

Arxiv Paper

**Project Page** 

**Code & Data** 

# MOSPA: Human Motion Generation Driven by Spatial Audio

Shuyang Xu\*, Zhiyang Dou\*†, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang†, Taku Komura†

**Evaluation Results** 

NEURAL INFORMATION PROCESSING SYSTEMS

Because it's not just about what they hear, hut where they hear it from!

MOSPA is a diffusion model. Think of it like a sculptor starting with a block of marble and carving out a masterpiece! it learns to reverse a process of adding 'noise' to create clean, realistic motion.

takes a random, 'noisy' motion and gradually refines it, guided by the audio context, into a

MOSPA

**Users Prefer MOSPA** 

■ MOSPA (Votes ■ All Other Methods Combine

Human Intent | Alignment

MOSPA: Teaching Virtual Humans to Hear and React in 3D Space

The Problem: Existing Al Missed the "Where" - Previous models ignored the crucial spatial information encoded in sound, leading to unrealistic reactions.

**Quick Overview** 

Okay, you have the data. Now, how does your special gadget, MOSPA, work?

That sounds cool, but is it really better than other models?

Amazing! What's the next step for this

Absolutely! In sideby-side comparisons, MOSPA's results are far more realistic.

We're just getting started! The future is

HKU · Shanghai AI Lab · HKUST · MUST · Shanghai Tech · TAMU

### Introduction

Human reactions fundamentally depend on the spatial direction and semantics of sound. Enabling virtual humans to react realistically to audio in a 3D environment is a critical, under-explored challenge.



We introduce a novel task, **Spatial Audio-Driven Human Motion Synthesis:** 

- We first curate the SAM (Spatial Audio-Driven Human Motion) dataset, a comprehensive new resource of paired motion and spatial audio data.
- We then present MOSPA (MOtion generation driven by SPatial Audio), a robust, diffusion-based model that effectively leverages complex spatial information. The model achieves SOTA performance against existing baselines.

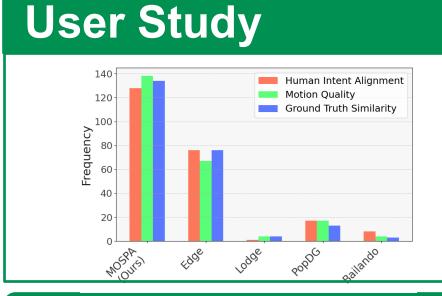
# Description: Look for the sound source and approach upon hearing miaow at the left-hand side. Description: Step away with ears covered upon hearing crowd yelling at the back. Description: Run away from the sound source upon hearing a gunshot on the right-hand side.

Dataset	SSL	$3 \mathrm{DJoint_{pos/rot}}$	Model	Joints	Subjects	Seconds
Dance with Melody	×	√/×	-	21	-	5640
DanceNet	×	√/×	-	55	2	3472
AIST++	×	$\sqrt{ }$	COCO/SMPL	17/24	30	18694
PopDanceSet	×	$\checkmark$ 1 $\checkmark$	COCO/SMPL	17/24	132	12819
FineDance	×	$\checkmark$ 1 $\checkmark$	SMPL+Hand	52	27	52560
SAM (Ours)	<b>√</b>	<b>√</b> 1 <b>√</b>	SMPL-X	55	12	34356

The **SAM** dataset provides paired human motion and spatial audio, covering 12 subjects and 34,356 seconds of motion capture at 120 FPS. It spans 27 spatial audio scenarios and 49 human reaction types across genres. The blue sphere marks the sound source.

# $\begin{array}{|c|c|c|c|c|} \hline Method & R-precision \uparrow & FID \downarrow & Diversity \to & APD \to \\ \hline \hline Real Motion & 1.000^{\pm0.000} & 1.000^{\pm0.000} & 1.000^{\pm0.000} & 0.001 & 23.616^{\pm0.188} & 59.435 \\ \hline EDGE & 0.886^{\pm0.005} & 0.960^{\pm0.003} & 0.977^{\pm0.002} & 13.993 & 23.099^{\pm0.196} & 43.882 \\ POPDG & 0.762^{\pm0.006} & 0.886^{\pm0.005} & 0.934^{\pm0.003} & 20.967 & 22.536^{\pm0.170} & 34.996 \\ I.ODGE & 0.444^{\pm0.006} & 0.594^{\pm0.005} & 0.679^{\pm0.004} & 102.289 & 21.101^{\pm0.141} & 11.801 \\ Bailando & 0.077^{\pm0.003} & 0.134^{\pm0.003} & 0.189^{\pm0.004} & 168.396 & 17.347^{\pm0.247} & 23.121 \\ MOSPA (Ours) & 0.937^{\pm0.005} & 0.984^{\pm0.002} & 0.996^{\pm0.001} & 7.981 & 23.575^{\pm0.188} & 53.915 \\ \hline Audio & Expected & EDGE & POPDG & LODGE & Bailando & MOSPA (Ours) \\ Bicycle Ring & Upon hearing a bicycle ring & Very limit & Ver$

