# SONIC-AMG: STREAMING HUMAN MOTION GENERATION FROM SPATIAL AUDIO

**Anonymous authors**
Paper under double-blind review

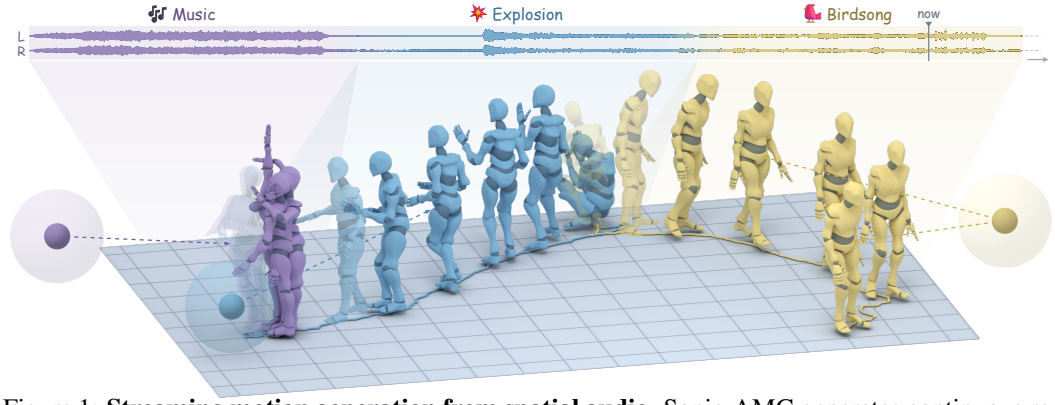



Figure 1: **Streaming motion generation from spatial audio.** Sonic-AMG generates continuous reactions to successive sounds, updating each source's position relative to the generated body. Colors link sound sources and audio segments to the corresponding motions. The unified scene shows successive motion snapshots from purple to blue to yellow. Translucent bands connect audio segments to their motion responses, while faded poses highlight motion transitions between audio segments.

## ABSTRACT

Modeling human reactions to spatial audio requires understanding both what is heard and where it comes from. Prior work generates motion offline from complete acoustic and spatial input sequences. Humans, however, react to sounds without knowing what they will hear next. We present Sonic-AMG, the first autoregressive framework for generating human motion from streaming spatial audio. The model encodes acoustic content and body-relative source geometry as separate tokens. A causal transformer with a flow head generates continuous motion latents, which a causal decoder reconstructs into motion segments. Teacher-forced training uses motion histories and body-relative source locations derived from recorded motion, whereas streaming inference relies on generated motion for both. We introduce *rehearsal training*, which combines full-sequence teacher-forced supervision with supervised prediction on histories constructed through short autoregressive self-rollouts, improving robustness to accumulated prediction errors. To keep spatial conditioning consistent with generated motion, we propose *spatial feedback*, which recomputes body-relative source locations from generated root poses during both rehearsal training and inference. On SAM, Sonic-AMG achieves state-of-the-art FID and R-precision. It further supports real-time online generation of coherent reactions across sound events and source locations. Our code and model will be released upon acceptance.

## 1 INTRODUCTION

Sound provides information about events beyond the visual field, making it an important cue for embodied interaction (Chen et al., 2020; 2021). Human reactions to sound depend on both its content and the source's location relative to the body. For example, hearing a voice from behind may prompt a listener to turn toward the speaker before the utterance ends and adjust the movement as new acoustic information arrives. As the listener turns, the source's direction relative to the body changes. Generating such reactions therefore requires responding as audio arrives and updating spatial conditions as the body moves.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent work (Xu et al., 2025) synthesizes complete motion sequences within predefined temporal windows, conditioned on fully observed acoustic and spatial inputs. Requiring the full input window prevents it from responding causally to an incoming audio stream. Autoregressive motion models provide a basis for streaming generation by predicting successive motion segments from preceding movements (Shi et al., 2024; Xiao et al., 2025). Extending autoregressive learning to spatial audio requires representations that capture both acoustic content and spatial cues. Furthermore, streaming generation requires each motion prediction to incorporate the acoustic and spatial information currently available as the input evolves.

We present Sonic-AMG, the first framework for streaming human motion generation from continuous spatial audio (Table 1). Motion is represented in the continuous latent space of a pretrained causal temporal VAE (Xiao et al., 2025). Given a binaural audio stream and the world-space sound source location (SSL), the model encodes acoustic content and body-relative source geometry as separate tokens. At each step, a causal transformer processes the incoming acoustic and spatial conditions together with preceding motion latents. Its output conditions a flow-matching head that samples the next motion latent, and a causal decoder reconstructs the corresponding motion segment. Repeating this process extends the motion as audio arrives, without requiring the full audio in advance or revising previously generated frames.

Table 1: **Method comparison.** Online SSL denotes spatial conditioning updated from generated motion.

| Method | Spatial Audio | Causal | Online SSL |
|---|---|---|---|
| EDGE (Tseng et al., 2023) | × | × | — |
| POPDG (Luo et al., 2024) | × | × | — |
| LODGE (Li et al., 2024a) | × | × | — |
| MOSPA (Xu et al., 2025) | ✓ | × | × |
| **Ours** | ✓ | ✓ | ✓ |

We treat acoustic content and spatial cues as distinct modalities of spatial-audio conditioning. We hypothesize that early fusion introduces interference between them. Sonic-AMG therefore encodes them as dedicated *what* and *where* tokens, respectively, and learns their interactions within the causal transformer. With sound source geometry supplied separately, the acoustic branch can focus on sound content rather than a comprehensive description of the binaural signal. For instance, simple log-mel features in our model yield better results than MOSPA's 2,272-dimensional audio features (Xu et al., 2025). To design the *where* token, we examine perceptual research to understand how body-relative SSL influences human reactions. We develop an 11-dimensional spatial event representation inspired by studies of nearby auditory space (Graziano et al., 1999), sound-source motion (Warren et al., 2002), and looming (Neuhoff, 1998).

With acoustic and spatial representations in place, learning the causal generator further requires addressing two sources of *training–inference mismatch*. The first concerns motion history: teacher-forced training uses ground-truth histories, whereas autoregressive inference relies on the model's own predictions. To reduce this mismatch, two-forward training replaces selected ground-truth inputs with predictions from a preliminary teacher-forced pass (Mihaylova & Martins, 2019; Xiao et al., 2025). However, these replacements are still predicted under ground-truth histories, where later replacements do not depend on earlier sampled outputs. We introduce *rehearsal training*, which combines teacher-forced supervision with supervised prediction on short histories generated autoregressively by the model. Training on self-rollouts improves robustness to accumulated prediction errors, while teacher-forced supervision provides a stable learning signal across the full sequence.

The second mismatch arises in spatial conditioning. While acoustic content is directly supplied by the incoming audio stream, body-relative SSL depends on the root position and orientation. MOSPA uses complete SSL sequences precomputed from recorded motion, which encode information about future ground-truth root motion. Such conditioning is incompatible with causal streaming generation, which requires SSL to be updated from the generated root. Training with SSL derived from ground-truth motion therefore creates a mismatch with the spatial conditions encountered at inference. Our experiments show that this mismatch substantially degrades generation quality. We therefore introduce *spatial feedback*, which recomputes body-relative source coordinates after each generated segment during both rehearsal training and inference. The updated coordinates condition the next prediction, keeping spatial conditions consistent with the generated motion history.

On SAM (Xu et al., 2025), Sonic-AMG achieves state-of-the-art FID and R-precision. It further supports causal and online motion generation from streaming spatial audio in real time. Although

trained on individual audio–motion clips, Sonic-AMG can generate coherent motion across successive sound events with varying content and source locations (Fig. 1), demonstrating robust streaming generation. Our contributions are:

- We present Sonic-AMG, the first framework for causal human motion generation from streaming spatial audio, with dedicated *what* and *where* tokens for acoustic content and spatial cues.
- We introduce *rehearsal training* to address the history mismatch between training and inference, combining teacher-forced supervision with supervised prediction from histories constructed through short autoregressive self-rollouts.
- We propose *spatial feedback* to address the mismatch in spatial conditioning, updating body-relative source coordinates from generated motion during both rehearsal training and inference.

## 2 Related Work

**Audio-conditioned motion generation.** Music-driven dance methods use autoregressive models to align motion with musical structure, either in pose space or through learned choreographic codes (Li et al., 2021; Siyao et al., 2022). Diffusion-based approaches support diverse choreography, motion editing, and hierarchical synthesis of long dance sequences (Alexanderson et al., 2023; Tseng et al., 2023; Luo et al., 2024; Li et al., 2024a). Co-speech gesture models combine acoustic, textual, and speaker or style information to generate gestures that match speech content and rhythm (Yoon et al., 2020; Ao et al., 2023). Diffusion models and learned motion priors capture gesture diversity while coordinating movements across body parts (Zhu et al., 2023; Ao et al., 2023; Liu et al., 2024; Chen et al., 2024b). Streaming systems extend these tasks to incrementally available music and speech (Saleem et al., 2026; Chen et al., 2026; Vu et al., 2026; Ji et al., 2026; Liu et al., 2026). MOSPA (Xu et al., 2025) introduces the SAM dataset of human reactions to spatial audio and a diffusion model conditioned on binaural audio, source coordinates, and reaction style. Its full-sequence conditioning requires future observations, precluding causal reactions to an incoming audio stream. Sonic-AMG instead generates motion causally and in real time as spatial audio arrives.

