Event-Aware Audio Generation for LLM-Driven Storytelling in Extended Reality

Mehmet Karaaslan, Meral Kuyucu, Bora Şenceylan, Gökhan İnce

24.05.2026

Mehmet Karaaslan (karaaslan18@itu.edu.tr)

Biography

Mehmet Karaaslan

B.Sc. & M.Sc. Computer Engineering in Istanbul Technical University

Mainly working on AI, and XR systems.

Introduction

- Sound is a fundamental modality for perceiving surroundings and spatial orientation.

Introduction

- Sound is a fundamental modality for perceiving surroundings and spatial orientation.
- Auditory feedback is not supplementary; it's vital for a mental model of presence.

Introduction

- Sound is a fundamental modality for perceiving surroundings and spatial orientation.
- Auditory feedback is not supplementary; it's vital for a mental model of presence.
- While visual fidelity in XR has reached high levels, auditory experiences lag behind.

Introduction

- Sound is a fundamental modality for perceiving surroundings and spatial orientation.
- Auditory feedback is not supplementary; it's vital for a mental model of presence.
- While visual fidelity in XR has reached high levels, auditory experiences lag behind.
- Current methods are often manual, rule-based, and disconnected from the narrative context.

The Sonification Gap

Without synchronized, contextually accurate sound, even the most visually stunning immersive experience feels hollow and artificial.

The Sonification Gap

Without synchronized, contextually accurate sound, even the most visually stunning immersive experience feels hollow and artificial.

Manual Effort

Retrieval-based mappings require substantial human authoring for every interaction.

The Sonification Gap

Without synchronized, contextually accurate sound, even the most visually stunning immersive experience feels hollow and artificial.

Manual Effort

Retrieval-based mappings require substantial human authoring for every interaction.

Temporal Misalignment

Generative models lack inherent synchronization with scene animations and events.

The Sonification Gap

Without synchronized, contextually accurate sound, even the most visually stunning immersive experience feels hollow and artificial.

Manual Effort

Retrieval-based mappings require substantial human authoring for every interaction.

Temporal Misalignment

Generative models lack inherent synchronization with scene animations and events.

Narrative Void

Most systems focus on isolated effects rather than continuity across multiple scenes.

Literature Review

- Focuses on synthesizing audio triggered by explicit physical interactions (e.g., sliding objects). **SonifyAR [Su, '24]**

Literature Review

- Focuses on synthesizing audio triggered by explicit physical interactions (e.g., sliding objects). **SonifyAR [Su, '24]**

- Introduces a standalone model optimized for precise timestamp controllability during sound generation. **PicoAudio [Xie, '24]**

Literature Review

- Focuses on synthesizing audio triggered by explicit physical interactions (e.g., sliding objects). **SonifyAR [Su, '24]**
- Introduces a standalone model optimized for precise timestamp controllability during sound generation. **PicoAudio [Xie, '24]**
- Tailored for synchronizing foley audio directly extracted from video motion data. **DreamFoley [Li, '25]**

Literature Review

- Focuses on synthesizing audio triggered by explicit physical interactions (e.g., sliding objects). **SonifyAR [Su, '24]**
- Introduces a standalone model optimized for precise timestamp controllability during sound generation. **PicoAudio [Xie, '24]**
- Tailored for synchronizing foley audio directly extracted from video motion data. **DreamFoley [Li, '25]**
- Generates spatial sound configurations derived from individual visual frames. **SEE-2-SOUND [Dagli, '25]**

Literature Review

- Focuses on synthesizing audio triggered by explicit physical interactions (e.g., sliding objects). **SonifyAR [Su, '24]**
- Introduces a standalone model optimized for precise timestamp controllability during sound generation. **PicoAudio [Xie, '24]**
- Tailored for synchronizing foley audio directly extracted from video motion data. **DreamFoley [Li, '25]**
- Generates spatial sound configurations derived from individual visual frames. **SEE-2-SOUND [Dagli, '25]**
- **Our Work:** Diverges by introducing Semantic Representation to maintain narrative continuity and persistent audio binding across multi-scene XR scenarios.

Research Objectives

RQ1: Narrative Alignment

Does using structured scene descriptions (Semantic Representation) improve alignment and immersion compared to video-based input?