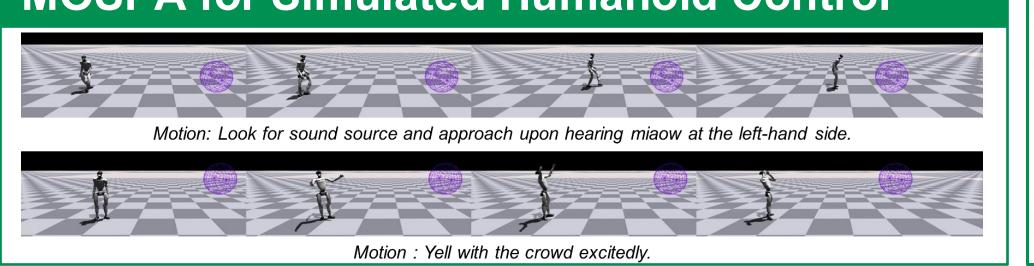
Motion: Walk back and look around upon hearing vehicle horn.



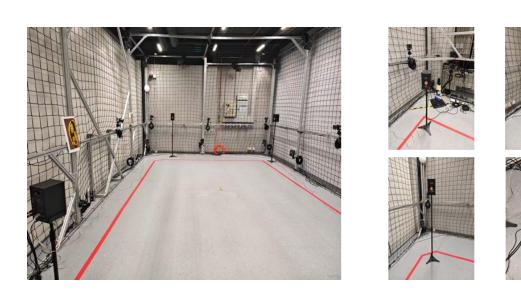
We conducted a user study with 25 participants to assess the perceptual quality of motion generation.

User study results: MOSPA outperforms other methods in intent alignment, motion quality, and similarity to ground truth. The bar chart shows the vote distribution across methods.

# **MOSPA for Simulated Humanoid Control**



## **Mocap Environment**



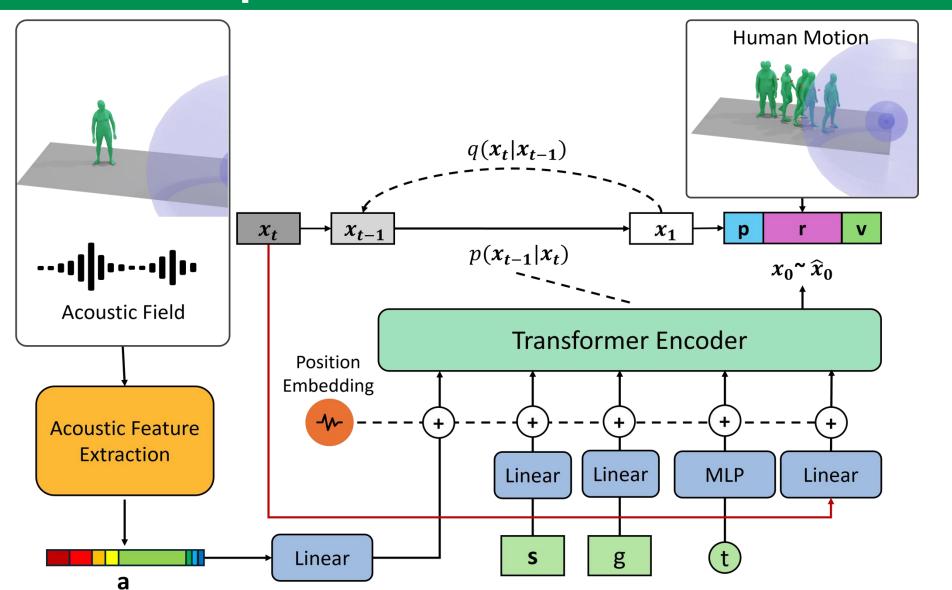
Vicon mocap system: 28 cameras, 4 speakers, 120 Hz.





Two microphones are placed at the ear positions and connected to a Deity PR-2 stereo audio recorder.





MOSPA is an encoder-only transformer diffusion model for spatial audio-conditioned human motion synthesis.

### **Inputs:**

- Noisy Motion  $x_t$
- Acoustic Feature *a*
- Sound Source Location **s**
- Motion Genre *g*
- Diffusion Timestep *t*

### **Output:**

• Predicted Motion  $\hat{x}_0$