

A Dataset Statistics

We present more statistics on our dataset regarding the audio scenarios and the the motion types implemented in the collection of the dataset.

Spatial Audio Scenarios The SAM dataset covers 27 common spatial audio scenarios, as listed in Tab. A1. Each scenario is accompanied by two or three audio clips and for each audio clip 48 10-second motion sequences are collected.

Scenario		
aircraft	bark	bicycle bell
bird	call	cat
church bell	cough	crowd yell
drum	engine	explosion
fire alarm	firework	gunshot
insect	instrument	laughter
music	phone ringing	shout
sing	siren	speech
thunder	vehicle	wind rain

Table A1: The 27 common spatial audio settings covered in SAM, each with two or three 10-second audio clips and around 100 motion sequences.

Motion Types The SAM dataset covers 20 common reactions to spatial audio in daily life, as presented in Tab. A2. Each motion type includes at least 10 minutes of motion sequences, ensuring a balanced dataset.

Motion		
look at	stand still	cover ears
step aside	run	look around
shake	dance	wave hands
change trajectory	squat	nod head
clap hands	laugh	stretch
shout	curl	cover nose
pray	jump	

Table A2: The 20 common reactions excluding the motion genres covered in SAM.

Motion Genres Three general motion genres are covered in the SAM dataset—dull, neutral, and sensitive. The intensity and reaction speed decreased from sensitive to neutral to dull. The motion types listed in Tab. A2 can thus be further divided into motion types with genres, as listed in Tab. A3, A4, A5.

Dull Motion		
change trajectory	clap hands	dance
laugh	look around	look at
nod head	pray	run
shout	stand still	step aside
stretch	wave hands	

Table A3: Dull motion types.

Neutral Motion		
change trajectory	clap hands	cover ears
cover nose	curl	dance
laugh	look around	look at
nod head	pray	run
shake	stand still	step aside
stretch	wave hands	

Table A4: Neutral motion types.

Sensitive Motion		
change trajectory	clap hands	cover ears
cover nose	curl	dance
jump	laugh	look around
look at	nod head	pray
run	shake	squat
stand still	step aside	wave hands

Table A5: Sensitive motion types.

B Implementation Details

B.1 More Implementation Details

We present more details regarding the audio features extracted. The full list of audio features and their corresponding dimensions are listed in Tab. B6. During feature extraction of the audio, we use a sampling rate of 30,720 Hz, an FFT window length of 2048, and a hop length of 256. The Short-Time Fourier Transform (STFT) windowing function is the Hann window [1]. The final shape of the audio features \mathbf{a} is thus $\mathbf{a} \in \mathbb{R}^{T \times 2272}$, where $T=240$ is the number of frames in each motion sequence.

Feature Name	Dimension
MFCC	20
MFCC delta	20
constant-Q chromagram	12
STFT chromagram	12
onset strength	1
tempogram	1068
one-hot beats	1
RMS energy	1
active frames	1

Table B6: The extracted audio features for a single ear comprise a 1136-dimensional vector. The total dimension of the audio feature vector \mathbf{a} is 2272.

B.2 Feature Extractor Implementation Details

Following the feature extractor architecture from [2], our framework incorporates two bidirectional GRUs (Bi-GRUs) and a motion sequence autoencoder, as illustrated in Fig. B1. The Bi-GRUs are applied to extracting condition and motion features respectively, each employ 4 layers with a hidden size of 1024. The audio features undergo an initial projection to 1020 dimensions, followed by concatenation with the sound source location and genre information, resulting in a 1024-dimensional condition feature vector. The motion autoencoder consists of a transformer encoder-decoder structure, where each component contains 4 layers with 4 attention heads and has a latent dimension of 512. Notably, the Bi-GRU used to extract features from motion sequences takes the motion autoencoder’s output as its input.

