



NeuroAI



SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding



Mingfei Chen^{1,2}, Israel D. Gebru², Ishwarya Ananthabhotla³, Christian Richardt²,
Dejan Markovic², Jake Sandakly², Steven Krenn², Todd Keebler², Eli Shlizerman¹, Alexander Richard²
¹University of Washington ²Codec Avatars Lab, Pittsburgh, Meta ³Reality Labs Research, Meta

Arbitrary Scene



N Ref Microphones
Walking Path

Please turn audio ON with stereo speakers or headphones!





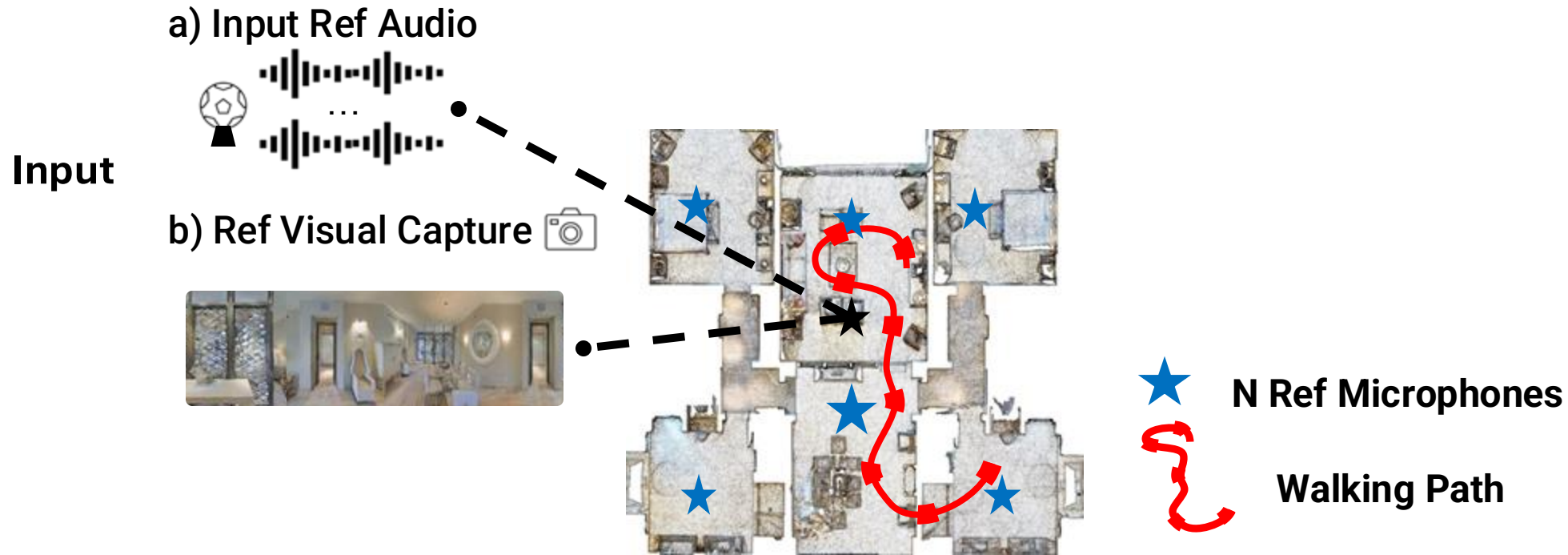
NeuroAI



SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding



Mingfei Chen^{1,2}, Israel D. Gebru², Ishwarya Ananthabhotla³, Christian Richardt²,
Dejan Markovic², Jake Sandakly², Steven Krenn², Todd Keebler², Eli Shlizerman¹, Alexander Richard²
¹University of Washington ²Codec Avatars Lab, Pittsburgh, Meta ³Reality Labs Research, Meta



Please turn audio ON with stereo speakers or headphones!



