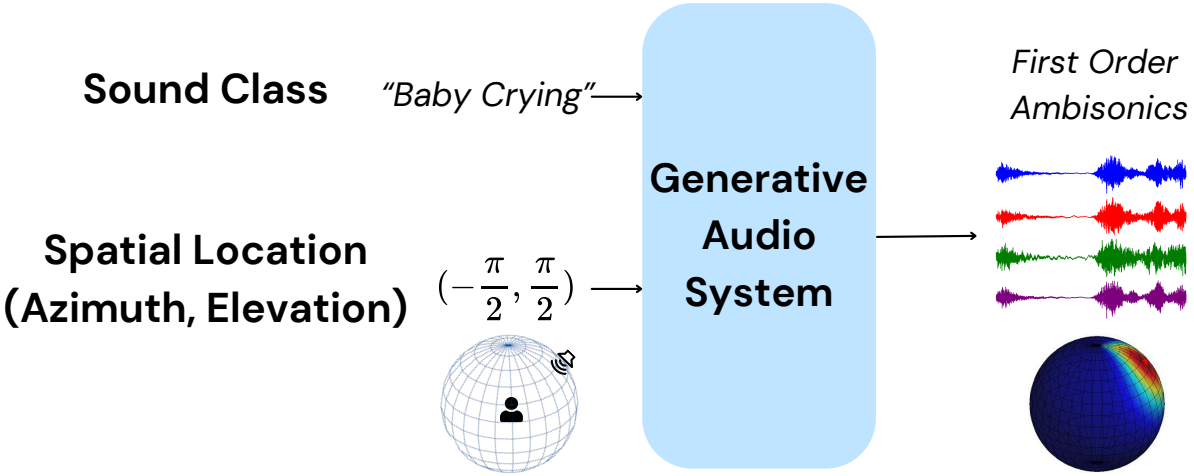
Traditional spatial audio-generation



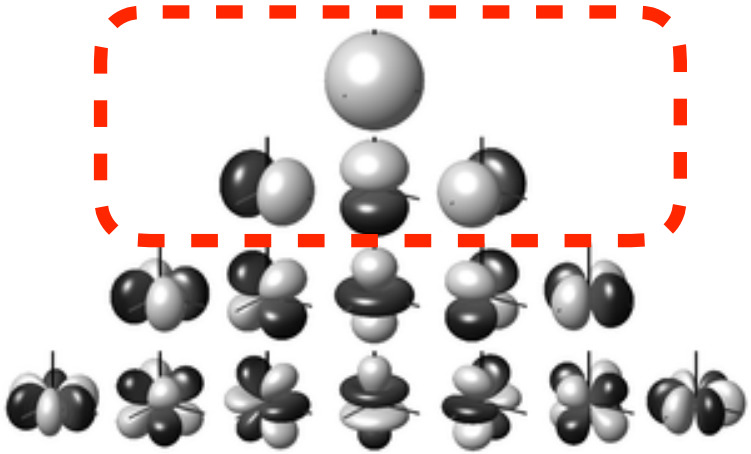
End-to-end spatial audio generation



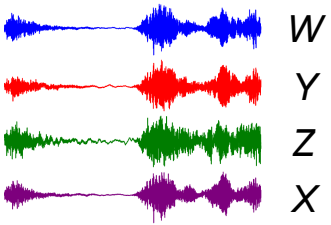
New Task: Ambisonics generation



Background: First Order Ambisonics



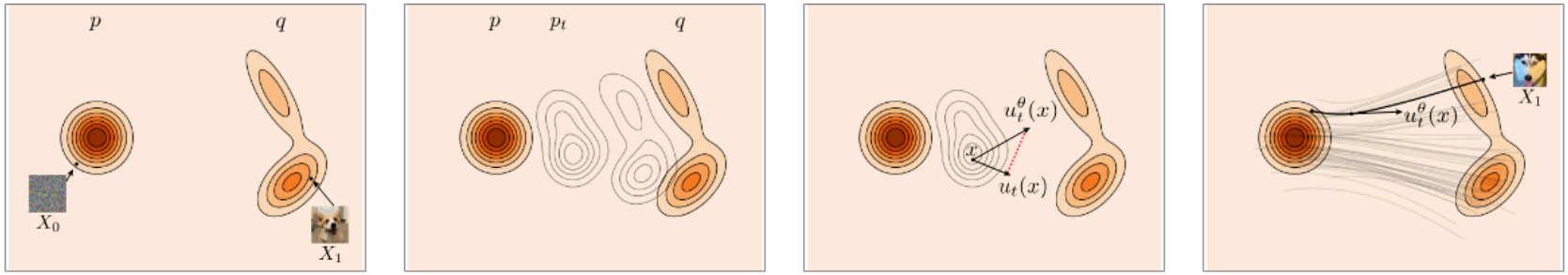
Visual representation of Ambisonics B-format [1]



