STASE: A Spatialized Text-to-Audio Synthesis Engine for Music Generation

An LLM-driven framework for generating musical compositions with user-specified spatial attributes

Tutti Chi[1] Letian Gao[2] Yixiao Zhang[3]

[1] University of Chinese Academy of Sciences, China

[2] Tsinghua University, China

[3] Centre for Digital Music (C4DM), Queen Mary University of London, UK

LLM4Music @ ISMIR 2025

Abstract

While many text-to-audio systems produce monophonic or fixed-stereo outputs, generating audio with user-defined spatial properties remains a challenge. We introduce STASE, a system that leverages a Large Language Model (LLM) as an agent to interpret spatial cues from text. A key feature of STASE is the decoupling of semantic interpretation from a separate, deterministic signal-processing-based spatial rendering engine, which facilitates interpretable and user-controllable spatial reasoning. The LLM processes prompts through two pathways: (i) Description Prompts for direct spatial information mapping, and (ii) Abstract Prompts with Retrieval-Augmented Generation (RAG) for spatial template retrieval.

1 Introduction

AI-powered music generation advances rapidly while immersive audio gains industry traction. However, current spatial audio synthesis workflows provide limited controllability. Existing deep learning spatialization methods operate solely on audio inputs without leveraging text descriptions. Their black-box nature limits precise control over critical psychoacoustic parameters.

We propose STASE, a hybrid neuro-symbolic approach. Our key innovation: **decoupling semantic interpretation from spatial rendering**. An LLM interprets natural language while deterministic signal processing ensures controllable spatialization, enabling precision for experts and intuitive interaction for novices.

2 Methodology

STASE Architecture

STASE is an LLM-driven spatial audio synthesis framework designed to generate musical compositions with user-specified spatial attributes from natural language prompts. The system operates in four sequential stages:

a. **Prompt Processing:** Natural language input with RAG-based template retrieval
b. **Conductor Agent:** LLM-based reasoning module that outputs structured plans
c. **Music Generation:** Synthesis of individual audio tracks (stems)
d. **Spatial Rendering:** Deterministic signal-processing-based spatialization

Two Processing Pathways

- **Description Prompts:** Direct spatial information mapping when precise positional details are given
- **Abstract Prompts:** RAG module retrieves relevant spatial templates for coherent layouts

Key Innovation

Decoupling of semantic interpretation from deterministic signal-processing-based spatial rendering supports interpretable and user-controllable spatial reasoning.

3 Signal Processing Components

Spatial Localization:

- Stereo amplitude panning
- ITD/ILD rendering
- HRTF convolution (KEMAR)

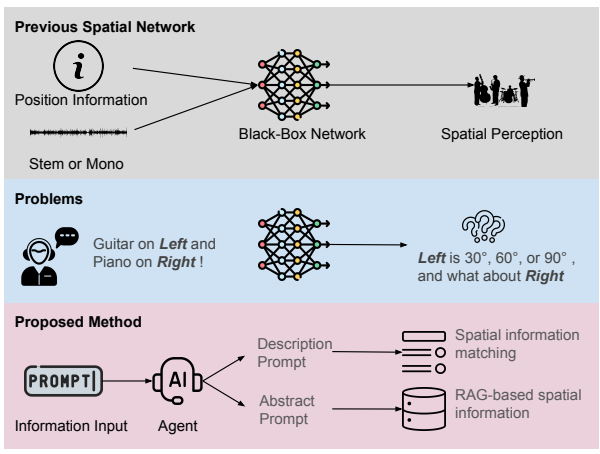
Environmental Acoustics:

- Parameterized reverb
- RIR convolution

4 Conclusion

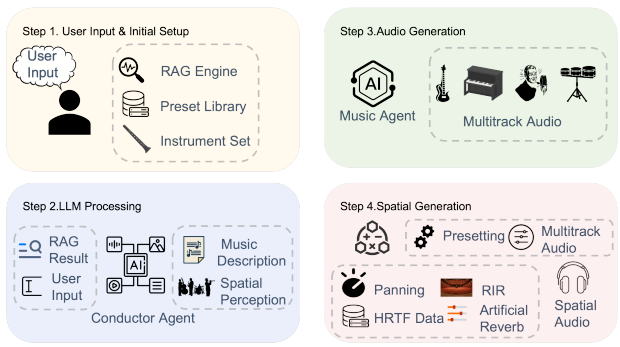
STASE presents a novel approach to text-driven spatial audio synthesis by decoupling semantic interpretation from deterministic signal processing. The system enables interpretable and controllable spatial reasoning, offering both precision for experts and accessibility for novices.

System Overview



Comparison: Latent-space vs STASE

STASE Workflow



Complete STASE pipeline

Key Features

- LLM-driven spatial interpretation
- Deterministic signal processing
- Template-based spatial layouts
- Supports panning, ITD/ILD, HRTF
- Multiple spatial configurations
- Modular, replaceable components

Results & Demo

Audio demonstrations:
`https://chengtopia.github.io/STASE.github.io/`