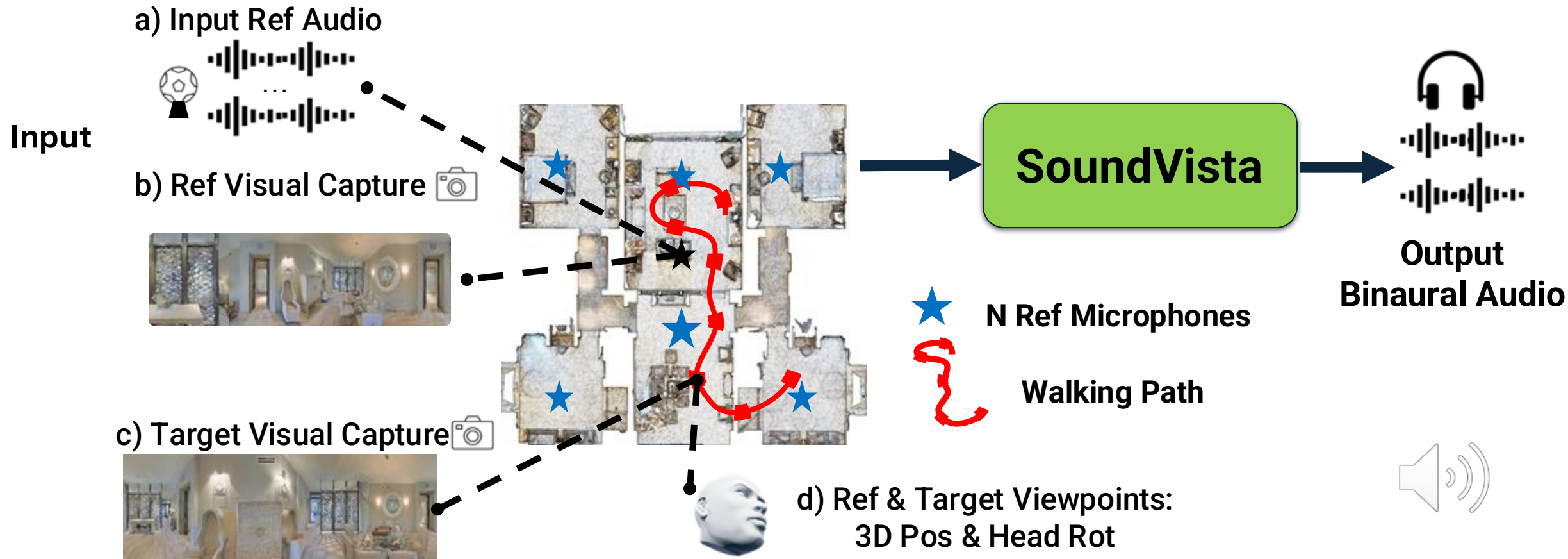
NeuroAI



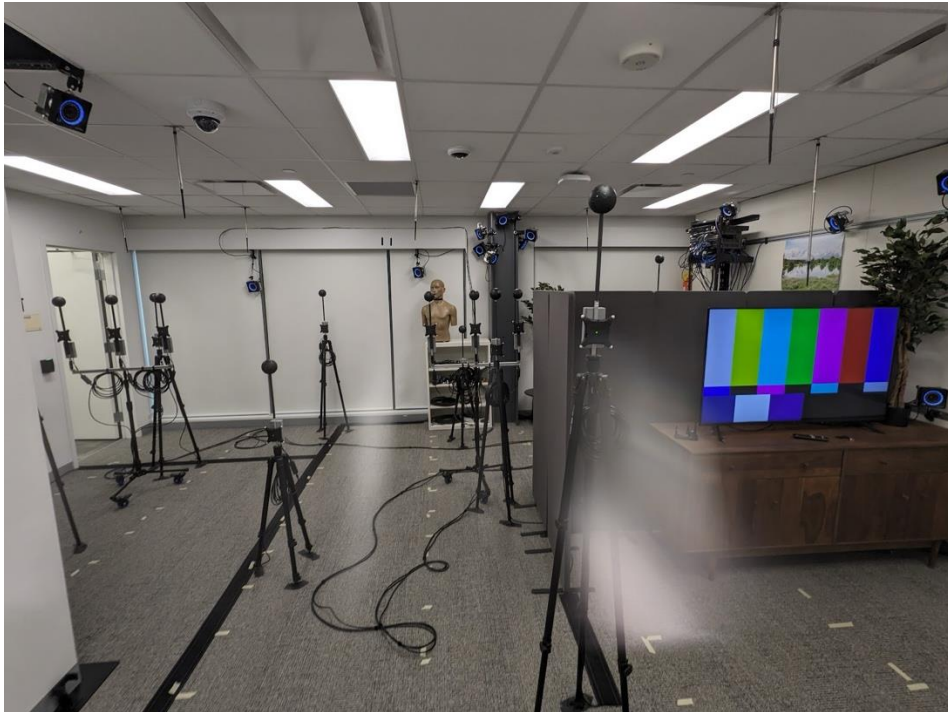
SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding



Mingfei Chen^{1,2}, Israel D. Gebru², Ishwarya Ananthabhotla³, Christian Richardt²,
Dejan Markovic², Jake Sandakly², Steven Krenn², Todd Keebler², Eli Shlizerman¹, Alexander Richard²
¹University of Washington ²Codec Avatars Lab, Pittsburgh, Meta ³Reality Labs Research, Meta



Real Demo: Reconstructing binaural ambient sounds of Walkable Spaces





Standard acoustic renderers with room impulse responses (RIRs) is not realistic solution

- **Ambient Sound - the acoustic state describing all sounds in a scene**
 - Dynamic nature with complex spatial and temporal changes
 - Hard to localize all sound sources in 3D
 - Lack of clean audio signals of each ambient source



Standard acoustic renderers with room impulse responses (RIRs) is not realistic solution

- **Ambient Sound - the acoustic state describing all sounds in a scene**
 - Dynamic nature with complex spatial and temporal changes
 - Hard to localize all sound sources in 3D
 - Lack of clean audio signals of each ambient source
- **High cost for fully reconstruct acoustic scene in real applications**
 - Impractical to obtain all impulse responses of the scenes



