SONOWORLD

From One Image to a 3D Audio-Visual Scene

Derong Jin*, Xiyi Chen*, Ming C. Lin, Ruohan Gao
University of Maryland, College Park

Project Page

Scan to try intractive demos & VR demos!

CVPR JUNE 3-7, 2026 DENVER COLORADO

TL;DR: Given a single image, SonoWorld generates an explorable 3D scene with spatial audio aligned to the scene's geometry and semantics.

Motivation & Task

Recent single-image 3D generation can create explorable visual worlds, but they remain *silent*. Real immersion requires spatial sound: what can be heard, where it comes from, and how it changes as you move.

New Task: Image2AVScene

Input: one RGB image

Output: an explorable *3D visual scene* + a *spatial audio field* aligned with scene semantics and 3D geometry.

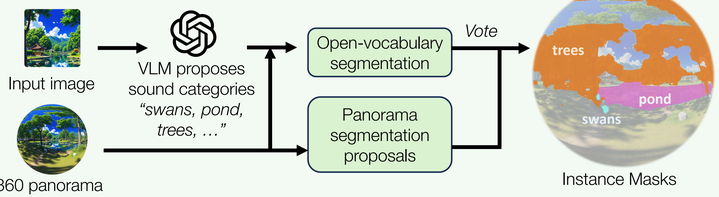
How SonoWorld Works

1 Visual Scene Generation

From a single image to a full 360 visual scene

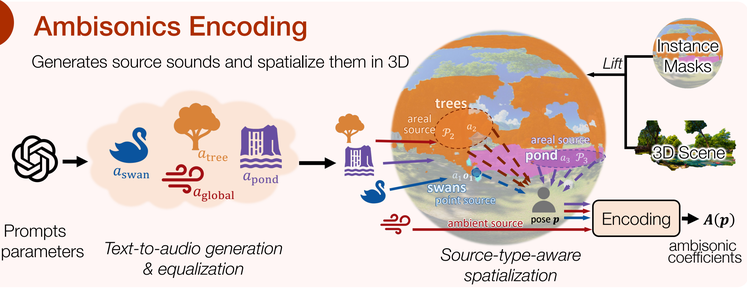
2 360 Semantic Grounding

Understand the scene and localize the sounding objects & regions



3 Ambisonics Encoding

Generates source sounds and spatialize them in 3D



4 Free-Viewpoint Rendering

Render synchronized audio-visual outputs at arbitrary pose

Evaluation & Results

SonoScene360 Dataset

- Calibrated camera & ambisonics microphone
- 68 synchronized 360 video + first-order ambisonics (FOA) clips across 6 real scenes
- Annotations of sound labels, descriptions, and coarse directions

Azimuth: 5.5 °
Elevation: 2.2 °
Rotation Quaternion: [0.61, -0.35, 0.32, 0.64]
Mic Calibration

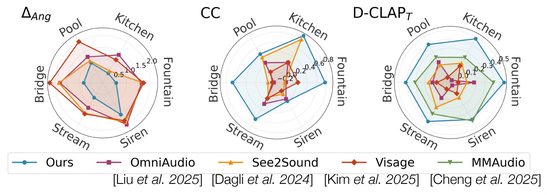
Source label: *"fountain"*
Direction: *"left"*
Description: *"Bright, sparkling water noise from fountain."*
Text Annotation

Capture Setup & Annotation Example

Metrics: **Spatial** Angular DoA Err (°) ↓ · CC ↑ · AUC ↑ | **Semantic** D-CLAP-A ↑ · D-CLAP-T ↑ · D-CLAP-R ↑

Quantitative Results

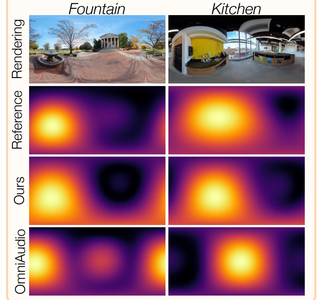
Numerical evaluation on SonoScene360 dataset:



- Consistently better semantic alignment & spatial grounding
- **Preferred by users in perceptual study**

Qualitative Results (Real)

Spatial Audio Energy Maps:



Qualitative Results (Synthetic)