Example FOA

[1] Zotter at. al., “Energy-Preserving Ambisonic Decoding”

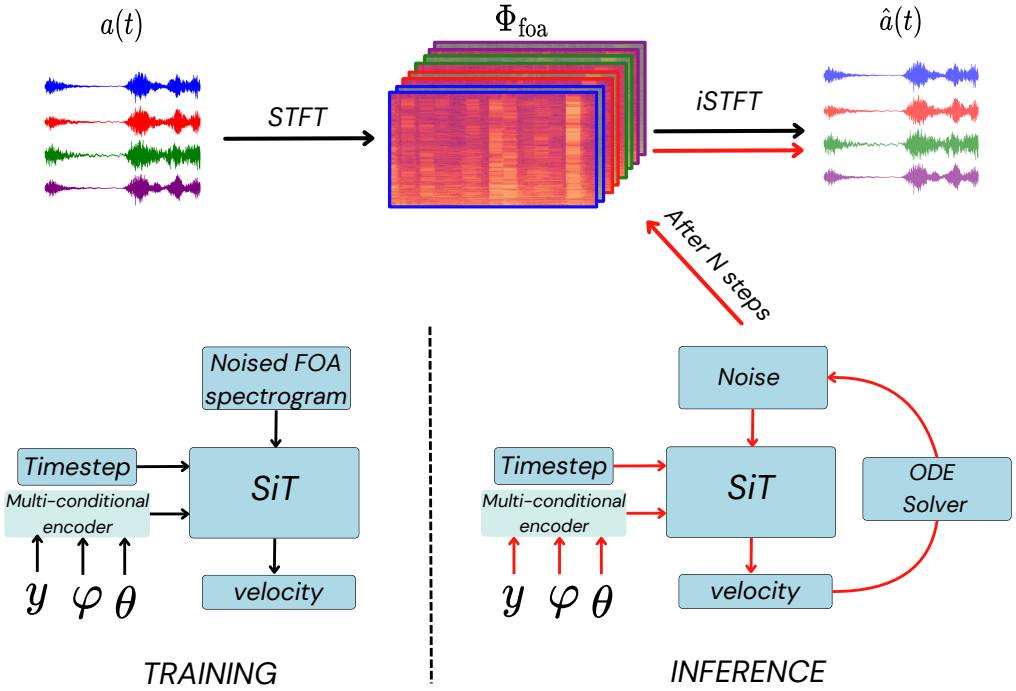
Background: Flow Matching



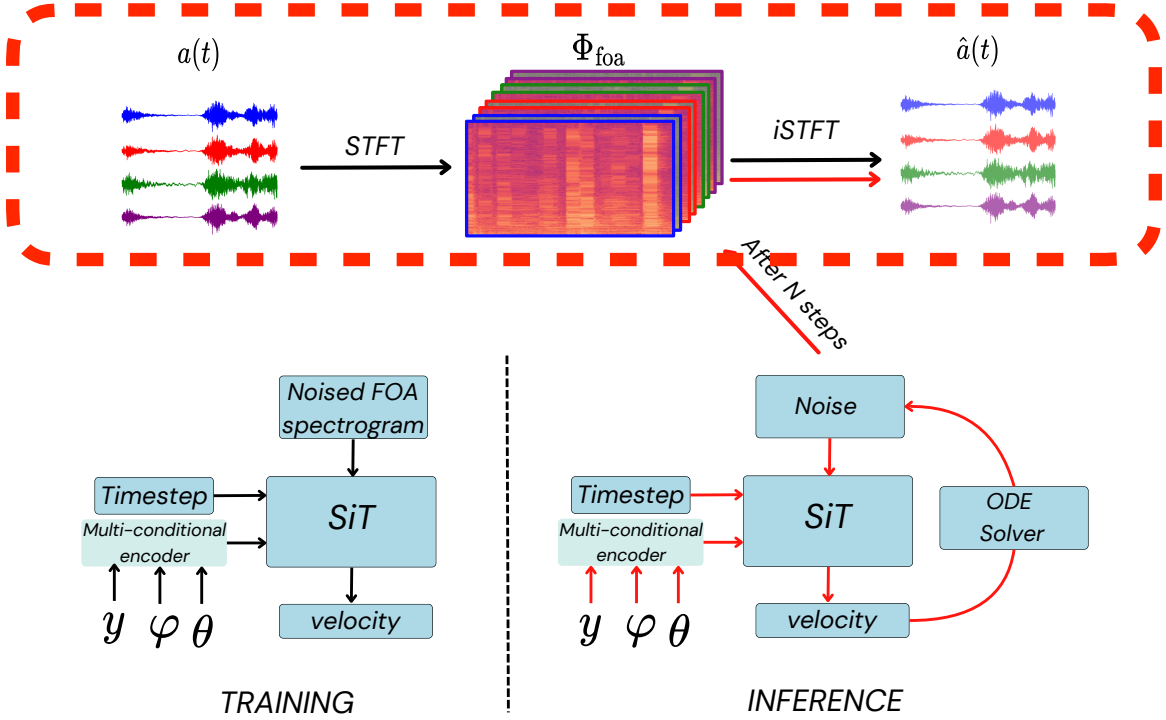
(a) Data. **(b)** Path design. **(c)** Training. **(d)** Sampling.

[1] Lipman et al., "Flow Matching Guide and Code"