Novel-view Ambient Sound Synthesis with Sparse Reference Audio Recordings

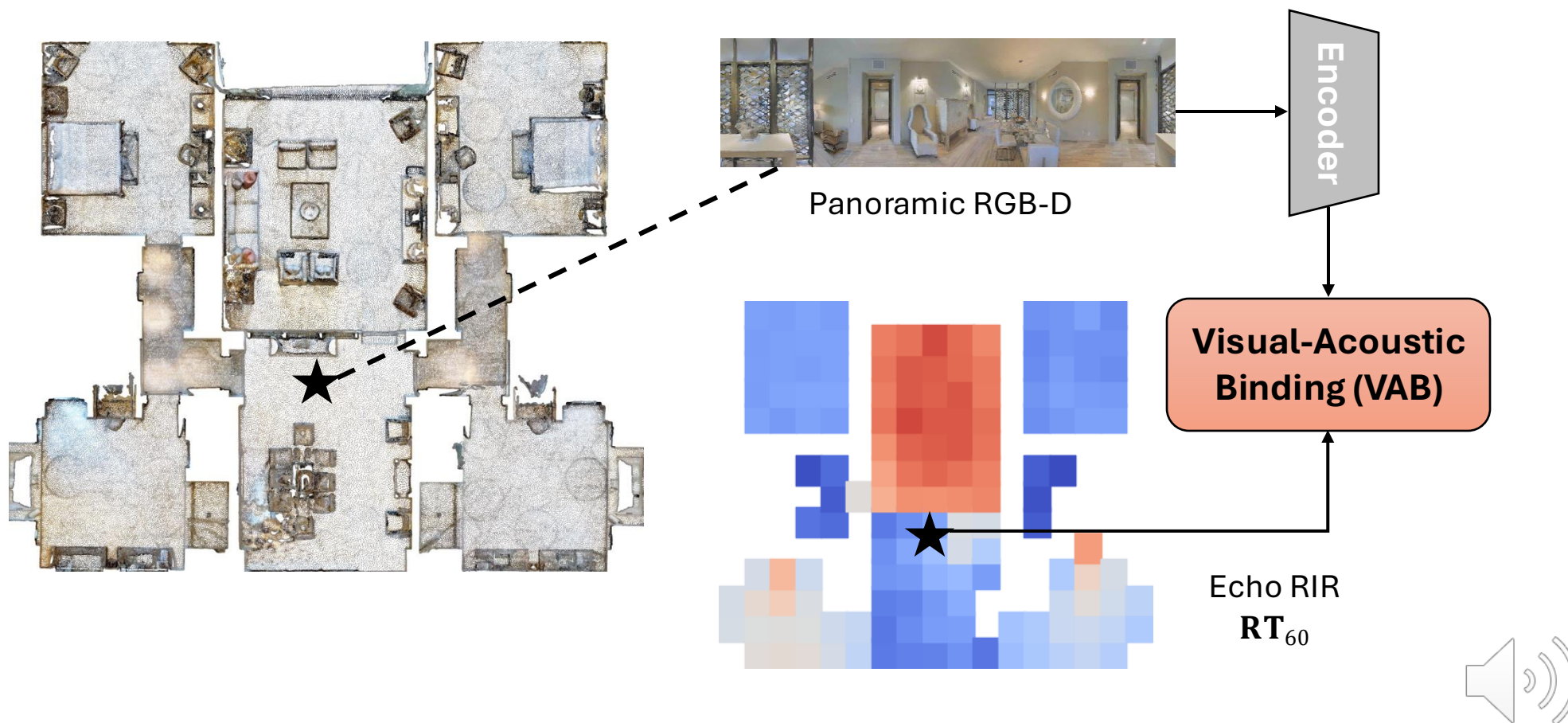


★ Reference

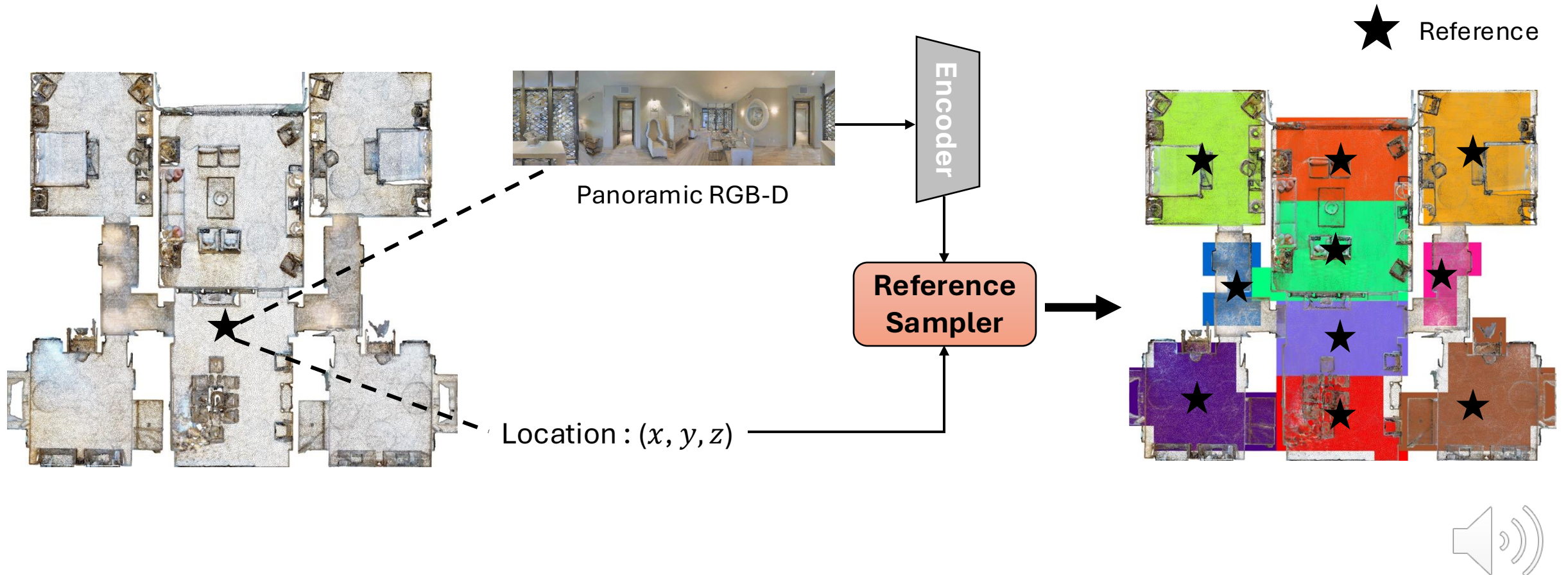
Where to place the N references?