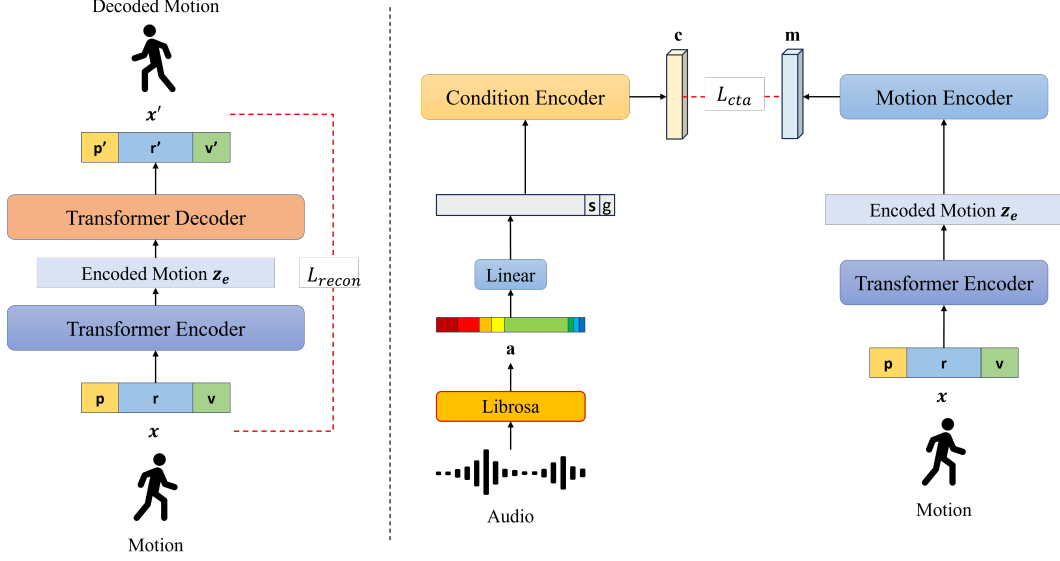


Figure B1: The feature extractor framework consists of a motion autoencoder co-trained with two Bi-GRU modules functioning as feature extractors (condition encoder and motion encoder). A reconstruction loss is applied between the ground-truth motions \mathbf{x} and the decoded motions \mathbf{x}' produced by the motion autoencoder. Additionally, a contrastive loss operates on the extracted condition features \mathbf{c} and motion features \mathbf{m} .

The feature extractor is optimized using two loss functions. Consistent with the implementation in [2], we employ a contrastive loss [3] defined as:

$$D_{\mathbf{c}, \mathbf{m}} = \|\mathbf{c} - \mathbf{m}\|_2$$

$$\mathcal{L}_{cta} = (1 - y)D_{\mathbf{c}, \mathbf{m}}^2 + (y)\max(0, m - D_{\mathbf{c}, \mathbf{m}})^2,$$

where \mathbf{c} denotes condition features, \mathbf{m} represents motion features, and $y = 0$ for matched condition-motion pairs, otherwise $y = 1$. This contrastive loss segregates the feature space by minimizing distances between matched pairs while enforcing a minimum separation margin m between mismatched pairs. We set $m = 10$ here. Additionally, we apply a reconstruction loss:

$$\mathcal{L}_{recon} = \|\mathbf{x}' - \mathbf{x}\|_2^2,$$

to guide the learning of the motion autoencoder, where \mathbf{x}' denotes reconstructed motion sequences from the transformer decoder and \mathbf{x} represents ground-truth motion sequences. These sequences maintain the identical $\mathbb{R}^{T \times 300}$ dimensionality format used during MOSPA's training.

We employ the Adam optimizer [4] with a learning rate of 5×10^{-5} . The feature extractor undergoes training for 1,500 epochs with a batch size of 64, with the motion autoencoder frozen after the initial 1,000 epochs to concentrate optimization efforts on the two Bi-GRU modules in the later 500 epochs.

R-precision For each motion sequence and its extracted motion features, a condition feature pool is constructed comprising the ground-truth matched condition feature along with 31 randomly selected mismatched condition features from the test dataset serving as distractors. The R-precision metrics are then computed by ranking the Euclidean distances between the extracted condition features and motion features, then determining the probability that the ground-truth condition appears among the top-1, top-2, and top-3 ranked positions.