**Causal motion generation.** Autoregressive (AR) models use next-token prediction for language modeling (Brown et al., 2020) and image synthesis with discrete visual tokens (Esser et al., 2021). To model continuous latents without vector quantization, image generators (Li et al., 2024b; Ren et al., 2025) and multimodal language models (Sun et al., 2026) pair an AR transformer with a conditional diffusion or flow head. In motion synthesis, discrete-token models predict quantized motion codes (Zhang et al., 2023; Jiang et al., 2023), whereas AR diffusion models sample successive frames or motion primitives (Shi et al., 2024; Zhao et al., 2025; Tevet et al., 2025; Zhang et al., 2025; Zhao et al., 2026). Streaming generation in latent space requires decoding segments without future latents. Systems combining causal decoding with transformer-conditioned generative heads support online text control, two-character interaction, and co-speech gestures (Xiao et al., 2025; Cen et al., 2025; Li et al., 2026; Liu et al., 2026). Causal decoding alone does not resolve exposure bias: teacher forcing uses recorded histories, whereas inference relies on model predictions. Diffusion forcing exposes models to contexts at varying noise levels (Chen et al., 2024a; Cai et al., 2026; Ji et al., 2026), while other strategies introduce model-generated histories (Bengio et al., 2015; Zhao et al., 2025; Huang et al., 2025). Two-forward training preserves parallel computation using replacement inputs from a preliminary teacher-forced pass (Mihaylova & Martins, 2019; Xiao et al., 2025). These replacements do not recursively depend on earlier sampled outputs. Our rehearsal training retains teacher forcing alongside short autoregressive self-rollouts. The combined objective accounts for error propagation across generated segments while retaining sequence-wide supervision.

## 3 Method

Given an incoming binaural audio stream $A$ and a known world-space sound-source position, Sonic-AMG generates motion autoregressively from an initial motion segment, without access to future audio or motion (Fig. 2). We represent motion with continuous latents and encode acoustic content and body-relative source geometry as separate conditioning tokens (Sec. 3.1). A causal transformer conditions a flow head to generate the next motion latent from these tokens and the motion history (Sec. 3.2). We train the generator using rehearsal training, which combines full-sequence

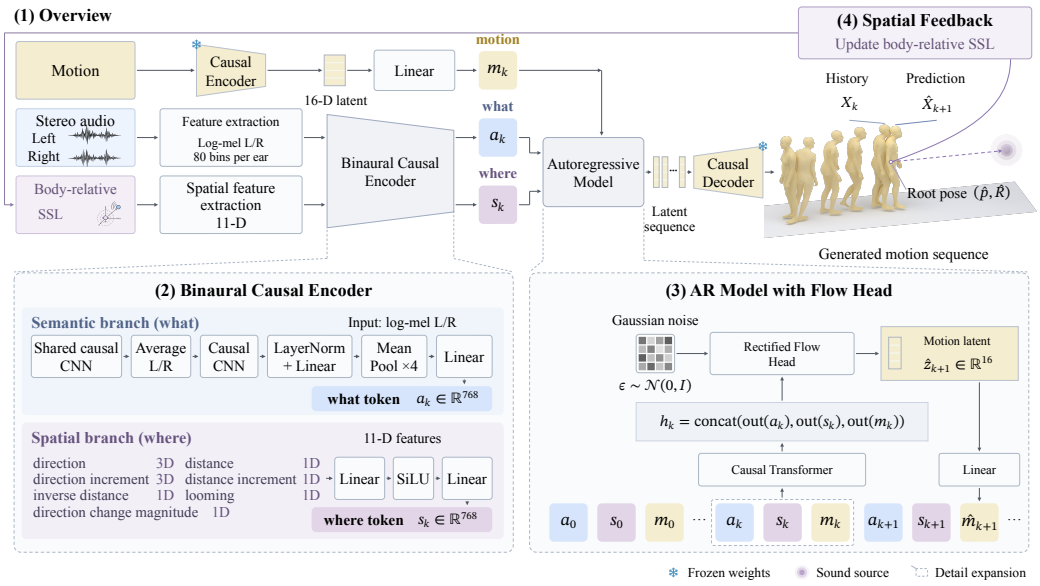



Figure 2: **Sonic-AMG architecture.** A causal transformer and flow head generate motion latents from interleaved acoustic (*what*), spatial (*where*), and motion tokens. Spatial feedback updates body-relative SSL from decoded root poses. $X_k$, $z_k$, and $m_k$ denote a four-frame motion block, its latent, and its projected token, respectively.

teacher-forced supervision with supervised prediction under histories constructed through short autoregressive self-rollouts (Sec. 3.3). We further apply spatial feedback to update body-relative SSL from the generated root pose during both rehearsal training and inference, keeping spatial conditions consistent with the generated motion history (Sec. 3.4).

### 3.1 Motion and Conditioning Representations

**Motion representation.** We represent motion at 30 Hz. Each frame $\mathbf{x}_t \in \mathbb{R}^{300}$ comprises the global 3D positions, local rotations in a 6D representation, and 3D velocities of 25 joints (Xu et al., 2025). We adopt MotionStreamer's causal temporal VAE (Xiao et al., 2025) and pretrain it on SAM (Xu et al., 2025). The VAE encodes motion into a sequence of $K$ latents $\mathbf{z}_k \in \mathbb{R}^{d_z}$ with $d_z = 16$ at a temporal stride of $r = 4$. Each latent $\mathbf{z}_k$ corresponds to a four-frame motion block $\mathcal{I}_k = \{rk, \ldots, rk + r - 1\}$. The encoder and decoder remain frozen during generator training.

For spatial audio, we regard acoustic content and source geometry as two distinct conditioning modalities. We hypothesize that the two modalities can act as noise for one another when fused into a single conditioning representation. We therefore encode them as separate tokens.

**Semantic branch: what.** The semantic branch in Figure 2(2) learns acoustic features for motion prediction from the incoming binaural stream. We represent each ear's spectral content with 80-dimensional log-mel features and process them with a shared causal temporal CNN. The shared encoder maps both ears into a common feature space, where we average the two streams to emphasize sound content shared across ears. A two-layer causal temporal CNN then aggregates recent acoustic context, followed by layer normalization and a linear projection. Mean pooling over each four-frame block produces $\mathbf{f}_k$ at the motion latent rate, so that each acoustic token aligns with a motion block. We obtain the acoustic token as $\mathbf{a}_k = W_a \mathbf{f}_k + \mathbf{e}_c$, where $W_a$ is a learned linear projection and $\mathbf{e}_c$ is a learned embedding indexed by SAM's annotated reaction-style label $c$.

**Spatial branch: where.** The spatial branch in Figure 2(2) encodes source location and relative motion. Our descriptor reflects perceptually relevant aspects of spatial sound: proximity (Graziano et al., 1999), relative motion (Warren et al., 2002), and approach (Neuhoff, 1998). We assume a sound source at known world-space position $\mathbf{g}$. For root position $\mathbf{p}_t$ and orientation $\mathbf{R}_t \in SO(3)$, the egocentric source location is $\boldsymbol{\ell}_t = \mathbf{R}_t^\top (\mathbf{g} - \mathbf{p}_t)$. We average source locations within each block,

$\bar{\boldsymbol{\ell}}_k = r^{-1} \sum_{t \in \mathcal{I}_k} \boldsymbol{\ell}_t$, and define distance $\rho_k = \|\bar{\boldsymbol{\ell}}_k\|_2$ and direction $\mathbf{d}_k = \bar{\boldsymbol{\ell}}_k / \rho_k$. The resulting 11-dimensional descriptor contains direction, distance, their increments, and three derived features:

$$\boldsymbol{\phi}_k = \left[ \mathbf{d}_k, \rho_k, \Delta\mathbf{d}_k, \Delta\rho_k, \rho_k^{-1}, \|\Delta\mathbf{d}_k\|_2, \frac{\max(-\Delta\rho_k, 0)}{\rho_k} \right]. \tag{1}$$

For $k > 0$, the direction increment $\Delta\mathbf{d}_k = \mathbf{d}_k - \mathbf{d}_{k-1}$ and distance increment $\Delta\rho_k = \rho_k - \rho_{k-1}$ measure changes between consecutive motion blocks. The final three features are inverse distance, direction change magnitude, and looming, respectively. Inverse distance emphasizes proximity, while direction change magnitude summarizes changes in source bearing. We represent looming by the decrease in distance per block normalized by the current distance, with zero assigned when the distance does not decrease. The spatial token is $\mathbf{s}_k = E_s((\boldsymbol{\phi}_k - \boldsymbol{\mu}_\phi)/\boldsymbol{\sigma}_\phi) + \mathbf{e}_c$, where $E_s$ is a two-layer MLP and $\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi$ are the training-set mean and standard deviation. At inference, acoustic tokens are computed from the incoming audio, while spatial feedback updates spatial tokens from the generated body state (Section 3.4).

### 3.2 Causal Motion Generation

Given an initial motion latent $\mathbf{z}_0$, we predict each subsequent latent $\mathbf{z}_{k+1}$ from the motion history $\mathbf{z}_{0:k}$, conditioned on acoustic and spatial tokens up to block $k$:

$$p_\theta(\mathbf{z}_{1:K-1} \mid A, \mathbf{g}, c, \mathbf{z}_0) = \prod_{k=0}^{K-2} p_\theta(\mathbf{z}_{k+1} \mid \mathbf{z}_{0:k}, \mathbf{a}_{0:k}, \mathbf{s}_{0:k}, c), \tag{2}$$

where $\theta$ denotes the generator parameters.

We use a learned linear projection $W_m$ to obtain motion tokens $\mathbf{m}_k = W_m \mathbf{z}_k$ and feed the interleaved sequence to a causal transformer (Figure 2(3)). We concatenate the transformer outputs at the three token positions of block $k$ to form the conditioning vector $\mathbf{h}_k$ for the flow head:

$$\mathbf{U}_{0:k} = [\mathbf{a}_0, \mathbf{s}_0, \mathbf{m}_0, \ldots, \mathbf{a}_k, \mathbf{s}_k, \mathbf{m}_k], \qquad \mathbf{h}_k = [\mathbf{o}_k^a; \mathbf{o}_k^s; \mathbf{o}_k^m], \tag{3}$$

where $\mathbf{o}_k^a, \mathbf{o}_k^s, \mathbf{o}_k^m$ denote the corresponding transformer outputs. At inference, we reuse cached keys and values for the history and process only the new token triplet at each step.