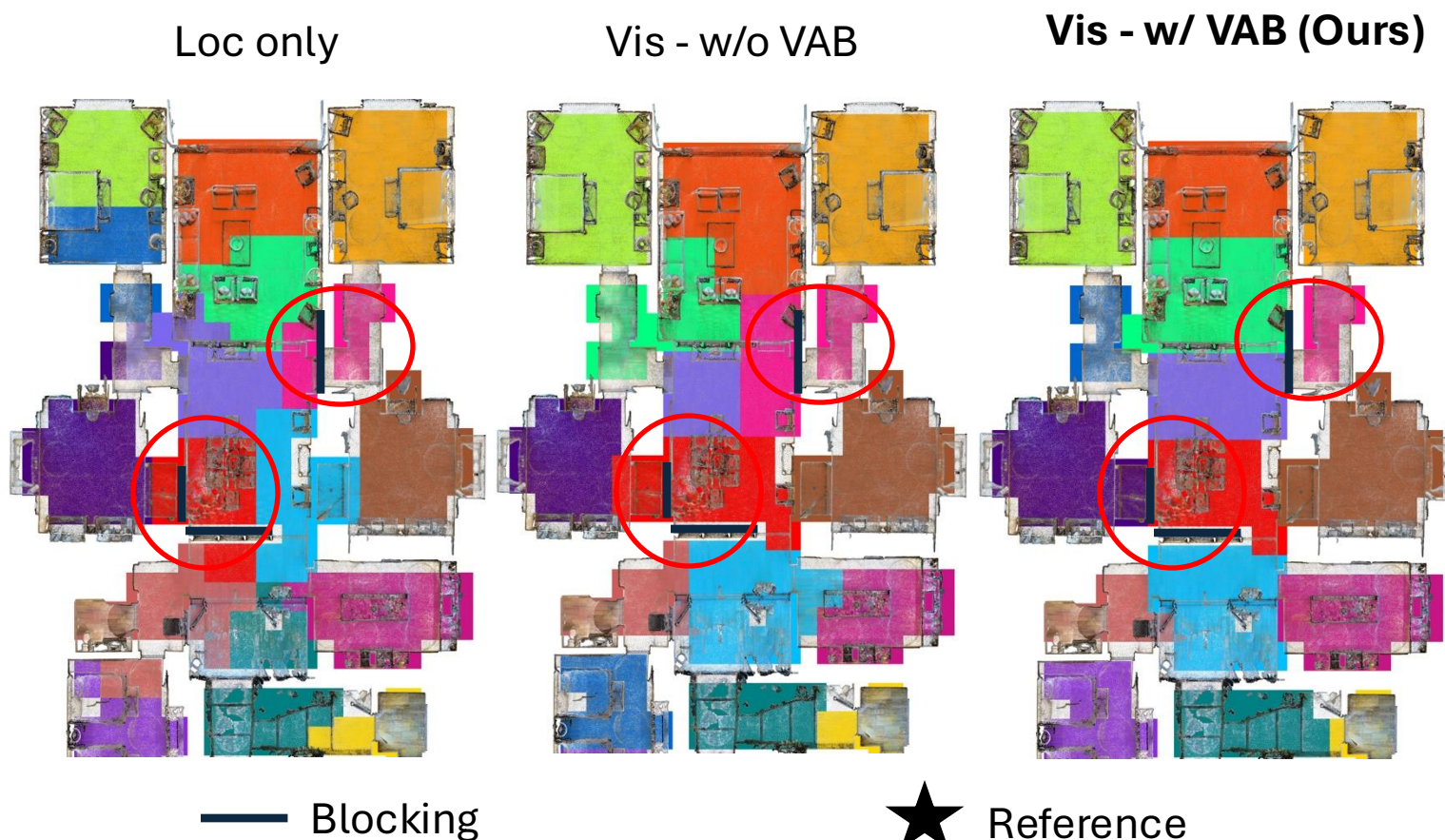
Reference Sampler Adaptive to Arbitrary Scenes via Visual-Acoustic Binding (VAB)



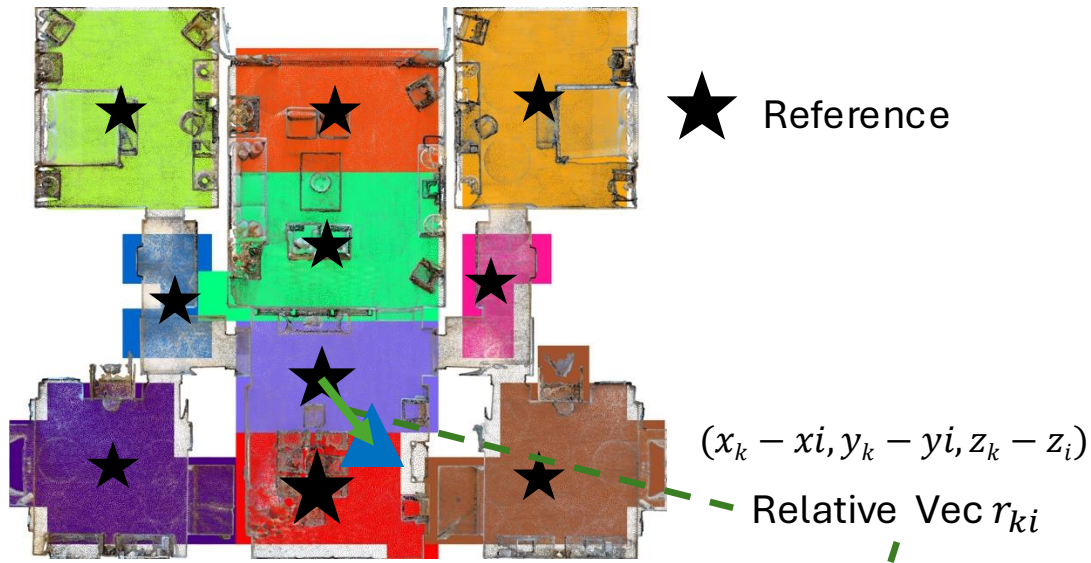
Reference Sampler Adaptive to Arbitrary Scenes via Visual-Acoustic Binding (VAB)