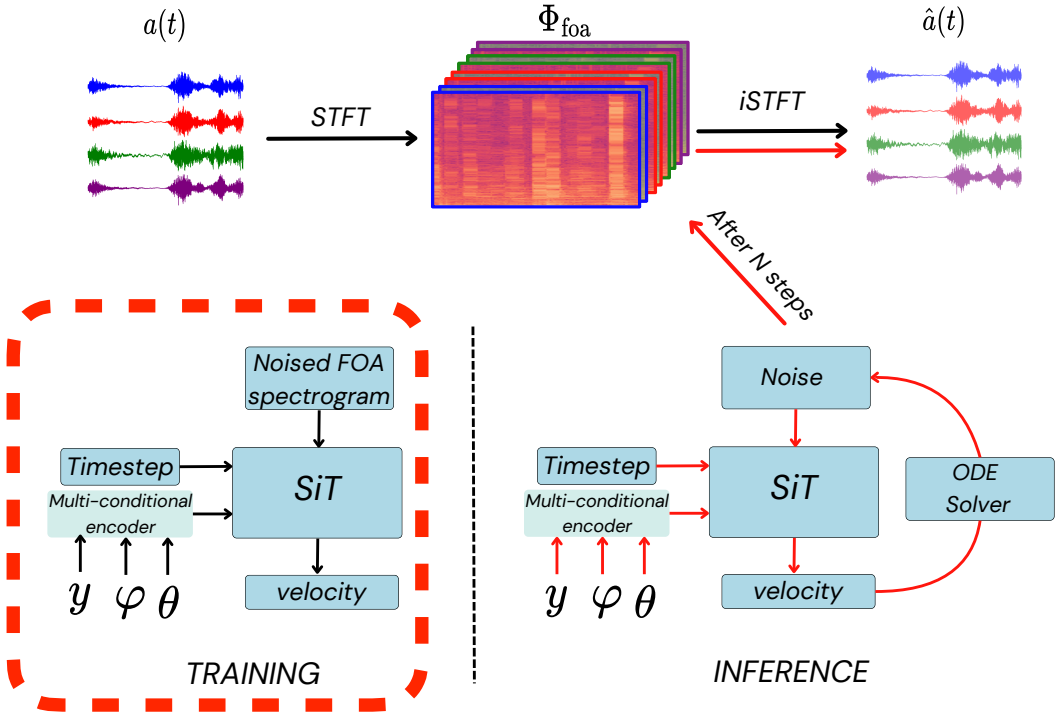
Approach


$a(t)$
STFT
Φ_{foa}
iSTFT
$\hat{a}(t)$
After N steps
Noised FOA spectrogram
Timestep
Multi-conditional encoder
SiT
velocity
y φ θ
TRAINING
Noise
Timestep
Multi-conditional encoder
SiT
ODE Solver
velocity
y φ θ
INFERENCE