Conditioned on $\mathbf{h}_k$, a rectified flow head samples the next motion latent from Gaussian noise. The head predicts a velocity field $v_\theta$, which we train with conditional flow matching (Liu et al., 2023; Lipman et al., 2023). For a next-latent target $\mathbf{z}$ and conditioning vector $\mathbf{h}$, the interpolation at flow time $\tau \in [0, 1]$ is $\mathbf{u}_\tau = (1 - \tau)\boldsymbol{\epsilon} + \tau\mathbf{z}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. We sample $\tau$ from a logit-normal distribution and minimize

$$\ell_{\mathrm{FM}}(\mathbf{z}, \mathbf{h}) = \mathbb{E}_{\boldsymbol{\epsilon}, \tau} \left[ \frac{1}{d_z} \|v_\theta(\mathbf{u}_\tau, \tau; \mathbf{h}) - (\mathbf{z} - \boldsymbol{\epsilon})\|_2^2 \right]. \tag{4}$$

Sampling integrates the learned velocity field from $\tau = 0$ to 1 with ten Euler steps. We write the sampled latent as $\widehat{\mathbf{z}}_{k+1} = G_\theta(\boldsymbol{\epsilon}; \mathbf{h}_k)$, where $G_\theta$ denotes the flow sampler. At inference, classifier-free guidance (Ho & Salimans, 2022) combines velocity predictions with and without the current acoustic and spatial conditions.

We project the sampled latent $\widehat{\mathbf{z}}_{k+1}$ to a motion token and append it to the motion history. The causal decoder reconstructs the new motion segment from the extended latent sequence without revising earlier frames. Spatial feedback then recomputes body-relative source geometry from the decoded root pose for the next prediction (Section 3.4).

### 3.3 Rehearsal training

In two-forward training (Mihaylova & Martins, 2019; Xiao et al., 2025), an initial teacher-forced pass estimates motion latents from GT histories. Selected estimates then replace GT motion inputs for a second supervised pass. Although this exposes the model to predicted inputs, those inputs are not generated recursively from earlier predictions, as they would be during autoregressive inference.

Rehearsal training instead constructs motion histories through short autoregressive self-rollouts. Each sampled latent extends the history used for the next prediction. We combine next-latent supervision under these recursively generated histories with full-sequence teacher-forced supervision. Both objectives use the corresponding GT next latent as the prediction target.

**Teacher-forced prediction and initialization.** A teacher-forced pass over the full sequence yields the loss

$$\mathcal{L}_{\mathrm{TF}} = \frac{1}{K-1} \sum_{k=0}^{K-2} \ell_{\mathrm{FM}}(\mathbf{z}_{k+1}^{\mathrm{GT}}, \mathbf{h}_k^{\mathrm{GT}}). \tag{5}$$

Here, $\mathbf{h}_k^{\mathrm{GT}}$ is computed from the GT motion history $\mathbf{z}_{0:k}^{\mathrm{GT}}$ and the corresponding acoustic and spatial tokens. We sample a start $s$ for an $H$-transition rollout independently for each sequence. A detached denoising estimate from the same teacher-forced pass initializes $\overline{\mathbf{z}}_s$. The transformer cache and acoustic context are reset at $s$.

**Recursive context construction.** At step $k$, the rollout history $\overline{\mathbf{z}}_{s:k}$ and the corresponding acoustic and spatial token sequences determine $\overline{\mathbf{h}}_k$. We compute the flow-matching loss against $\mathbf{z}_{k+1}^{\mathrm{GT}}$ under this context. We then independently sample a latent under the same context and append it to the rollout history:

$$\overline{\mathbf{z}}_{k+1} = \mathrm{sg}\big[G_\theta(\boldsymbol{\epsilon}_k; \overline{\mathbf{h}}_k)\big], \qquad \overline{\mathbf{z}}_{s:k+1} = [\overline{\mathbf{z}}_{s:k}, \overline{\mathbf{z}}_{k+1}], \tag{6}$$

where sg stops gradients and $\boldsymbol{\epsilon}_k$ is fresh Gaussian noise. We average the flow-matching losses over the $H$ transitions and combine the two objectives with equal weight:

$$\mathcal{L}_{\mathrm{reh}} = \frac{1}{H} \sum_{k=s}^{s+H-1} \ell_{\mathrm{FM}}(\mathbf{z}_{k+1}^{\mathrm{GT}}, \overline{\mathbf{h}}_k), \qquad \mathcal{L} = \tfrac{1}{2}(\mathcal{L}_{\mathrm{TF}} + \mathcal{L}_{\mathrm{reh}}). \tag{7}$$

The teacher-forced term supervises the full recorded sequence, while the rollout term supervises continuations of increasingly long sampled histories. We use $H = 16$ and unguided sampling during rehearsal. Gradients from both objectives update the conditioning encoders, transformer, and flow head. Gradients propagate through cached transformer states but not through sampled latents or the sampling procedure.

### 3.4 Spatial feedback

Spatial feedback recomputes the source's body-relative coordinates from generated motion during both training rollouts and inference (Figure 2(4)). Given the decoder input prefix $\mathbf{z}_{0:k}$, the causal decoder $\mathcal{D}_\psi$ with frozen parameters $\psi$ reconstructs block $k$, whose root positions $\widehat{\mathbf{p}}_t$ and orientations $\widehat{\mathbf{R}}_t$ determine the updated source coordinates:

$$\widehat{\mathbf{x}}_{\mathcal{I}_k} = \mathcal{D}_\psi(\mathbf{z}_{0:k})|_{\mathcal{I}_k}, \qquad \widehat{\boldsymbol{\ell}}_t = \widehat{\mathbf{R}}_t^\top(\mathbf{g} - \widehat{\mathbf{p}}_t), \quad t \in \mathcal{I}_k. \tag{8}$$

We average $\widehat{\boldsymbol{\ell}}_t$ within block $k$ and recompute the descriptor in Eq. (1) from the current and preceding pooled locations. During training rollouts, the recomputed geometry is detached before spatial encoding. The spatial encoder in Section 3.1 maps this descriptor to $\mathbf{s}_k$, which conditions the prediction of block $k+1$.

## 4 Experiments

**Dataset.** We evaluate on SAM (Xu et al., 2025), which pairs spatial audio with human reactions. Autoregressive models receive four initial ground-truth motion frames and the reaction style.

**Spatial conditioning inputs.** We consider two spatial-conditioning protocols: **GT** uses body-relative SSL precomputed from recorded motion, whereas **Online** recomputes SSL from generated root poses during inference. GT SSL relies on ground-truth root information that is unavailable during causal streaming generation. We include a GT SSL variant to enable fair comparison with MOSPA and other baselines under their established evaluation protocol. Online evaluation is more challenging, as inaccuracies in generated root poses perturb the SSL used to condition subsequent predictions, introducing an additional pathway for error propagation. All qualitative results of Sonic-AMG shown in the experiments are generated under the Online protocol.

Following the established SAM evaluation protocol, SSL inputs are available to all methods in our main comparisons. Such inputs can be supplied by dataset annotations, an upstream localization module (Krause et al., 2024a;b), or virtual-scene metadata. Our modular conditioning design also

Table 2: **Quantitative comparison on SAM.** Causal denotes support for causal generation. † denotes results from Xu et al. (2025).

| Method | Causal | FID↓ | R@1↑ | R@2↑ | R@3↑ | Diversity→ |
|---|---|---|---|---|---|---|
| GT motion† | — | 0.001 | $1.000^{\pm 0.000}$ | $1.000^{\pm 0.000}$ | $1.000^{\pm 0.000}$ | $23.616^{\pm 0.188}$ |
| EDGE† | × | 13.993 | $0.886^{\pm 0.005}$ | $0.960^{\pm 0.003}$ | $0.977^{\pm 0.002}$ | $23.099^{\pm 0.196}$ |
| POPDG† | × | 20.967 | $0.762^{\pm 0.006}$ | $0.886^{\pm 0.005}$ | $0.934^{\pm 0.003}$ | $22.536^{\pm 0.170}$ |
| LODGE† | × | 102.289 | $0.444^{\pm 0.006}$ | $0.594^{\pm 0.005}$ | $0.679^{\pm 0.004}$ | $21.101^{\pm 0.141}$ |
| Bailando† | × | 168.396 | $0.077^{\pm 0.003}$ | $0.134^{\pm 0.003}$ | $0.182^{\pm 0.004}$ | $17.347^{\pm 0.247}$ |
| MOSPA | × | <u>8.242</u> | <u>$0.942^{\pm 0.004}$</u> | <u>$0.984^{\pm 0.002}$</u> | <u>$0.994^{\pm 0.001}$</u> | **$23.412^{\pm 0.169}$** |
| Sonic-AMG (Ours) | ✓ | **6.505** | **$0.957^{\pm 0.004}$** | **$0.991^{\pm 0.002}$** | **$0.998^{\pm 0.001}$** | <u>$23.241^{\pm 0.158}$</u> |

supports a No-SSL variant that derives spatial tokens directly from binaural audio, without requiring SSL inputs at inference, detailed in Appendix D.

**Evaluation metrics.** Following MOSPA (Xu et al., 2025), we report Fréchet Inception Distance (FID), R-precision (R@1–3), and feature-space diversity. Lower FID indicates better distributional fidelity, while higher R-precision indicates better conditioning alignment. Diversity closer to real motion is preferred. Error bars denote 95% confidence intervals; bold and underlined values mark the best and second-best results within each comparison.