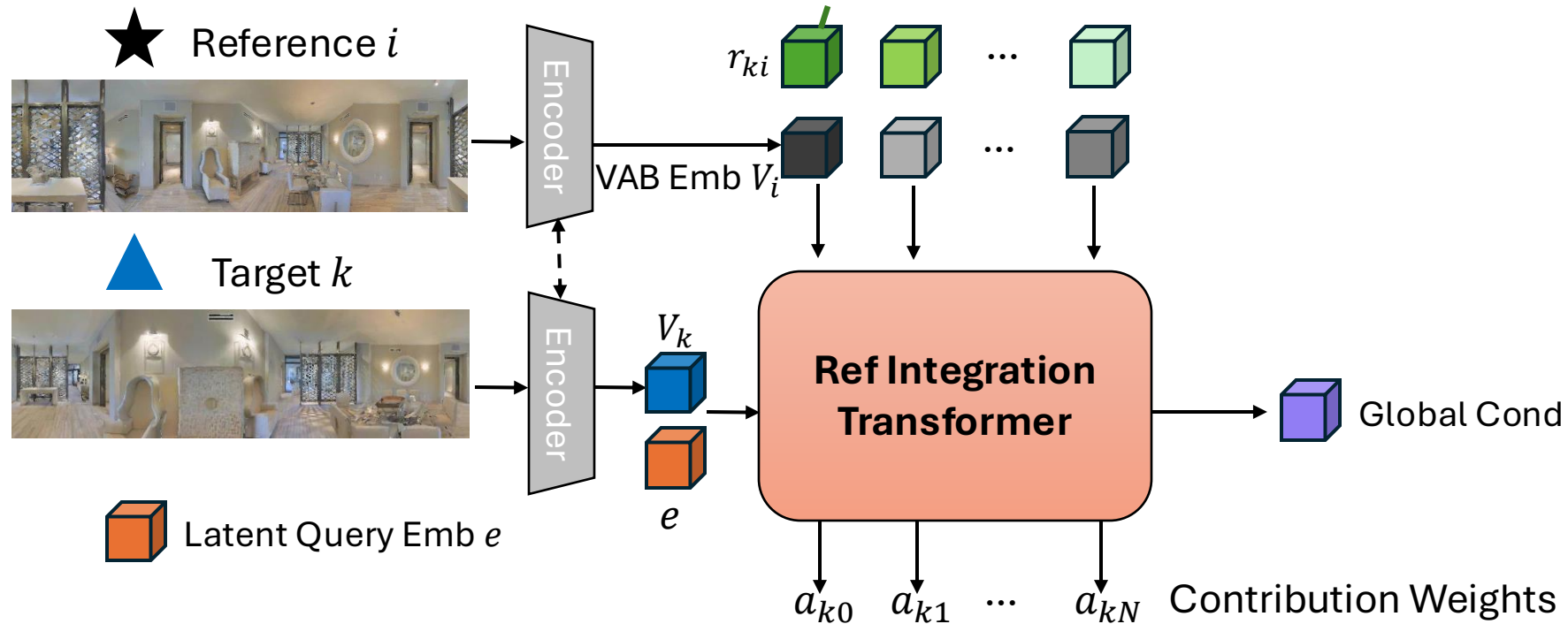
VAB recognizes acoustic partitions automatically aligned with scene geometry