Fréchet Inception Distance (FID) The Fréchet Inception Distance (FID) is computed between motion features derived from ground-truth motion sequences in the test dataset and corresponding generated motion sequences, serving as a metric for assessing the quality and fidelity of the synthesized motions.

Diversity Following the methodology in [2], diversity is computed to quantify the variance of the generated motions. Two sets of generated motions, each of size $S_d = 64$, are randomly sampled

from the test dataset: $\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{S_d}\}$ and $\{\mathbf{m}'_1, \mathbf{m}'_2, \dots, \mathbf{m}'_{S_d}\}$. The diversity metric is then calculated as:

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \|\mathbf{m}_i - \mathbf{m}'_i\|.$$

C User Study Design

We conducted a user study to evaluate the perceptual quality of motion generation. A total of 25 participants assessed five models (MOSPA, EDGE, POPDG, LODGE, and Bailando) alongside the ground truth (GT). They were asked to select the best motion based on the following criteria:

- **Human Intent Alignment:** Does the motion align with real-world intent?
- **Motion Quality:** Which motion exhibits the highest movement quality?
- **GT Similarity:** Which motion best matches the GT?

To facilitate evaluation, we provided both the GT motion and a textual description as references. Fig. C2 presents a screenshot of the user study interface.

Further Details on the User Study for Spatial Audio-driven Motion Generation.

Before starting the study, users are instructed to:

- Stay in a quiet environment;
- Wear **binaural headphones**;
- Carefully read all instructions.

Each case presents **six videos** with spatial audio:

- **Videos 1–5** are generated by different methods.
- The **rightmost video** is the **Ground Truth (GT)**, shown in green.

After watching and listening to each row of videos, participants are asked to answer:

- (1) Which video best aligns with real-world human **Intent** in motion generation?
- (2) Which video exhibits the highest **Motion Quality** of body movement?
- (3) Which video best **Matches the Ground Truth** motion (rightmost)?

The following are the 10 testing cases and their corresponding instructions: ...

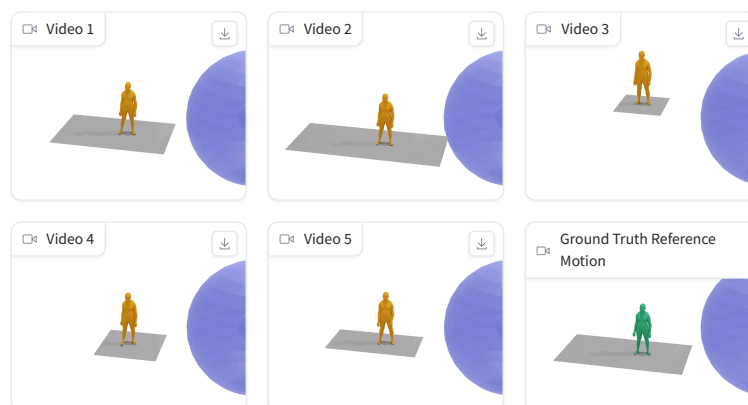
User Study: Spatial Audio Driven Human Motion Generation

Before starting the study, ensure you are in a quiet environment, wearing headphones, and carefully read the following instructions:

- You will watch six videos **with spatial audio**: the rightmost is the ground truth (GT) in green, while the remaining five are generated by different methods.
- Carefully **watch and listen** to each row of videos and answer the three corresponding questions.

Binaural headphones are required!

Case 1



Text Description:

Cover ears and step back when hearing engine (sensitive)

Among Videos 1 to 5, which one best aligns with real-world human Intent in motion generation?

☐ Video 1

☐ Video 2

☐ Video 3

☐ Video 4

☐ Video 5

Among Videos 1 to 5, which one exhibits the highest motion Quality of body movement?

☐ Video 1

☐ Video 2

☐ Video 3

☐ Video 4

☐ Video 5

Figure C2: Screenshot of the user study interface.

References

- [1] RB Blackman and JW Tukey. The measurement of power spectra dover publications. *Inc, New York*, 1958.
- [2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.
- [3] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006.
- [4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.