### 4.1 Comparison with Existing Methods

**Quantitative comparison.** Sonic-AMG supports real-time streaming generation (Appendix C) while achieving the best distributional fidelity and conditioning alignment among the compared methods under the GT SSL protocol (Table 2). We attribute this advantage to the complementary roles of our separate conditioning representations and rehearsal training. Dedicated conditioning tokens allow each motion prediction to distinguish acoustic content from body-relative source geometry, supporting reactions that account for both cues. Rehearsal training combines sequence-wide supervision with local autoregressive consistency. Teacher forcing provides supervision across complete motion sequences, supporting motion fidelity and long-range temporal structure, while short self-rollouts expose the model to the temporal consequences of its own predictions. The joint objective directly targets the recursive prediction process required for streaming generation, which is absent from MOSPA's full-sequence training. The ablations in Sections 4.2 and 4.4 support the contributions of the representation and training.

**Qualitative comparison.** Sonic-AMG produces recognizable and spatially appropriate reactions under the Online protocol (Fig. 3). It covers the ears in response to crowd noise and combines movement away from the source with a lowered posture for an explosion, whereas several baselines exhibit limited postural change or displacement in these examples. These comparisons illustrate the model's ability to coordinate the type of reaction with movement relative to the sound source. Additional visual results are provided in Appendix B, with videos available in the supplementary material.

**Generation over complete audio clips.** Sonic-AMG generates complete reactions to audio clips of different durations using a single trained model (Fig. 4). While MOSPA is restricted to a fixed output window, Sonic-AMG extends motion as audio arrives. Postural changes and body displacement can therefore unfold throughout the audio, yielding coherent and complete reaction sequences.

**Physical simulation.** We retarget reactions to Unitree G1 with GMR (Araújo et al., 2026) and track them with pretrained GEAR-SONIC (Luo et al., 2026) in MuJoCo (Todorov et al., 2012) (Fig. 5).

### 4.2 Rehearsal training

**Comparison with two-forward training.** Rehearsal training improves motion fidelity and conditioning alignment over all two-forward schedules under both SSL protocols (Table 3). All variants share the generator architecture, and Online two-forward baselines also update spatial conditions. The gains support recursive history construction, which exposes the model to error propagation across generated segments that teacher-forced replacements do not capture.

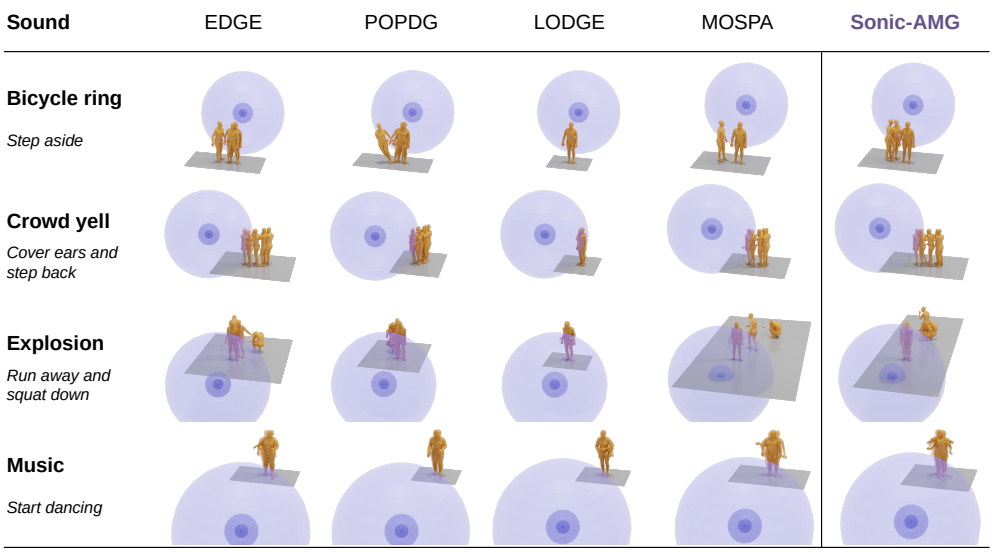



Figure 3: **Qualitative comparison on SAM.** EDGE, POPDG, and LODGE panels are reproduced from Xu et al. (2025); purple markers indicate sound sources.

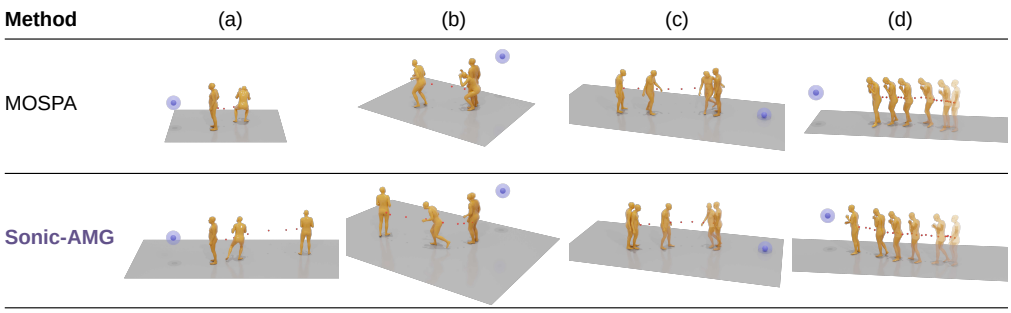
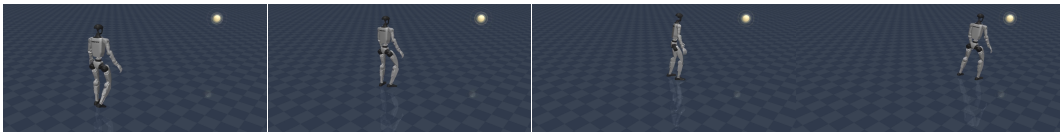
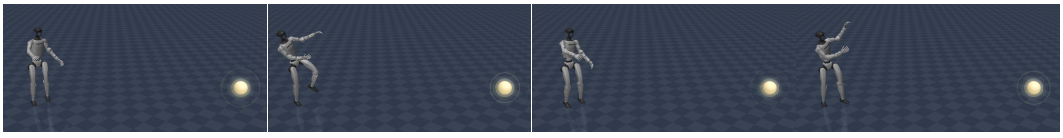



Figure 4: **Reactions over the full audio duration.** MOSPA (top) generates 8 s of motion; Sonic-AMG (bottom) covers the full 10–12 s audio clips. Six poses are sampled uniformly over each sequence's own duration. Purple markers denote sound sources; red dots trace root trajectories.



(a) Birdsong: approaching the sound source.



(b) Singing: performing arm and leg gestures.

Figure 5: **Physical simulation of generated reactions.** See Appendix B.3 for details.

**Role of teacher-forced supervision.** Self-rollouts are most effective when combined with full-sequence teacher-forced supervision. Removing this supervision degrades fidelity and alignment. Restoring the motion and conditioning prefix in Rollout Only with History (Appendix F) improves performance but does not recover rehearsal's fidelity or top-rank retrieval under either protocol. Additional context therefore cannot fully replace supervision over complete sequences.

Table 3: **Training strategies.** Each group uses the same SSL protocol for predicted training histories and evaluation. Two-forward schedules increase ($0 \to 1$), decrease ($1 \to 0$), or fix ($p = 1$) the fraction of replaced motion inputs.

| Training strategy | Eval. SSL | FID↓ | R@1↑ | R@2↑ | R@3↑ | Diversity→ |
|---|---|---|---|---|---|---|
| GT motion† | | 0.001 | $1.000^{\pm 0.000}$ | $1.000^{\pm 0.000}$ | $1.000^{\pm 0.000}$ | $23.616^{\pm 0.188}$ |
| Two-forward ($0 \to 1$) | GT | 10.797 | $0.839^{\pm 0.006}$ | $0.924^{\pm 0.004}$ | $0.956^{\pm 0.003}$ | $22.856^{\pm 0.160}$ |
| Two-forward ($1 \to 0$) | GT | 8.001 | $0.926^{\pm 0.004}$ | $0.972^{\pm 0.002}$ | $0.983^{\pm 0.001}$ | $23.106^{\pm 0.136}$ |
| Two-forward ($p = 1$) | GT | 8.033 | $0.934^{\pm 0.004}$ | $0.981^{\pm 0.002}$ | $0.992^{\pm 0.001}$ | $23.223^{\pm 0.172}$ |
| Rollout Only | GT | 14.854 | $0.860^{\pm 0.005}$ | $0.937^{\pm 0.004}$ | $0.966^{\pm 0.003}$ | $22.550^{\pm 0.161}$ |
| Rollout Only with History | GT | 8.585 | $0.915^{\pm 0.004}$ | $0.968^{\pm 0.003}$ | $0.984^{\pm 0.003}$ | $22.883^{\pm 0.142}$ |
| Rehearsal (Ours) | GT | **6.505** | $\mathbf{0.957}^{\pm \mathbf{0.004}}$ | $\mathbf{0.991}^{\pm \mathbf{0.002}}$ | $\mathbf{0.998}^{\pm \mathbf{0.001}}$ | $\mathbf{23.241}^{\pm \mathbf{0.158}}$ |
| Two-forward ($0 \to 1$) | Online | 15.367 | $0.796^{\pm 0.005}$ | $0.887^{\pm 0.004}$ | $0.923^{\pm 0.003}$ | $22.974^{\pm 0.147}$ |
| Two-forward ($1 \to 0$) | Online | 12.695 | $0.846^{\pm 0.005}$ | $0.933^{\pm 0.003}$ | $0.963^{\pm 0.002}$ | $23.239^{\pm 0.151}$ |
| Two-forward ($p = 1$) | Online | 11.766 | $0.887^{\pm 0.004}$ | $0.958^{\pm 0.003}$ | $\mathbf{0.981}^{\pm \mathbf{0.002}}$ | $23.195^{\pm 0.166}$ |
| Rollout Only | Online | 17.036 | $0.792^{\pm 0.006}$ | $0.903^{\pm 0.005}$ | $0.945^{\pm 0.003}$ | $22.660^{\pm 0.151}$ |
| Rollout Only with History | Online | 12.320 | $0.891^{\pm 0.005}$ | $0.961^{\pm 0.002}$ | $0.980^{\pm 0.002}$ | $22.684^{\pm 0.110}$ |
| Rehearsal (Ours) | Online | **8.714** | $\mathbf{0.909}^{\pm \mathbf{0.005}}$ | $\mathbf{0.966}^{\pm \mathbf{0.002}}$ | $\mathbf{0.981}^{\pm \mathbf{0.001}}$ | $\mathbf{23.353}^{\pm \mathbf{0.189}}$ |