Research Objectives

RQ1: Narrative Alignment

Does using structured scene descriptions (Semantic Representation) improve alignment and immersion compared to video-based input?

RQ2: Tool Viability

Do technically literate users perceive the proposed framework as a viable tool for automated audio authoring in XR?

Research Objectives

RQ1: Narrative Alignment

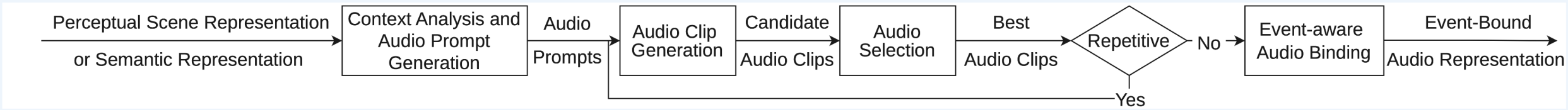
Does using structured scene descriptions (Semantic Representation) improve alignment and immersion compared to video-based input?

RQ2: Tool Viability

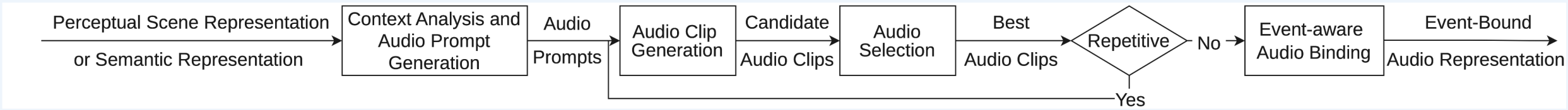
Do technically literate users perceive the proposed framework as a viable tool for automated audio authoring in XR?

Our study evaluates an LLM-driven pipeline designed to extract events and generate context-specific audio.

Proposed Framework

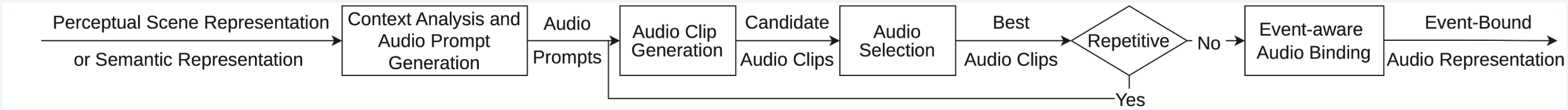


Proposed Framework



1. Context Analysis: LLM analyzes scene data to extract sound-inducing events, then generates corresponding prompts.

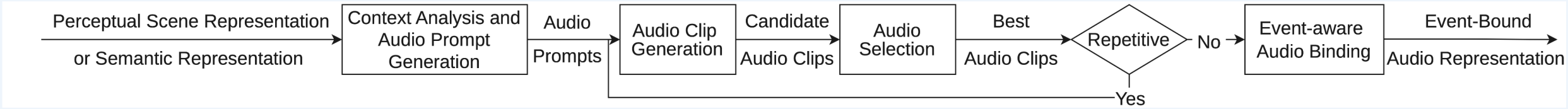
Proposed Framework



1. Context Analysis: LLM analyzes scene data to extract sound-inducing events, then generates corresponding prompts.

2. Audio Generation: Latent Diffusion Model (AudioLDM 2) synthesizes N candidates.

Proposed Framework

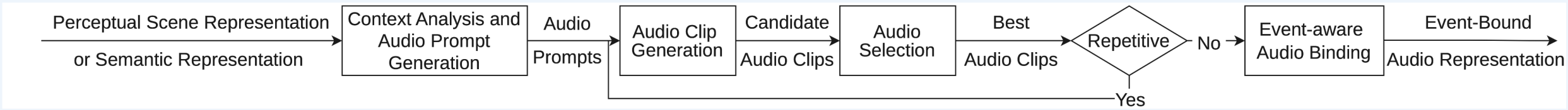


1. Context Analysis: LLM analyzes scene data to extract sound-inducing events, then generates corresponding prompts.

2. Audio Generation: Latent Diffusion Model (AudioLDM 2) synthesizes N candidates.

3. Selection: LLM-as-a-Judge selects the best candidate based on thematic and technical quality.

Proposed Framework



1. Context Analysis: LLM analyzes scene data to extract sound-inducing events, then generates corresponding prompts.

2. Audio Generation: Latent Diffusion Model (AudioLDM 2) synthesizes N candidates.

3. Selection: LLM-as-a-Judge selects the best candidate based on thematic and technical quality.

4. Binding: Integrates generated audio files into the XR environment

Input Representations

Perceptual Scene (PSR)

Scene Video + Short Summary

Approximates vision-based approaches. Often misses temporal nuance of narrative events.