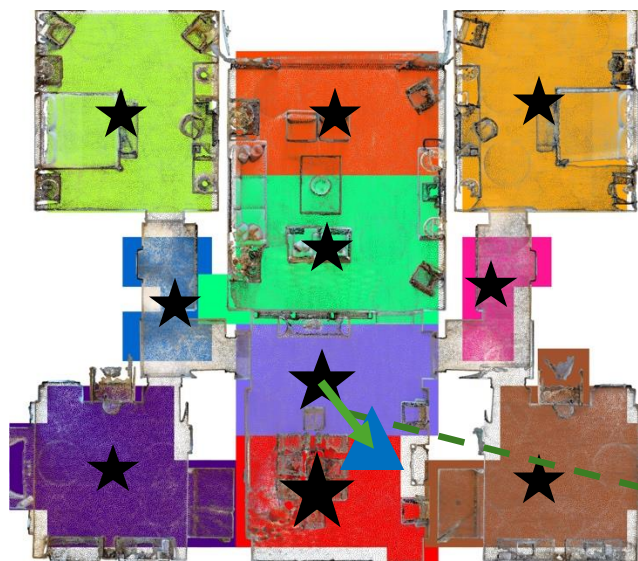


SoundVista Pipeline Overview



(a) Reference Integration Transformer





★ Reference Recording

STFT

(b) Reweighting

Relative Rot
 q_{ki}

$(x_k - x_i, y_k - y_i, z_k - z_i)$

Relative Vec r_{ki}

Local Cond

Reweight

Reweighted Channels