Table 4: **Cross-condition evaluation.**

| Train SSL | Eval. SSL | FID↓ | R@1↑ | R@2↑ | R@3↑ | Div.→ |
|---|---|---|---|---|---|---|
| GT motion† | | 0.001 | 1.000 | 1.000 | 1.000 | 23.616 |
| GT | GT | 6.505 | 0.957 | 0.991 | 0.998 | 23.241 |
| GT | Online | 25.872 | 0.714 | 0.827 | 0.868 | 22.067 |
| Online | Online | 8.714 | 0.909 | 0.966 | 0.981 | 23.353 |

Table 5: **Conditioning representations.**

| Configuration | | FID↓ | R@1↑ | R@2↑ | R@3↑ | Div.→ |
|---|---|---|---|---|---|---|
| GT motion† | | 0.001 | 1.000 | 1.000 | 1.000 | 23.616 |
| Spatial | Fused | 11.445 | 0.881 | 0.951 | 0.972 | 23.083 |
| | Separate | **8.714** | **0.909** | **0.966** | **0.981** | **23.353** |
| Acoustic | MOSPA | 12.759 | 0.818 | 0.907 | 0.940 | 22.694 |
| | Ours | **8.714** | **0.909** | **0.966** | **0.981** | **23.353** |

### 4.3 Effect of spatial feedback

Adapting to generated motion histories alone is insufficient for Online generation. Table 4 compares two rehearsal-trained models. For the GT-trained model, switching evaluation from GT to Online SSL with fixed generation settings substantially degrades motion fidelity and conditioning alignment. During training, it encounters prediction errors in its motion history but not the resulting changes in spatial conditions. Training with spatial feedback substantially improves fidelity and alignment under the same Online protocol. The comparison supports updating body-relative SSL during rehearsal, so that motion histories and spatial conditions reflect the same generated trajectory.

### 4.4 Acoustic and Spatial Representation

**Separate spatial tokens.** All representation ablations use Online SSL. *Fused* linearly projects concatenated acoustic and spatial embeddings into a single conditioning token. Separate tokens improve fidelity and alignment (Table 5). Preserving distinct inputs allows the transformer to attend to acoustic and spatial cues separately.

**Learning reaction-relevant acoustic features.** MOSPA aggregates diverse audio descriptors into a 2,272-dimensional feature vector. Our learned log-mel features achieve better fidelity and conditioning alignment (Table 5). This improvement is consistent with our hypothesis that redundant or task-irrelevant acoustic variation can interfere with condition modeling. Separating what from where helps the model associate acoustic events with spatially appropriate responses.

## 5 Conclusion

We presented Sonic-AMG for real-time autoregressive motion generation from streaming spatial audio, with separate acoustic and spatial representations. Rehearsal training combines full-sequence teacher-forced supervision with short autoregressive self-rollouts, while spatial feedback keeps source coordinates consistent with generated motion. Sonic-AMG achieves state-of-the-art FID and R-precision on SAM and generates coherent reactions across sound events and source locations.

### AI USE STATEMENT

Generative AI tools were used to polish the manuscript and assist with code development. The authors take full responsibility for the final manuscript and code.

### REPRODUCIBILITY STATEMENT

Section 3 describes the model and training objectives, while Section 4 specifies the evaluation protocols and metrics. Appendices E and F provide implementation and training details. Runtime measurement procedures are described in Appendix C. Our code and model will be released upon acceptance.

## REFERENCES

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics*, 42(4): 44:1–44:20, 2023. doi: 10.1145/3592458. URL https://doi.org/10.1145/3592458.

Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *ACM Transactions on Graphics*, 42(4):42:1–42:18, 2023. doi: 10.1145/3592097. URL https://arxiv.org/abs/2303.14613.

João Pedro Araújo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026. URL https://pei-xu.github.io/GMR.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yiyi Cai, Yuhan Wu, Kunhang Li, You Zhou, Bo Zheng, and Haiyang Liu. FloodDiffusion: Tailored diffusion forcing for streaming motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2295–2304, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Cai_FloodDiffusion_Tailored_Diffusion_Forcing_for_Streaming_Motion_Generation_CVPR_2026_paper.pdf.

Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-React: Online reaction policy for two-character interaction generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mm0cqJ2O3f.

Bohong Chen, Yumeng Li, Yinglin Xu, Youyi Zheng, Yanlin Weng, and Kun Zhou. EchoAvatar: Real-time generative avatar animation from audio streams. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '26. Association for Computing Machinery, 2026. doi: 10.1145/3799902.3811066. URL https://doi.org/10.1145/3799902.3811066.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-0759. URL https://arxiv.org/abs/2407.01392.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *Proceedings of the European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/1912.11474.

Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15516–15525, 2021. URL https://arxiv.org/abs/2012.11583.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7352–7361, June 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_DiffSHEG_A_Diffusion-Based_Approach_for_Real-Time_Speech-driven_Holistic_3D_Expression_CVPR_2024_paper.pdf.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

Michael S. A. Graziano, Lina A. J. Reiss, and Charles G. Gross. A neuronal representation of the location of nearby sounds. *Nature*, 397(6718):428–430, 1999. doi: 10.1038/17115. URL https://doi.org/10.1038/17115.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.08009.

Kaiyang Ji, Bingsheng Qian, Binghuan Wu, Kangyi Chen, Ye Shi, and Jingya Wang. DiscoForcing: A unified framework for real-time audio-driven character control with diffusion forcing. In *Proceedings of the Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.28491.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.14795.

Daniel Aleksander Krause, Guillermo García-Barrios, Archontis Politis, and Annamaria Mesaros. Binaural sound source distance estimation and localization for a moving listener. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:996–1011, 2024a. doi: 10.1109/TASLP.2023.3346297. URL https://doi.org/10.1109/TASLP.2023.3346297.

Daniel Aleksander Krause, Archontis Politis, and Annamaria Mesaros. Sound event detection and localization with distance estimation. In *32nd European Signal Processing Conference (EUSIPCO)*, pp. 286–290, 2024b. doi: 10.23919/EUSIPCO63174.2024.10715220. URL https://doi.org/10.23919/EUSIPCO63174.2024.10715220.

Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1524–1534. IEEE, 2024a.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412, 2021.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://arxiv.org/abs/2406.11838.

Zekun Li, Sizhe An, Chengcheng Tang, Chuan Guo, Ivan Shugurov, Linguang Zhang, Amy Zhao, Srinath Sridhar, Lingling Tao, and Abhay Mittal. LLaMo: Scaling pretrained language models for unified motion understanding and generation with continuous autoregressive tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2209–2220, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Li_LLaMo_Scaling_Pretrained_Language_Models_for_Unified_Motion_Understanding_and_CVPR_2026_paper.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1144–1154, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EMAGE_Towards_Unified_Holistic_Co-Speech_Gesture_Generation_via_Expressive_Masked_CVPR_2024_paper.html.

Pinxin Liu, Haiyang Liu, Jiahao Luo, Junhua Huang, Chunhao Zou, and Luchuan Song. GestureFAR: Streaming co-speech gesture generation with flow autoregression, 2026. URL https://arxiv.org/abs/2609.21576.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. doi: 10.1126/scirobotics.aed4592. URL https://arxiv.org/abs/2511.07820.

Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. POPDG: Popular 3D dance generation with PopDanceSet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26984–26993, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Luo_POPDG_Popular_3D_Dance_Generation_with_PopDanceSet_CVPR_2024_paper.html.

Tsvetomila Mihaylova and André F. T. Martins. Scheduled sampling for transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 351–356, 2019. doi: 10.18653/v1/P19-2049. URL https://aclanthology.org/P19-2049/.

John G. Neuhoff. Perceptual bias for rising tones. *Nature*, 395(6698):123–124, 1998. doi: 10.1038/25862. URL https://doi.org/10.1038/25862.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. FlowAR: Scale-wise autoregressive image generation meets flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 51489–51502, 2025. URL https://proceedings.mlr.press/v267/ren25e.html.

Muhammad Usama Saleem, Mayur Jagdishbhai Patel, Ekkasit Pinyoanuntapong, Zhongxing Qin, Li Yang, Hongfei Xue, Ahmed Helmy, Chen Chen, and Pu Wang. LiveGesture: Streamable co-speech gesture generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2264–2273, June 2026. URL

https://openaccess.thecvf.com/content/CVPR2026/papers/Saleem_LiveGesture_Streamable_Co-Speech_Gesture_Generation_Model_CVPR_2026_paper.pdf.

Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models. *ACM Transactions on Graphics*, 43(4):143:1–143:14, 2024. doi: 10.1145/3658140. URL https://xbpeng.github.io/projects/AMDM/AMDM_2024.pdf.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Siyao_Bailando_3D_Dance_Generation_by_Actor-Critic_GPT_With_Choreographic_Memory_CVPR_2022_paper.html.