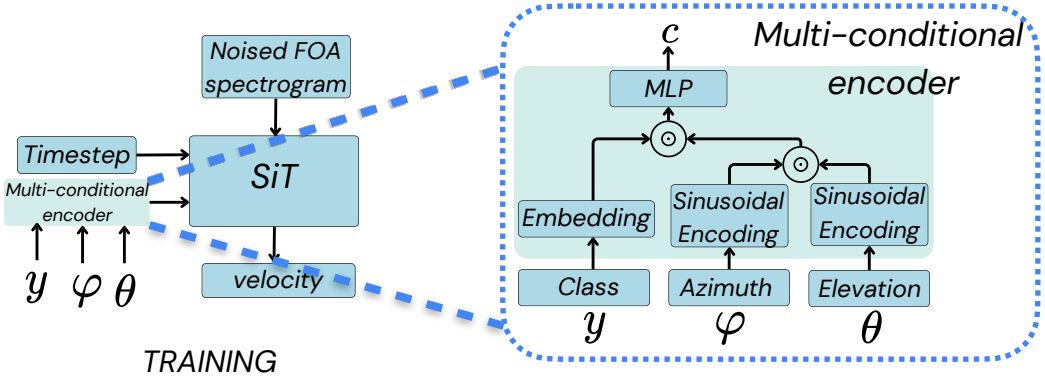
Approach: FOA representation



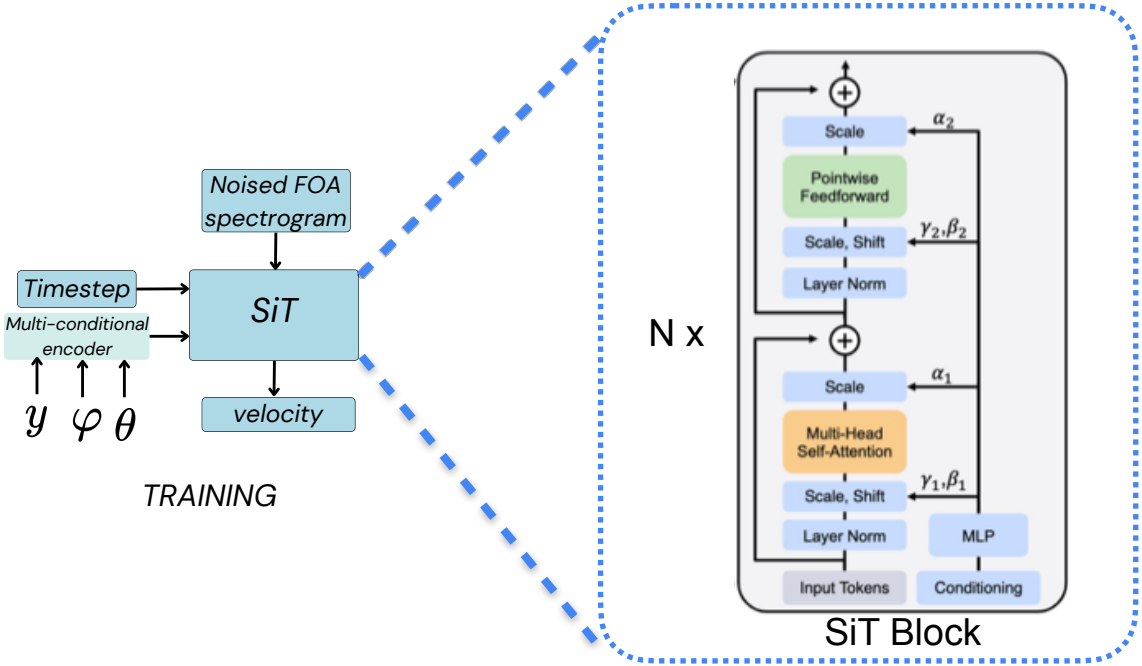
Approach: Training



Approach: Training

