🎧 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†

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

Project Page | Code & Data | Arxiv Paper

NEURAL INFORMATION PROCESSING SYSTEMS

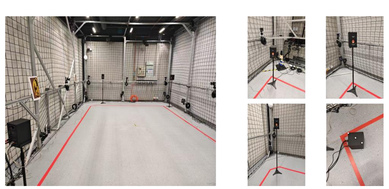
Introduction

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.

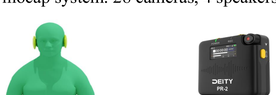
We introduce a novel task, **Spatial Audio-Driven Human Motion Synthesis:**

- 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 (*MO*tion generation driven by *SP*atial *A*udio)**, a robust, diffusion-based model that effectively leverages complex spatial information. The model achieves **SOTA** performance against existing baselines.

Mocap Environment



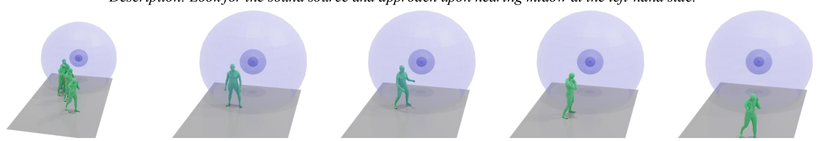
Vicon mocap system: 28 cameras, 4 speakers, 120 Hz.



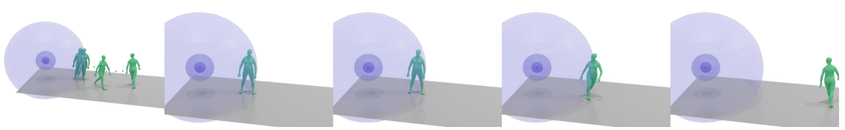
Two microphones are placed at the ear positions and connected to a Deity PR-2 stereo audio recorder.

SAM Dataset

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.

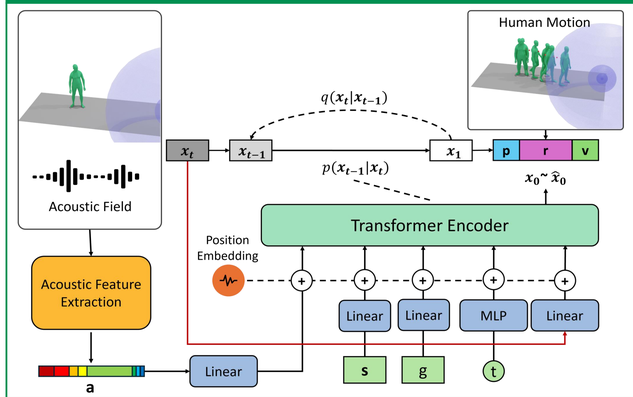


Description: Run away from the sound source upon hearing a gunshot on the right-hand side.

Dataset	SSL	3DJoint$_{pos/rot}$	Model	Joints	Subjects	Seconds
Dance with Melody	×	✓/×	-	21	-	5640
DanceNet	×	✓/×	-	55	2	3472
AIST++	×	✓/✓	COCO/SMPL	17/24	30	18694
PopDanceSet	×	✓/✓	COCO/SMPL	17/24	132	12819
FineDance	×	✓/✓	SMPL+Hand	52	27	52560
SAM (Ours)	✓	✓/✓	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.

MOSPA Pipeline



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

Inputs:

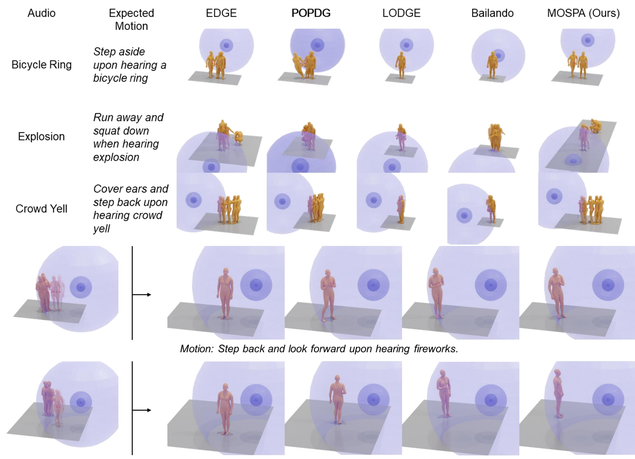
- Noisy Motion x_t
- Acoustic Feature a
- Sound Source Location s
- Motion Genre g
- Diffusion Timestep t

Output:

- Predicted Motion $\hat{x}_0$

Evaluation Results

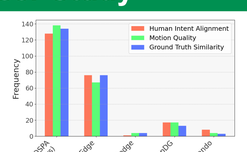
Method	R-precision ↑			FID ↓	Diversity →	APD →
	Top1	Top2	Top3			
Real Motion	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	0.001	$23.616^{\pm 0.188}$	59.435
EDGE	$0.886^{\pm 0.005}$	$0.960^{\pm 0.003}$	$0.977^{\pm 0.002}$	13.993	$23.099^{\pm 0.196}$	43.882
POPDG	$0.762^{\pm 0.006}$	$0.886^{\pm 0.005}$	$0.934^{\pm 0.003}$	20.967	$22.536^{\pm 0.170}$	34.996
LODGE	$0.444^{\pm 0.006}$	$0.594^{\pm 0.005}$	$0.679^{\pm 0.004}$	102.289	$21.101^{\pm 0.141}$	11.801
Bailando	$0.077^{\pm 0.003}$	$0.134^{\pm 0.003}$	$0.182^{\pm 0.004}$	168.396	$17.347^{\pm 0.247}$	23.121
MOSPA (Ours)	$\mathbf{0.937}^{\pm \mathbf{0.005}}$	$\mathbf{0.984}^{\pm \mathbf{0.002}}$	$\mathbf{0.996}^{\pm \mathbf{0.001}}$	**7.981**	$\mathbf{23.575}^{\pm \mathbf{0.188}}$	**53.915**



Motion: Step back and look forward upon hearing fireworks.

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

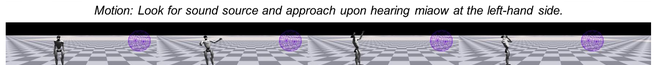
User Study



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

Motion: Look for sound source and approach upon hearing miaow at the left-hand side.



Motion : Yell with the crowd excitedly.

Quick Overview