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Yaoyao Chang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. In *Proceedings of the Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=PnTXyTR2VG.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H. Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7319b7561ffe5e2f6419acd4a2f52d6b-Abstract-Conference.html.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. EDGE: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–458, 2023. URL https://arxiv.org/abs/2211.10658.

Evgeniia Vu, Andrei Boiarov, and Dmitry Vetrov. Streaming generation of co-speech gestures via accelerated rolling diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26054–26061, 2026. doi: 10.1609/aaai.v40i31.39807. URL https://ojs.aaai.org/index.php/AAAI/article/view/39807.

Jason D. Warren, Brandon A. Zielinski, Gary G. R. Green, Josef P. Rauschecker, and Timothy D. Griffiths. Perception of sound-source motion by the human brain. *Neuron*, 34(1):139–148, 2002. doi: 10.1016/S0896-6273(02)00637-2. URL https://doi.org/10.1016/S0896-6273(02)00637-2.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. MotionStreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10086–10096, 2025. URL https://arxiv.org/abs/2503.15451.

Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, and Taku Komura. MOSPA: Human motion generation driven by spatial audio. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.11949.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6):222:1–222:16, 2020. doi: 10.1145/3414685.3417838. URL https://arxiv.org/abs/2009.02119.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14730–14740, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Generating_Human_Motion_From_Textual_Descriptions_With_Discrete_Representations_CVPR_2023_paper.html.

Yan Zhang, Yao Feng, Alpár Cseke, Nitin Saini, Nathan Bajandas, Nicolas Heron, and Michael J. Black. PRIMAL: Physically reactive and interactive motor model for avatar learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12725–12736, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Zhang_PRIMAL_Physically_Reactive_and_Interactive_Motor_Model_for_Avatar_Learning_ICCV_2025_paper.pdf.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05260.

Kaifeng Zhao, Mathis Petrovich, Haotian Zhang, Tingwu Wang, Siyu Tang, and Davis Rempe. Autoregressive diffusion with hybrid representation for interactive human motion generation. *ACM Transactions on Graphics*, 45(4):86:1–86:14, 2026. doi: 10.1145/3811284. URL https://doi.org/10.1145/3811284.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10544–10553, 2023. URL https://arxiv.org/abs/2303.09119.

# A Limitations and Future Work

Our main experiments assume known world-space source positions during inference. Our No-SSL variant (Appendix D) provides an initial demonstration that spatial conditioning can be derived from binaural audio. A systematic evaluation of alternative spatial inference methods and their effects on motion generation remains future work. Our evaluation is limited to SAM, so generalization to broader acoustic environments and human reactions remains to be assessed. Evidence for long-horizon coherence and event transitions is primarily qualitative. Future work should establish quantitative benchmarks for evaluating motion quality over longer audio streams with more diverse event transitions.

# B Additional Visual Results and Physics-Based Simulation

The supplementary material includes HTML and videos of four comparisons with MOSPA, ten single-sound examples, and a continuous sequence spanning three sound events. To view the videos in the accompanying HTML page, open `index.html` after extracting the supplementary package. The single-sound and continuous examples pair human motion visualizations with humanoid physics simulations, synchronized with stereo audio.

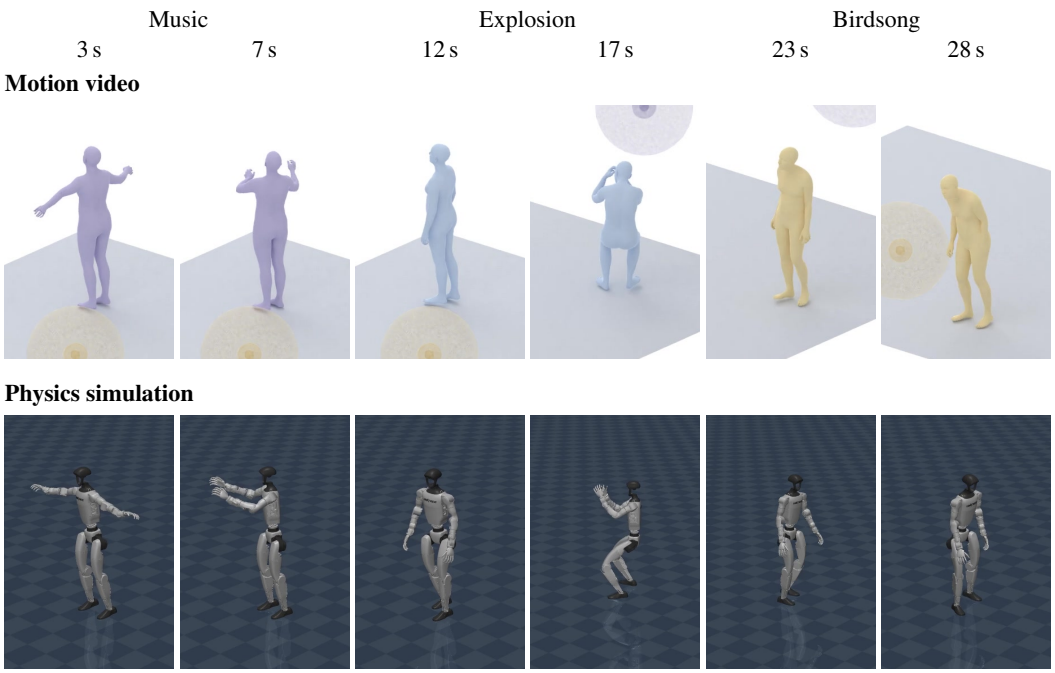
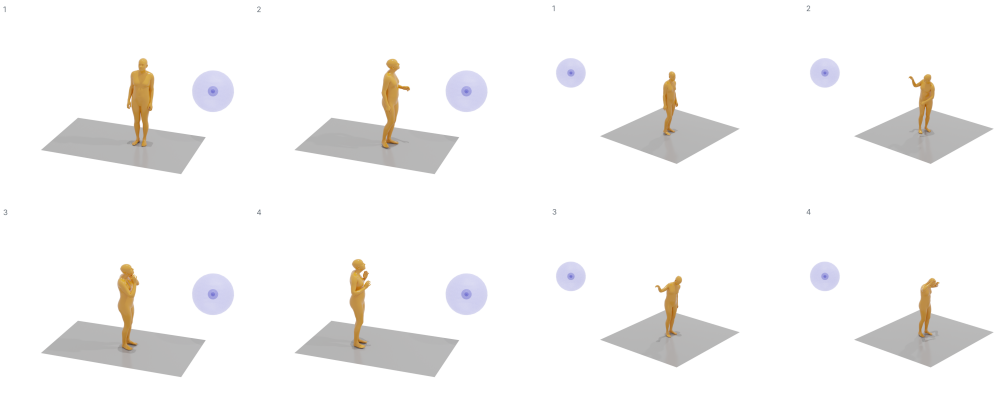



Figure 6: **Human reactions and humanoid tracking in physics simulation.** Aligned snapshots of the continuous Music → Explosion → Birdsong sequence show human motion (top) and G1 simulation (bottom). Robot tracking is described in Appendix B.3.

## B.1 Reactions to Different Sounds

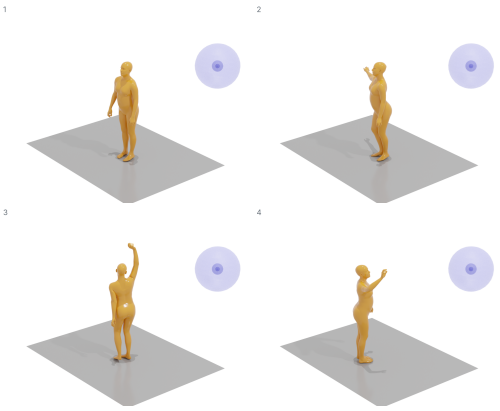
Across diverse sounds, Sonic-AMG combines recognizable gestures with changes in body orientation and position. Figures 7 and 8 show all ten single-sound examples from Part 3 of the supplementary HTML, including hands raised toward the ears, an arm raised in response to a call, and a hand brought toward the mouth. The full videos show how these gestures develop alongside stepping and body turns. All examples use the feedback-trained model with online spatial updates.



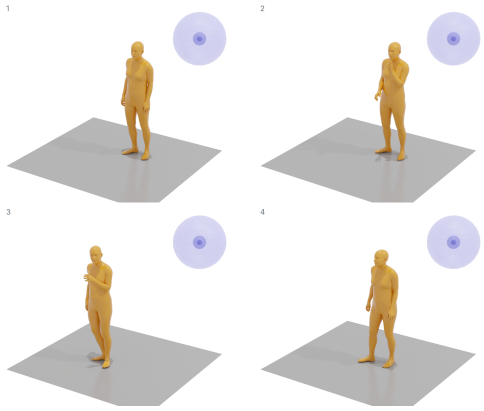
(a) Aircraft: cover ears (0012002)



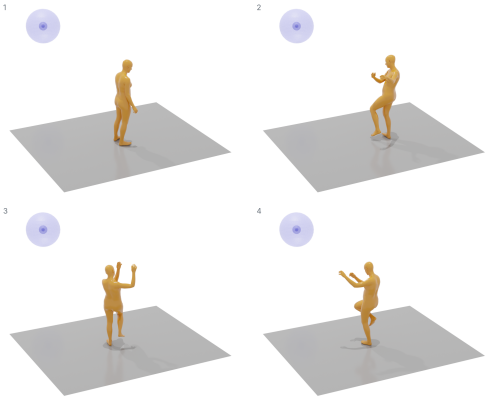
(b) Insect: swatting gestures (0473312)