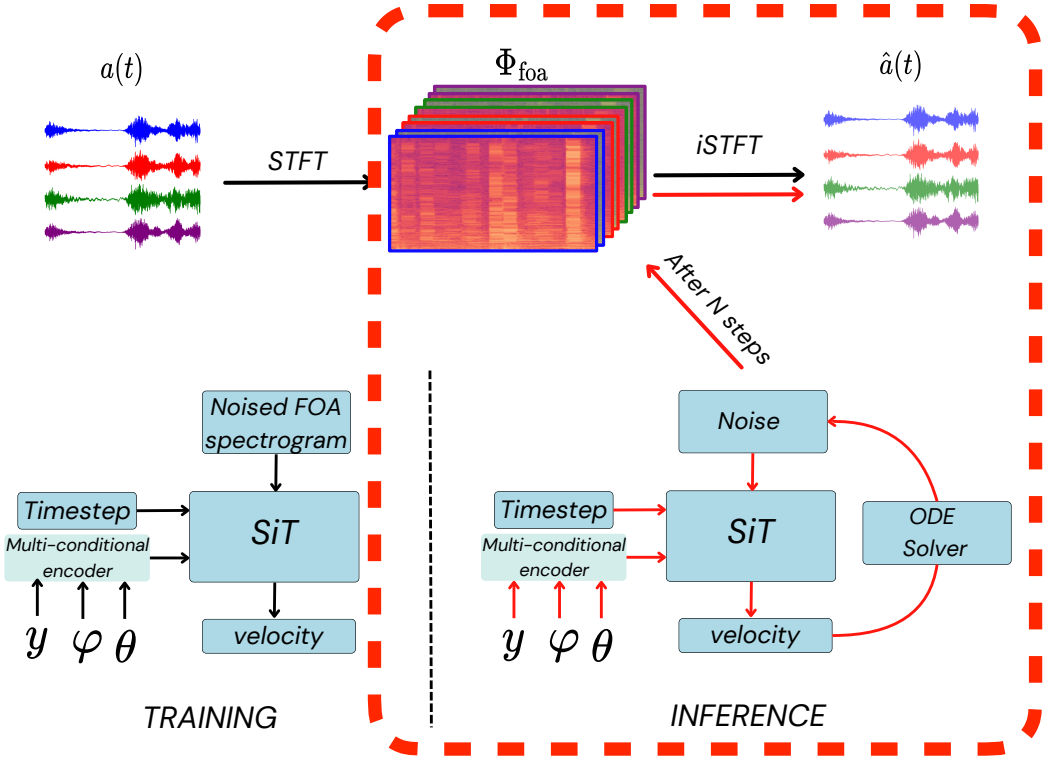


Approach: Training

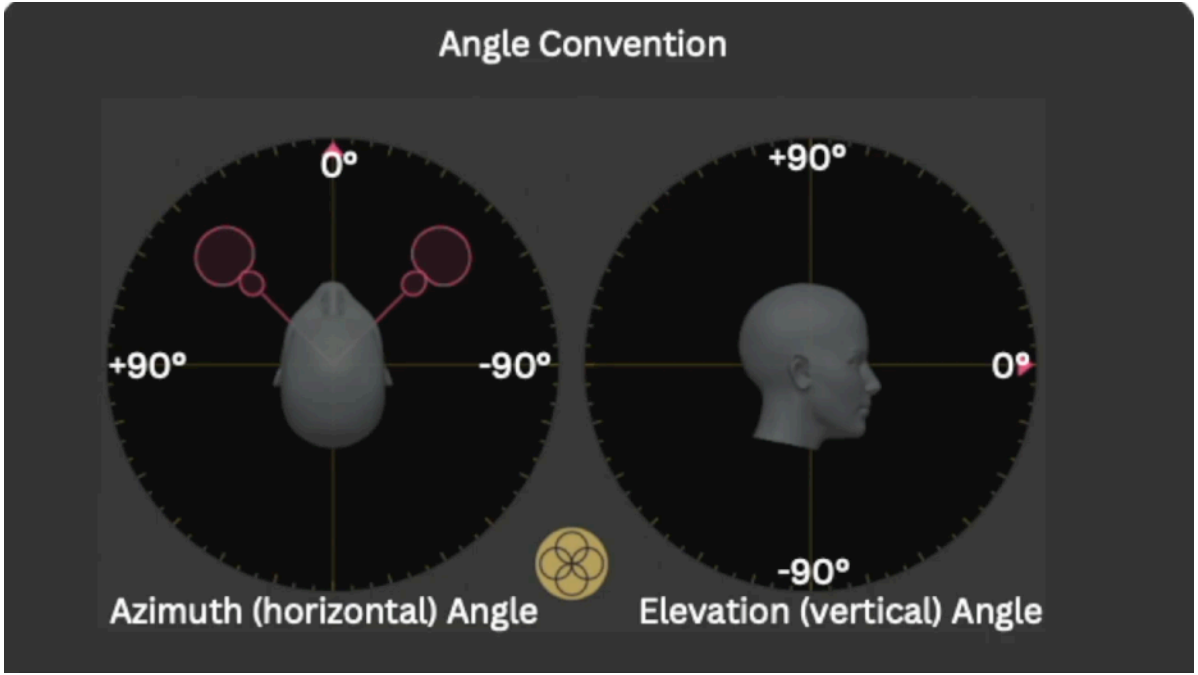


[1] Peebles et al., "Scalable Diffusion Models with Transformers "