Input Representations

Perceptual Scene (PSR)

Scene Video + Short Summary

Approximates vision-based approaches. Often misses temporal nuance of narrative events.

Semantic (SR)

Objects + Actions + Temporal Data

Encoded environmental data (Contextual). Allows the LLM to act as a "Digital Sound Designer".

LLM-as-a-Judge

To reduce randomness in diffusion models, we produce N candidate clips and rank them using LLM.

LLM-as-a-Judge

To reduce randomness in diffusion models, we produce N candidate clips and rank them using LLM.

Thematic Accuracy

Matches textual prompt.

LLM-as-a-Judge

To reduce randomness in diffusion models, we produce N candidate clips and rank them using LLM.

Thematic Accuracy

Matches textual prompt.

Technical Quality

Clarity & No Artifacts.

LLM-as-a-Judge

To reduce randomness in diffusion models, we produce N candidate clips and rank them using LLM.

Thematic Accuracy

Matches textual prompt.

Technical Quality

Clarity & No Artifacts.

Atmospheric Coherence

Immersion contribution.

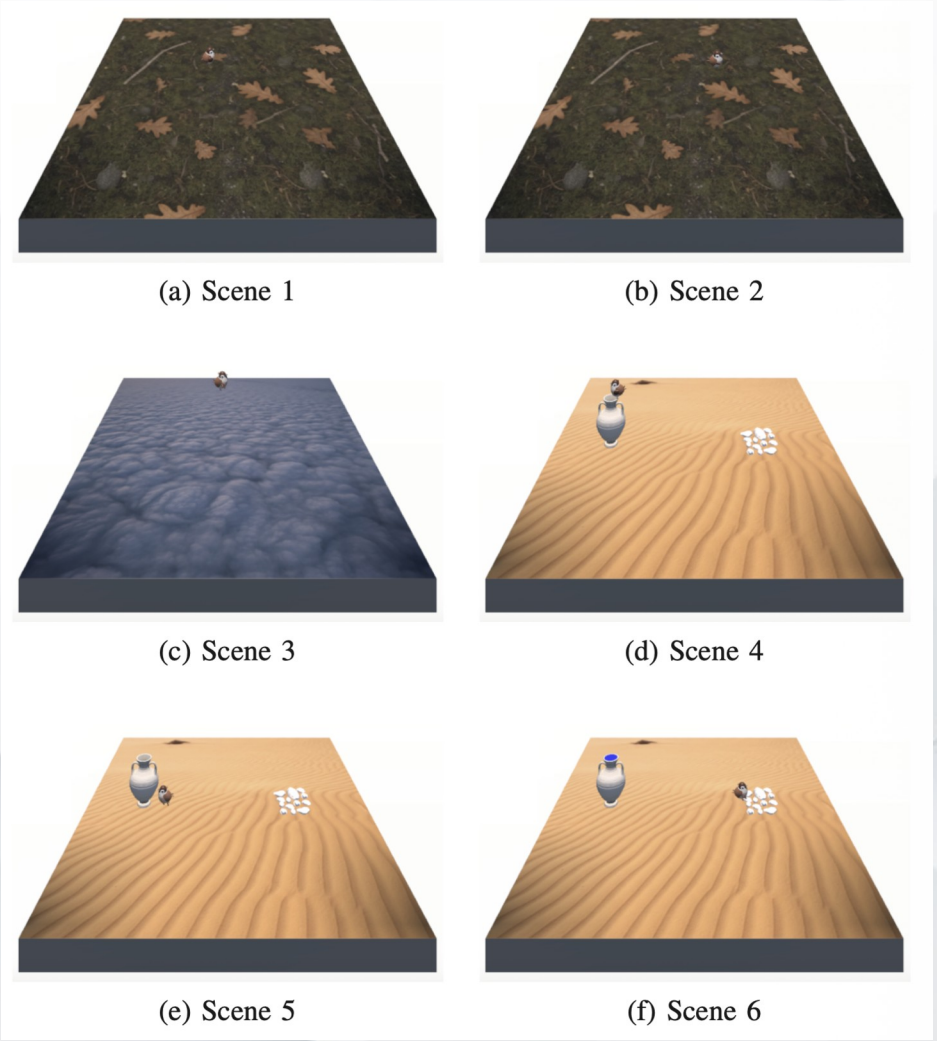
Evaluation (Unity Demo)

Narrative: 6 sequential scenes.