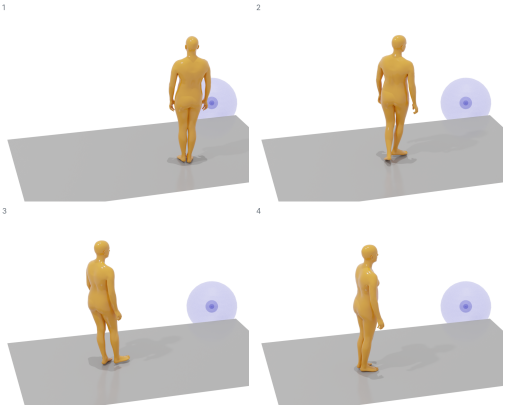
(c) Call: raise an arm (0133102)



(d) Cough: hand toward mouth (0213301)

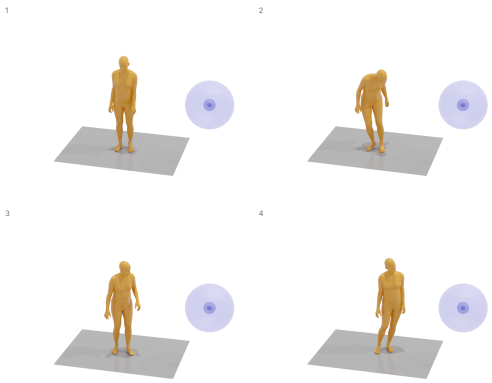


(e) Singing: lift a knee (0632311)

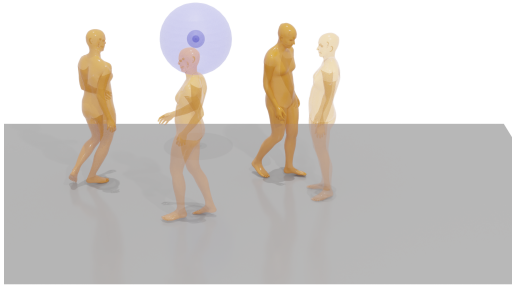


(f) Siren: step and turn (0681200)

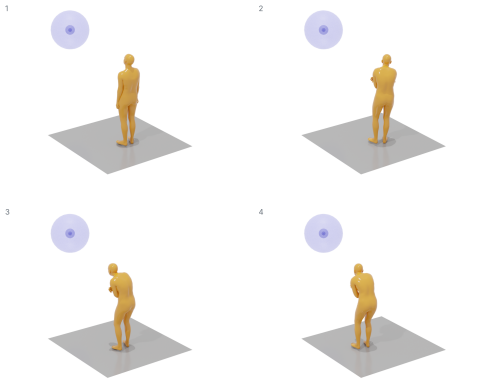
Figure 7: **Additional reactions under online spatial feedback.** Original generated motions for six sounds, with poses ordered chronologically. Purple markers denote sound sources; clip identifiers match the supplementary videos.



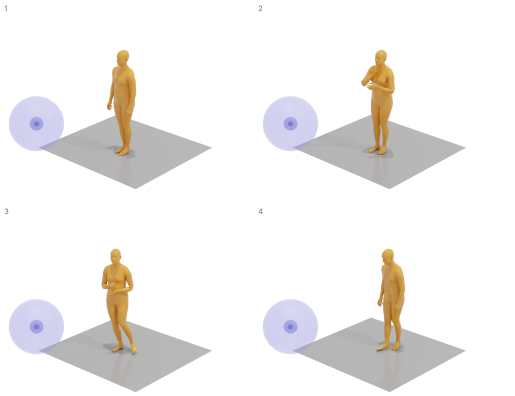
(a) Phone ringing: bend forward (0582000)



(b) Fire alarm: turn and return (0370110)



(c) Wind and rain: draw arms inward (0780302)



(d) Speech: gesture with both hands (0710002)

Figure 8: **Additional reactions under online spatial feedback (continued).**

### B.2 Continuous Reactions across Sound Events

Sonic-AMG sustains coherent reactions as sound content and source location change, without a new ground-truth motion initialization at event transitions. The continuous example spans 30.13 s and combines three SAM audio events—Music → Explosion → Birdsong—at different source locations (Table 6). Generation begins with four GT motion frames; later events continue from the generated body state.

Table 6: **Continuous demonstration setup.** Source locations change at event boundaries. Intervals refer to the generated 30 Hz sequence; style identifiers follow SAM.

| Sound | SAM clip | Style | Interval (s) |
|---|---|---|---|
| Music | 0541212 | 2 | $[0, 10.00)$ |
| Explosion | 0330202 | 2 | $[10.00, 20.27)$ |
| Birdsong | 0113311 | 1 | $[20.27, 30.13)$ |

At each transition, generation resumes from the generated motion history in a root-aligned frame. Spatial feedback updates the active source's body-relative coordinates, while previously emitted frames remain fixed.

### B.3 Execution in a Humanoid Physics Simulation

Generated reactions can be tracked by a pretrained humanoid controller without additional policy training. We retarget the human motions to a Unitree G1 humanoid using General Motion Retar-

geting (GMR) (Araújo et al., 2026) and track the references offline with the released pretrained GEAR-SONIC controller (Luo et al., 2026) in MuJoCo (Todorov et al., 2012). References are interpolated from 30 to 50 Hz, with physics integrated at 1,000 Hz. The simulation uses gravity, ground contact, and PD torques clipped to actuator limits. After initialization, the free root evolves through physics integration.

Figure 6 pairs human reactions with their simulated humanoid execution. The robot tracks the references through all three sound events in a single uninterrupted simulation.

## C Streaming Efficiency and Latency

Each autoregressive step generates four motion frames at 30 Hz, giving a computation budget of $4/30$ s, or 133.3 ms, per block. Table 7 reports steady-state computation on RTX A6000 and TITAN RTX GPUs, measured for our model and for MOSPA under the same conditions. With online spatial feedback, our P95 step times are 53.2 and 125.6 ms, both below the block duration. MOSPA achieves an RTF below one on the RTX A6000 but requires the complete audio window before generation, precluding causal streaming.

Table 7: **Streaming computation.** For our model, times are in milliseconds over 680 steps after warm-up, starting from precomputed acoustic tokens; Online includes motion decoding and spatial updates, and Offline measures latent generation with precomputed SSL. MOSPA emits a 240-frame block per sampler call and has no per-step quantity, so its per-step columns are blank; its RTF comes from five timed blocks of its released 1,000-step sampler on the same GPUs. RTF is computation time divided by the duration of motion produced, and is comparable across both methods. Real-time denotes support for causal streaming with an RTF below 1.

| Method | GPU | SSL type | Mean↓ | Median↓ | P95↓ | RTF↓ | Real-time |
|---|---|---|---|---|---|---|---|
| MOSPA | RTX A6000 | Offline | — | — | — | 0.357 | × |
| MOSPA | TITAN RTX | Offline | — | — | — | 1.427 | × |
| Ours | RTX A6000 | Offline | 41.2 | 41.1 | 42.5 | 0.309 | ✓ |
| Ours | RTX A6000 | Online | 50.2 | 50.9 | 53.2 | 0.376 | ✓ |
| Ours | TITAN RTX | Offline | 102.5 | 105.8 | 108.1 | 0.769 | ✓ |
| Ours | TITAN RTX | Online | 118.0 | 117.9 | 125.6 | 0.885 | ✓ |

**Timing protocol.** We measure one 700-step run per GPU and spatial protocol, discarding the first 20 steps as warm-up. Generation starts from a zero motion latent. The run measures computation over 93.3 s of additional motion. The KV cache retains the full history within a configured context of 4,096 token positions. We use FP32 tensors with TF32 matrix multiplication enabled and synchronize the GPU around each wall-clock measurement. The Offline and Online runs use the GT-SSL and feedback-trained rehearsal models, respectively.

**Measured operations.** At batch size one, each step includes both guidance branches, ten Euler sampling steps, and the persistent KV-cache update. Offline spatial tokens are precomputed. Online timing additionally includes decoding the generated latent prefix and computing spatial tokens from its last two poses; quality evaluation uses four-frame pooled coordinates (Section 3.4). We time the motion-generation stage from precomputed acoustic tokens, excluding audio preprocessing and encoding, initialization, rendering, and output delivery. The acoustic frontend and feature-availability delay are specified in Appendix E.

**Results.** Online feedback increases mean step time from 41.2 to 50.2 ms on RTX A6000 and from 102.5 to 118.0 ms on TITAN RTX. The corresponding online real-time factors are 0.376 and 0.885, with $\mathrm{RTF} = \bar{t}_{\mathrm{step}}/(4/30\,\mathrm{s})$, where $\bar{t}_{\mathrm{step}}$ denotes mean per-step computation time.

**MOSPA timing protocol.** MOSPA jointly denoises a complete 240-frame sequence using the full audio window (Xu et al., 2025), so its generation protocol does not support causal blockwise streaming. We time its released sampler on the same two GPUs: batch size one, the published 1,000-step denoising schedule, and five timed blocks after one warm-up block, with the GPU synchronized around each measurement. Because a block is emitted in a single call, MOSPA has no per-step time to report, and the per-step columns of Table 7 are left blank. The real-time factor is defined for both methods as computation time divided by the duration of motion produced, so that column is directly comparable.

**Cross-method comparison.** One 240-frame block spans 8.0 s of motion and takes 2,857 ms on the RTX A6000 and 11,415 ms on the TITAN RTX, giving real-time factors of 0.357 and 1.427. Our online configuration reaches 0.376 and 0.885 on the same hardware. The two methods are therefore

comparable in throughput on the RTX A6000, while on the TITAN RTX MOSPA no longer produces motion faster than it plays. RTF measures computational throughput. Causal streaming additionally requires generation from incoming audio without waiting for a complete input window.

