

1 USER CONDITION / INPUTS

User condition

$u = (u^{text}, u^{img})$

Text

"a calming evening"

Image

Conditioning prompts (optional)

duration: 25 s
channels: stereo
format: .wav
...

2 AGENT-GUIDED SCENE PLANNING

LLM-based Soundscape Agent

Multimodal intent parsing → Concise-audible grounding → Scene planning

Grounded example:
"a calming evening" → explicit events

- soft rain **bg** 0–25 s -30 dB
- distant traffic **mid** 10–25 s -32 dB
- page-turn **fg** ~ 12 s -22 dB

bg **mid** **fg** background / midground / foreground

OFFLINE PRIOR MODE

Scene priors π (reusable)

Sample scene plan

- e_1 _____
- e_2 _____
- $\vdots$
- e_N _____

3 ASSET LIBRARY & HYBRID ACQUISITION

Asset library and hybrid acquisition

$L = L^{open} \cup L^{priv} \cup L^{gen}$

L^{open} Open-source recordings
 L^{priv} in-house assets
 L^{gen} Generated assets

CLAP similarity + Production complexity (For score)

Fuzzy retrieval (For ambience)

Strict retrieval (For salient constrained events)

On-demand generation (Optional user guidance)

add back to library

Selected assets A

4 STRUCTURED RENDERING

Deterministic renderer

Multi-track timeline (0–25 s)

0 10 20 25 s

bg **mid** **fg**

Rendering operations (per event)

- ✂ crop
- ⏸ time silence
- 🔄 loop/extend
- 🔊 loudness remap
- 📶 fades
- ⚙ augmentation
- 📍 spatialization
- ✅ safety checks
- ⬆ light onset/gain perturbation
- 🔄 bounded retry for invalid clips

Mixed waveform y

5 ALIGNED OUTPUTS

Soundscape waveform y

Scene specification E

Metadata m

Aligned description d

"Soft rain continues throughout the scene as distant traffic emerges around 10 seconds and remains audible until the end, with a single page turn occurring around 12 seconds.."

Additional exports

- Stems (per event)
- Role-specific statistics (bg/mid/fg)

User-guided Editing (optional)

EVALUATION Trak A

★★★★★

Generation quality: listener studies + objective metrics

EVALUATION Trak B

Scalable audio-language training: MMAU, sound track