Participants: 20 people (purposive sampling, XR-familiar).

Method: Within-subjects comparison of PSR vs. SR input methods.

Device: Meta Quest 3



Subjective Preference Results

Measure (Likert 1-5)	Perceptual (PSR)	Semantic (SR)
Audio Narrative Alignment	2.80 ± 1.41	**3.75 ± 1.33**
Audio-Visual Synchronization	2.56 ± 1.28	**4.30 ± 0.78**
Immersion Contribution	2.94 ± 1.36	**3.80 ± 1.43**

Results on Tool

The secondary objective of our study was to evaluate the system from a developer’s perspective, specifically targeting its potential as an automated authoring tool.

Results on Tool

The secondary objective of our study was to evaluate the system from a developer's perspective, specifically targeting its potential as an automated authoring tool.

Adaptation Intent

Results on Tool

The secondary objective of our study was to evaluate the system from a developer's perspective, specifically targeting its potential as an automated authoring tool.

Adaptation Intent

Efficiency

Results on Tool

The secondary objective of our study was to evaluate the system from a developer's perspective, specifically targeting its potential as an automated authoring tool.

Adaptation Intent

Efficiency

Usefulness

Conclusion and Contribution

- **End-to-End Modular Pipeline:** We built an end-to-end, LLM-integrated pipeline that automates context-specific audio generation and binding directly within immersive XR narratives.

Conclusion and Contribution

- **End-to-End Modular Pipeline:** We built an end-to-end, LLM-integrated pipeline that automates context-specific audio generation and binding directly within immersive XR narratives.

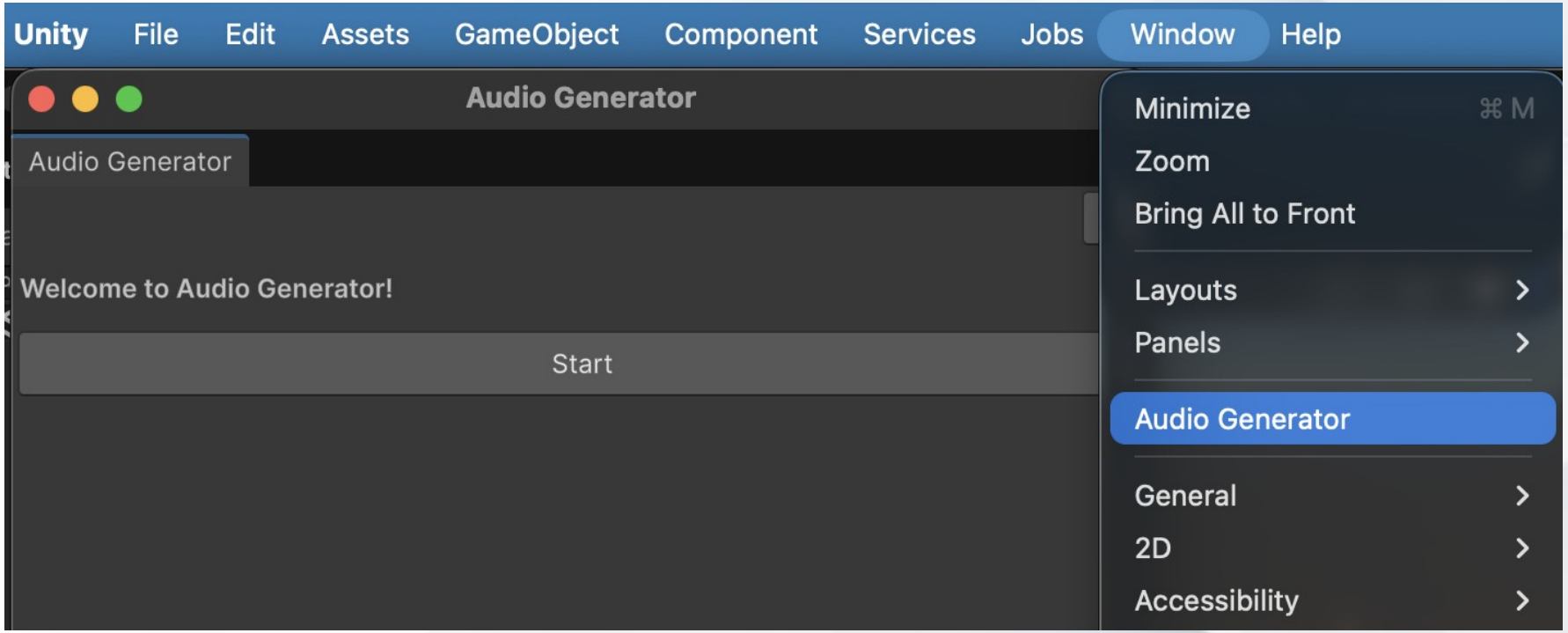
- **Semantic Structure Over Video:** Utilizing Semantic Representation (SR) significantly improves narrative-audio alignment and immersion compared to video-based inputs.