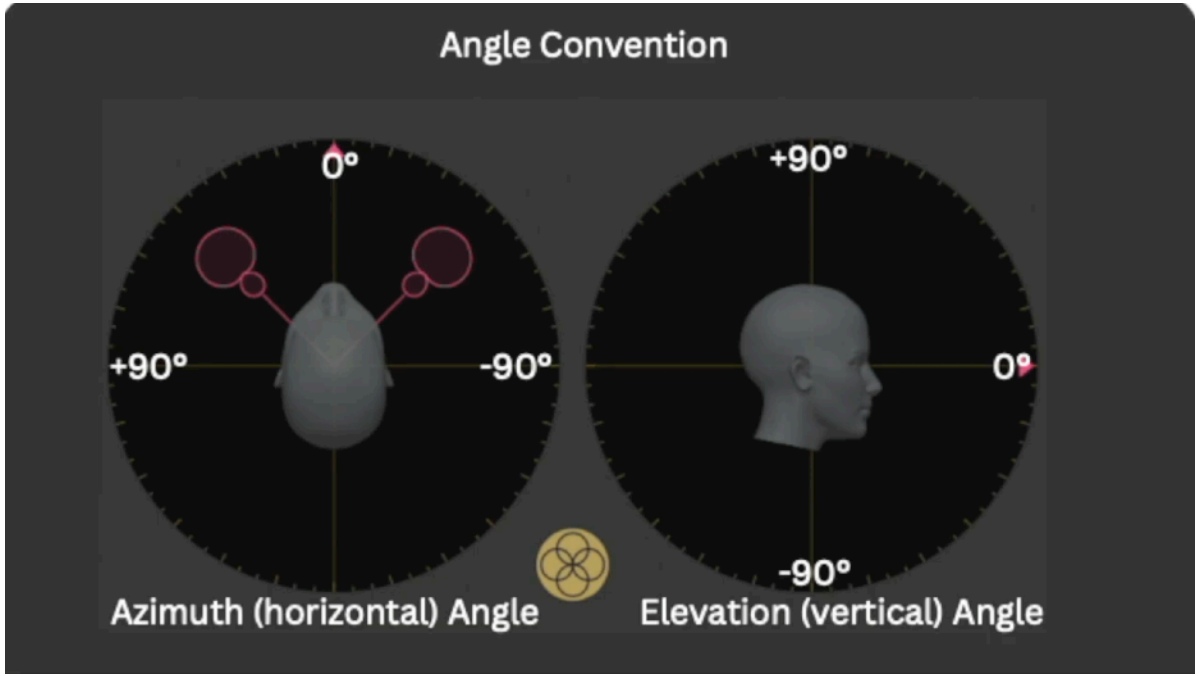
Approach: Inference



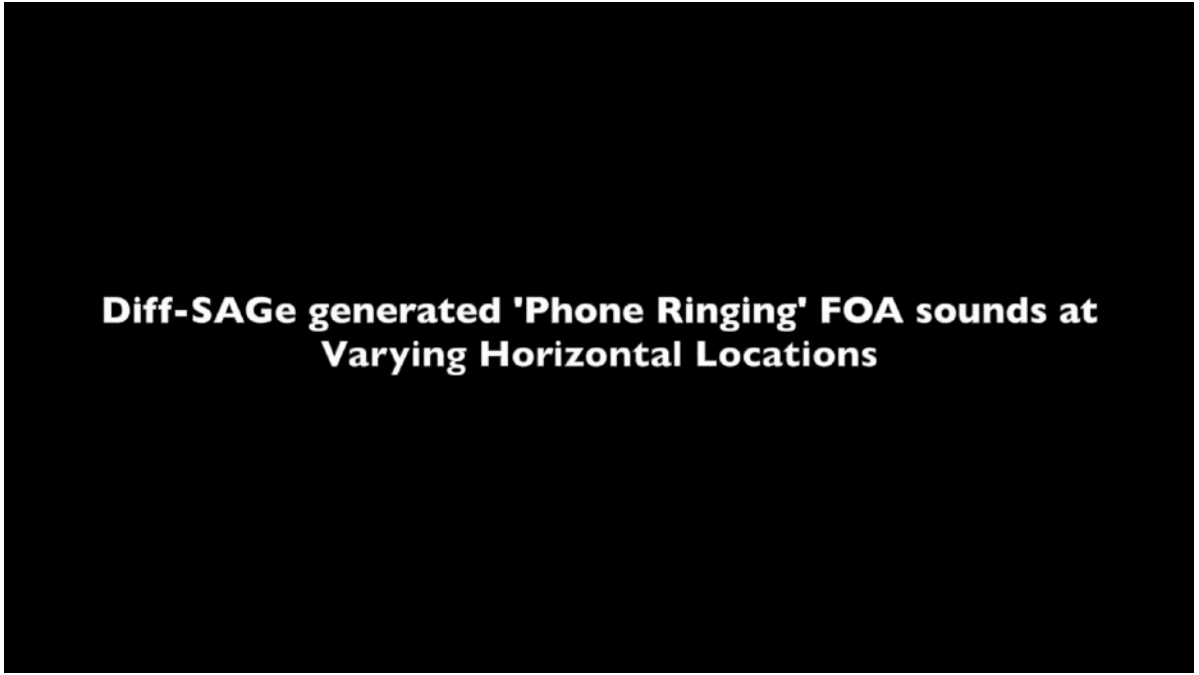
Demo examples : Along Azimuth



Demo examples : Along Elevation



Demo examples : randomly selected



Results: Comparison with Baseline

Comparison between Diff-SAGe and baseline approaches. Evaluation is conducted on the test set of TAU-19 and TAU-NIGENS-20.

		TAU-19					TAU-NIGENS-20				
		Condition		Distribution alignment			Condition		Distribution alignment		
		Acc(%)$_{\uparrow}$	DoA Error$_{\downarrow}$	FD$_{\downarrow}$	FAD$_{\downarrow}$	KL$_{\downarrow}$	Acc(%)$_{\uparrow}$	DoA Error$_{\downarrow}$	FD$_{\downarrow}$	FAD$_{\downarrow}$	KL$_{\downarrow}$
Ground-Truth	Human	87.19	32.06°	-	-	-	68.43	37.37	-	-	-
Simulated (Baseline)	Reference audio	62.66	3.33°	10.79	2.47	1.70	85.14	4.12°	15.94	3.05	1.90
	AudioLDM	14.57	**3.22°**	32.62	4.67	2.80	37.29	**3.07°**	22.64	3.11	1.98
	Tango	35.58	3.92°	23.03	8.04	2.05	52.00	3.39°	11.80	5.37	1.85
(Ours)	Diff-SAGe	**76.52**	22.97°	**3.93**	**0.64**	**1.44**	**85.29**	31.96°	**6.46**	**0.98**	**1.66**

Thank You

- Project Page: https://sakshamsingh1.github.io/diff-sage/
- Paper: https://arxiv.org/pdf/2410.11299