(a) Reference Integration Transformer

★ Reference i



▲ Target k



Latent Query Emb e

Encoder

VAB Emb V_i

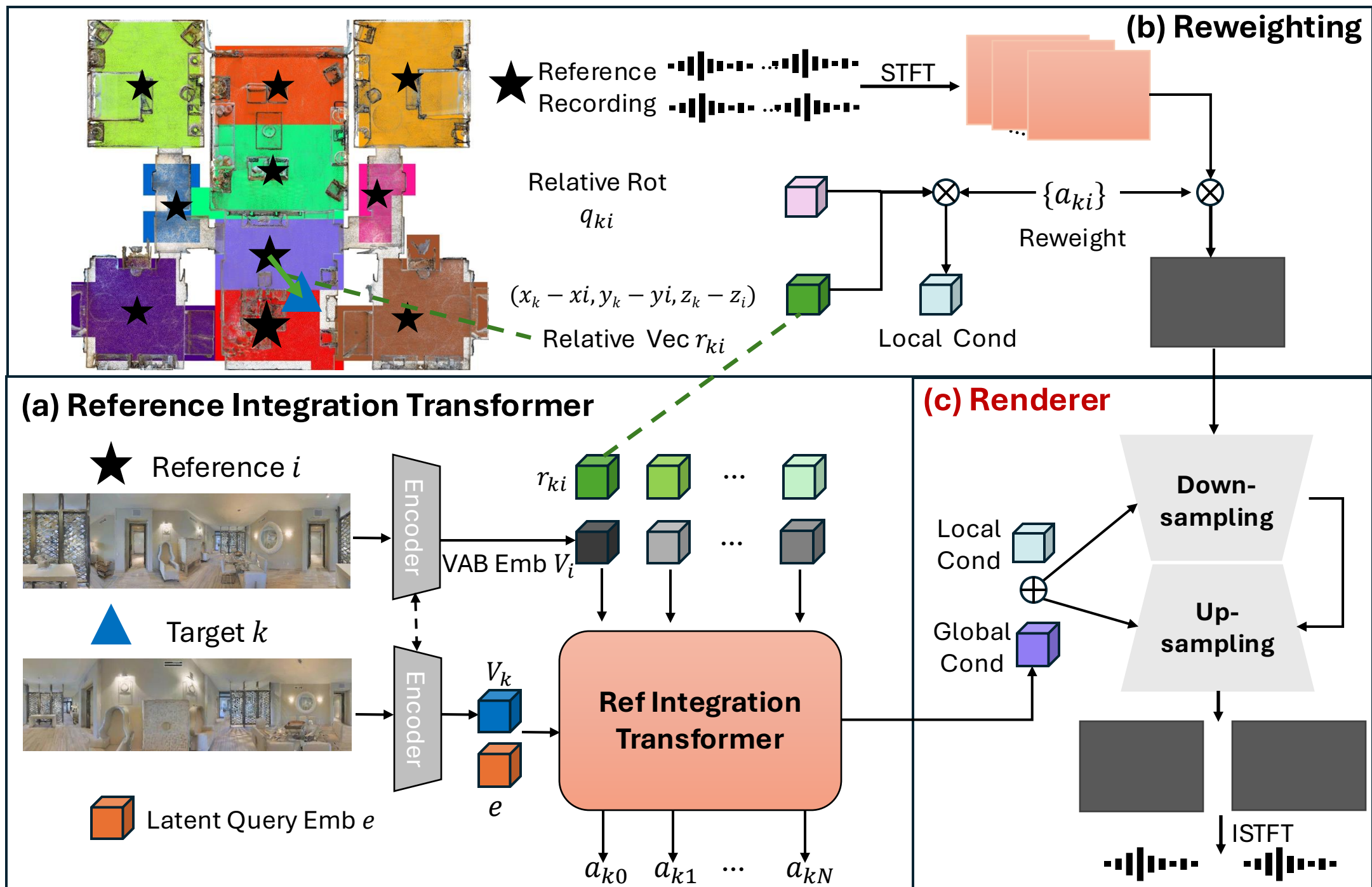
r_{ki}

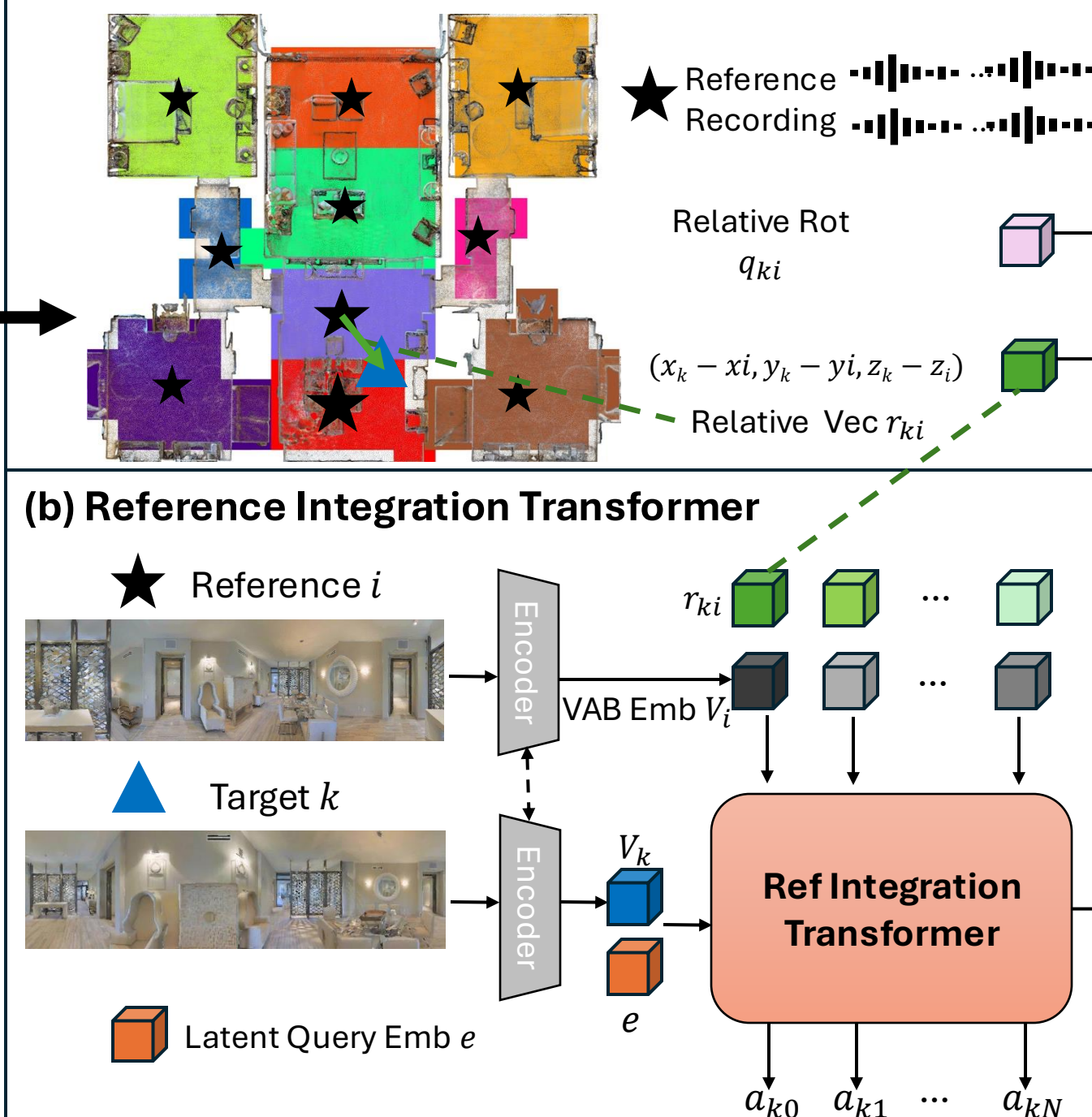
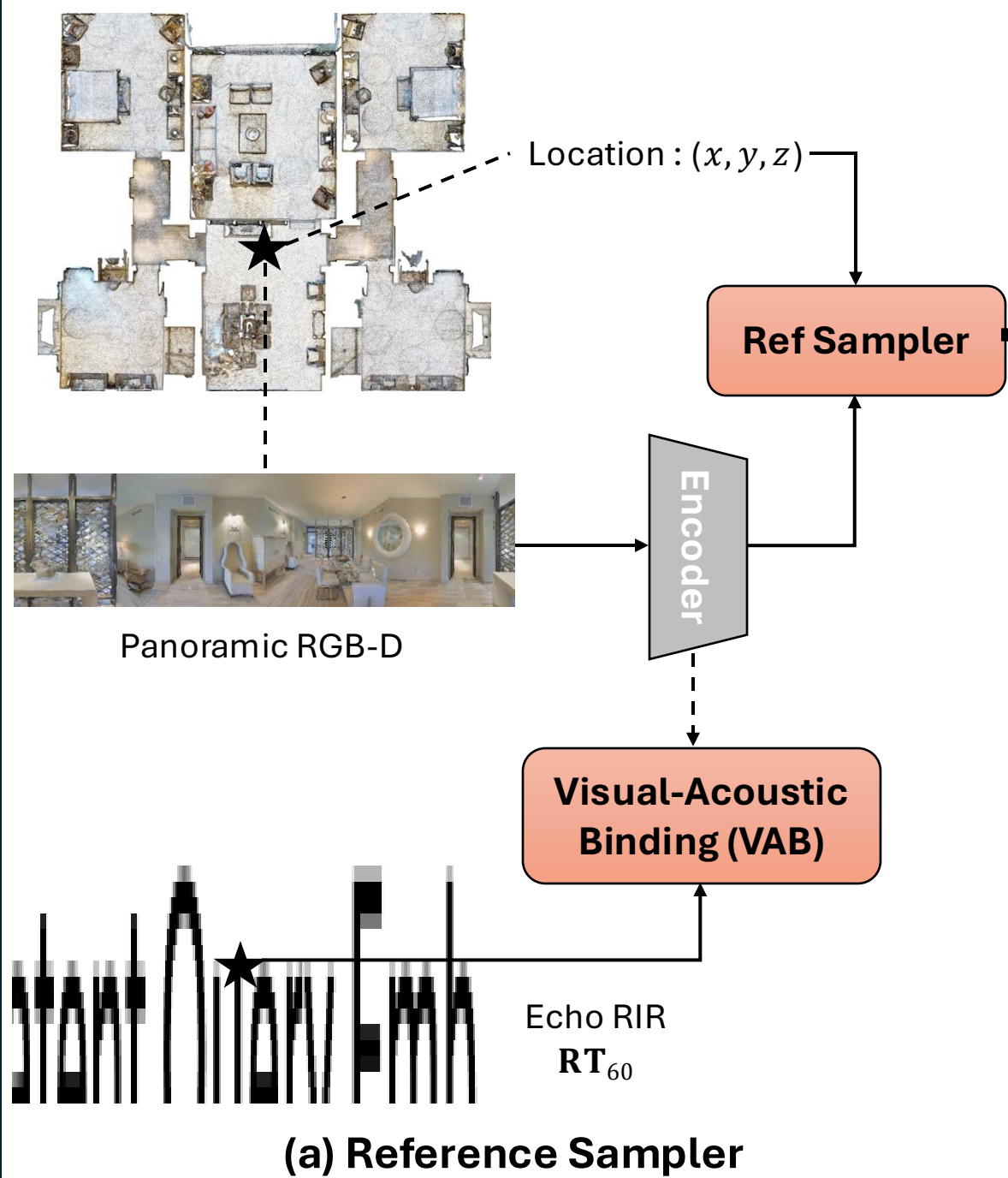
Ref Integration Transformer