Conclusion and Contribution

- **End-to-End Modular Pipeline:** We built an end-to-end, LLM-integrated pipeline that automates context-specific audio generation and binding directly within immersive XR narratives.
- **Semantic Structure Over Video:** Utilizing Semantic Representation (SR) significantly improves narrative-audio alignment and immersion compared to video-based inputs.
- **High Developer Utility:** The system is validated as a viable authoring tool that reduces manual effort while preserving creative control in large-scale immersive storytelling.

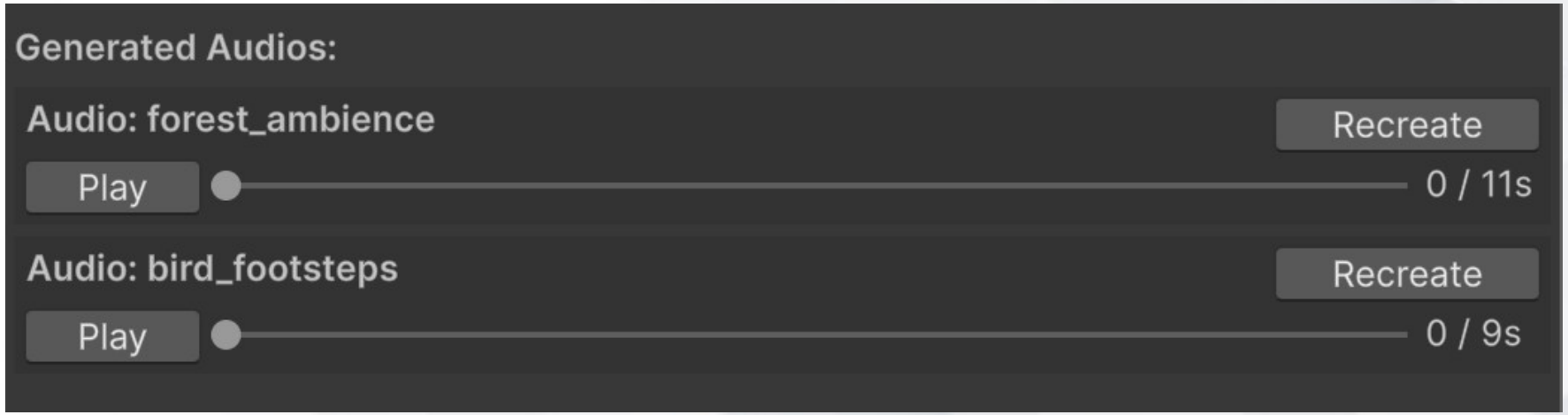
Future Work

- **Runtime Introspection:** Develop a custom mechanism to automatically derive semantic representations by capturing narrative events directly from the live XR environment.



Future Work

- **Author-Facing Interface:** Build a dedicated user interface for creators to preview, refine, and maintain creative control over the generated sounds.



İTÜ

Event-Aware Audio Generation for LLM-Driven Storytelling in Extended Reality

Thank You For Listening

24.05.2026

Mehmet Karaaslan (karaaslan18@itu.edu.tr)