Global Cond

a_{k0} a_{k1} ... a_{kN}

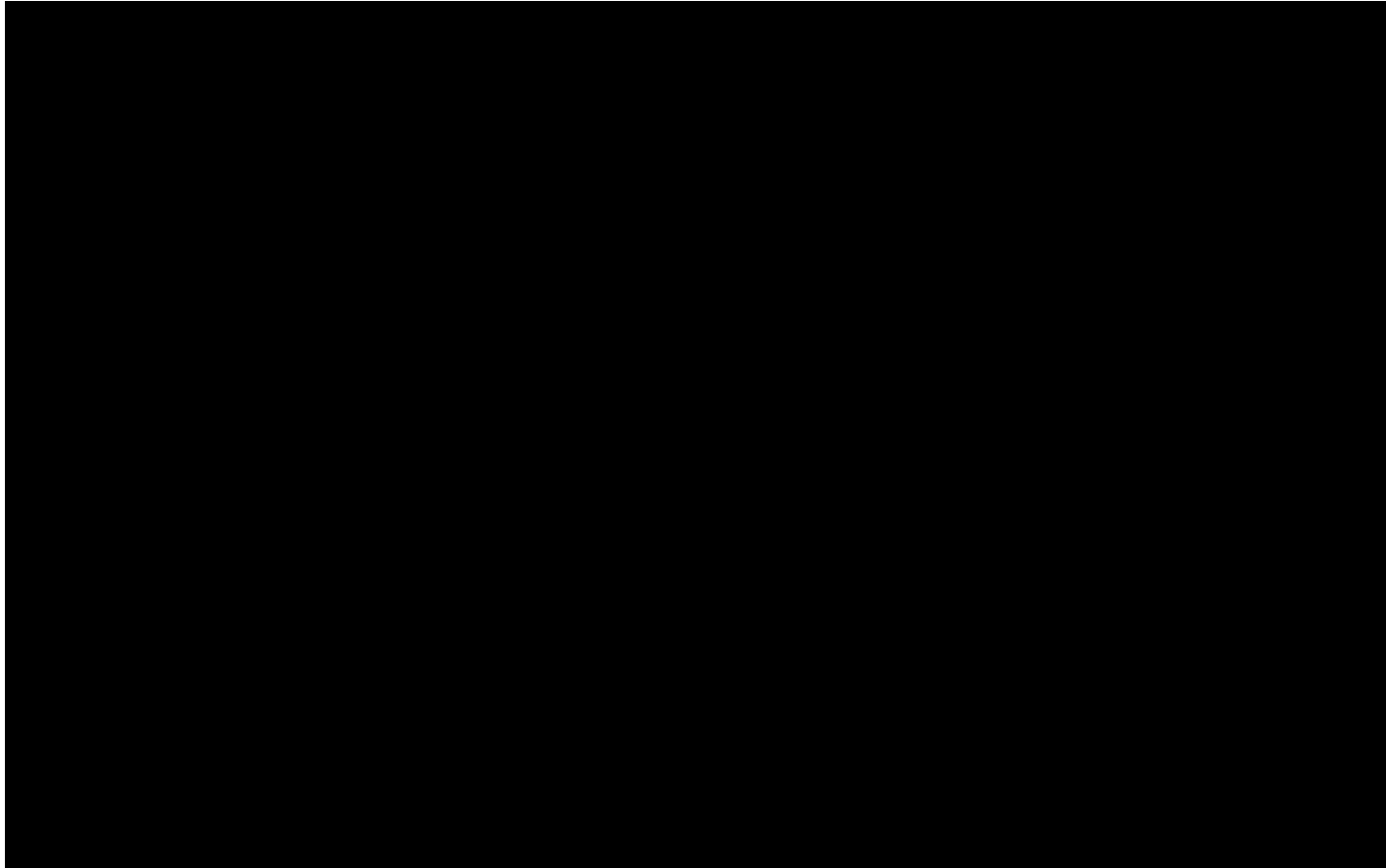






Qualitative Comparison with Baselines

- SOTA Baseline for Comparison: ViGAS [1]
- Demo Setting: Soundspace-Matterport3D [2] scene with 5 sources



[1] Novel-View Acoustic Synthesis. Chen et al. CVPR2023.

[2] SoundSpaces: Audio-Visual Navigation in 3D Environments. Chen et al. ECCV2020.

Qualitative Comparison with Baselines