The gap widens with weaker hardware. Moving from the RTX A6000 to the TITAN RTX multiplies MOSPA's block time by 3.99 and our online step time by 2.35. Per second of generated motion, MOSPA evaluates its network 125 times over the full 240-token sequence, whereas our model takes 7.5 autoregressive steps, each extending the KV cache by three tokens and invoking the rectified-flow head for ten Euler steps.

Initial response latency also depends on input availability. With the initial motion segment supplied, the acoustic token for the first new block becomes available 142.7 ms after audio onset (Appendix E). Adding the measured mean online generation time of 50.2 ms gives an estimated response latency of 192.9 ms on the RTX A6000. For MOSPA's 8.0 s input window, the corresponding estimate is 10,857 ms, including 2,857 ms of generation. Sonic-AMG thus reduces estimated initial response latency by a factor of approximately 56. Both estimates combine input waiting with mean generation time, using steady-state computation measurements. Acoustic preprocessing and encoding, initialization, rendering, and output delivery are excluded.

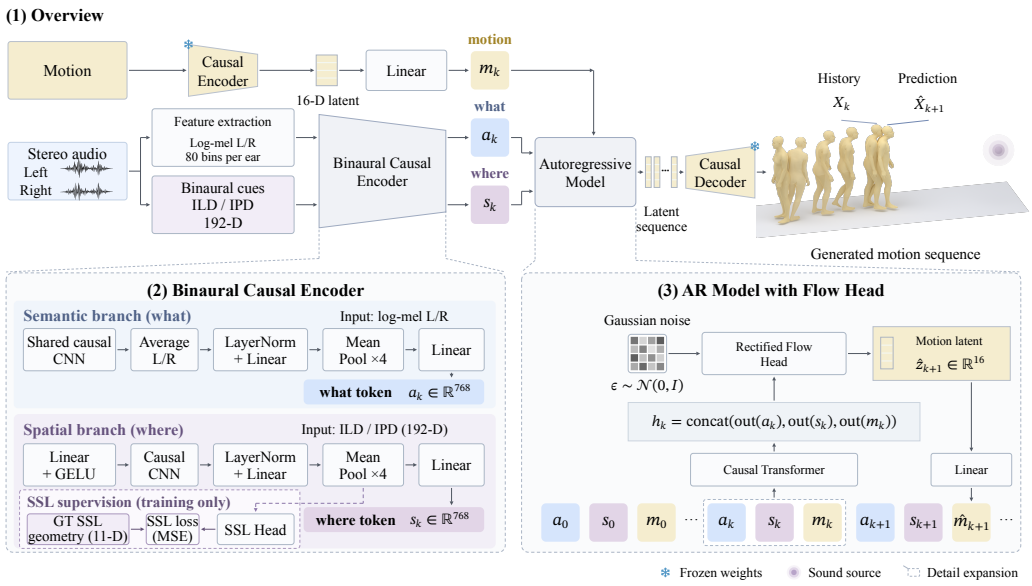



Figure 9: **Sonic-AMG with audio-derived spatial conditioning.** The semantic and spatial branches derive *what* and *where* tokens from log-mel and ILD/IPD features, respectively. These tokens use the same interface to the autoregressive motion generator. The training-only SSL head predicts the normalized 11-dimensional geometric descriptor from pooled spatial features, supervised by ground-truth geometry through an MSE loss. Inference uses the learned spatial features directly, without the SSL head or spatial feedback. Motion poses are schematic.

## D Generation without SSL Inputs

**Motivation and setting.** Sonic-AMG separates the source of spatial conditions from the autoregressive motion generator. In our SAM setting, source-location annotations are available alongside binaural audio (Xu et al., 2025) and provide the geometry used by the main model. At deployment, the source position can instead be supplied by an upstream sound source localization module, after transformation into the required coordinate frame, or by scene metadata in a virtual environment. To examine an alternative realization of this interface, we instantiate a variant that derives spatial tokens directly from binaural audio. We refer to this variant as *No-SSL*, denoting the absence of source-location inputs at inference. It retains the initial four-frame motion segment and the reaction-style label used by the main model.

**Audio-derived spatial tokens.** Figure 9 shows the variant's architecture. For each audio frame, we concatenate interaural level differences (ILD) and the sine and cosine of interaural phase differences (IPD) over 64 frequency bands, giving a 192-dimensional input. A linear projection with GELU, a two-layer causal CNN, and a layer-normalized linear projection encode these cues. Mean pooling over four frames produces a spatial feature $\mathbf{f}_k^s \in \mathbb{R}^{128}$ at the motion-latent rate. The spatial token is $\mathbf{s}_k = W_s \mathbf{f}_k^s + \mathbf{e}_c$, where $W_s$ projects to the transformer width of 768 and $\mathbf{e}_c$ is the reaction-style embedding. The semantic branch encodes log-mel features with a shared per-ear causal CNN, averages the left and right streams, and applies projection and pooling to obtain the acoustic token. Unlike the main encoder, this variant omits the additional causal CNN after left/right averaging in the semantic branch.

**Motion generation.** The variant preserves the interleaved sequence $[\mathbf{a}_k, \mathbf{s}_k, \mathbf{m}_k]$, the causal transformer, the rectified-flow head, and the frozen causal motion autoencoder. We train the generator jointly with the audio encoders using the teacher-forced and 16-step rehearsal objectives described in Section 3.3. Both conditioning streams are computed from the audio, while sampled motion latents extend the autoregressive history. This variant therefore does not decode root poses to recompute

spatial conditions during rehearsal or inference. Its spatial conditions follow the recorded binaural stream rather than an explicit source position updated relative to the generated body.

**Geometric supervision during training.** No-SSL retains source-location annotations as training targets. An auxiliary head $H_{\mathrm{SSL}}$ predicts the normalized 11-dimensional descriptor of Eq. (1) from $\mathbf{f}_k^s$, before its projection to a spatial token. For valid training positions $\mathcal{V}$, the SSL objective is

$$\mathcal{L}_{\mathrm{SSL}} = \frac{1}{11|\mathcal{V}|} \sum_{k \in \mathcal{V}} \left\| H_{\mathrm{SSL}}(\mathbf{f}_k^s) - \frac{\boldsymbol{\phi}_k^{\mathrm{GT}} - \boldsymbol{\mu}_\phi}{\boldsymbol{\sigma}_\phi} \right\|_2^2, \tag{9}$$

where normalization statistics are computed from the training set. The semantic features also support audio energy, onset, and activity regression with loss $\mathcal{L}_{\mathrm{event}}$, alongside audio-class classification with loss $\mathcal{L}_{\mathrm{class}}$. The complete training objective is

$$\mathcal{L}_{\text{No-SSL}} = \tfrac{1}{2}(\mathcal{L}_{\mathrm{TF}} + \mathcal{L}_{\mathrm{reh}}) + 0.10\mathcal{L}_{\mathrm{SSL}} + 0.05\mathcal{L}_{\mathrm{event}} + 0.05\mathcal{L}_{\mathrm{class}}. \tag{10}$$

The auxiliary heads are used only for training. The predicted geometric descriptor is not passed to the motion generator: at inference, the learned spatial feature directly supplies the *where* token.

**Scope of the experiment.** This configuration instantiates the same motion-generation interface with an audio-derived provider of spatial conditions. It removes the need for source-position inputs at inference while retaining geometric supervision during training.

**Quantitative comparison.** No-SSL must infer spatial cues from incoming binaural audio while generating motion causally, without source-location inputs or access to future audio. Despite these demands, it surpasses EDGE in motion fidelity and achieves conditioning alignment and diversity comparable to MOSPA (Table 8). These results show that audio-derived spatial tokens provide effective conditioning for autoregressive reactions. Sonic-AMG can therefore support streaming motion generation in scenarios where explicit source geometry is unavailable.

Table 8: **No-SSL comparison on SAM.** Baselines are reused from Table 2. † denotes results from Xu et al. (2025).

| Method | FID↓ | R@1↑ | R@2↑ | R@3↑ | Diversity→ |
|---|---|---|---|---|---|
| GT motion† | 0.001 | 1.000 | 1.000 | 1.000 | 23.616 |
| EDGE† | 13.993 | 0.886 | 0.960 | 0.977 | 23.099 |
| POPDG† | 20.967 | 0.762 | 0.886 | 0.934 | 22.536 |
| LODGE† | 102.289 | 0.444 | 0.594 | 0.679 | 21.101 |
| Bailando† | 168.396 | 0.077 | 0.134 | 0.182 | 17.347 |
| MOSPA | **8.242** | **0.942** | **0.984** | **0.994** | **23.412** |
| Sonic-AMG (No-SSL) | <u>11.917</u> | <u>0.915</u> | <u>0.973</u> | <u>0.987</u> | <u>23.371</u> |

# E Implementation Details

**Dataset.** SAM (Xu et al., 2025) contains binaural recordings paired with human reactions across 27 sound scenarios and three reaction styles.

**Architecture and optimization.** The generator uses a 12-layer, 12-head transformer of width $d = 768$, with rotary positional embeddings and query–key normalization. Motion latents have dimension $d_z = 16$. The conditional flow head contains nine residual MLP blocks of width 1792 with adaptive layer normalization. We optimize the generator and conditioning modules with AdamW, using batch size 64, learning rate $10^{-5}$, weight decay $10^{-6}$, and momentum coefficients $(0.9, 0.95)$. Each supervised transition uses four independently sampled noise–time pairs. Padded prediction positions are excluded from all losses.

**Training hardware and time.** We train the online rehearsal model from scratch for 150,000 updates on a single NVIDIA L40S GPU, taking approximately 63 hours. The 130,000-update checkpoint used for the reported online results is reached after approximately 54 hours. These wall-clock times include periodic validation and checkpointing, but exclude motion-autoencoder pretraining.

**Causal motion autoencoder.** The temporal autoencoder is pretrained using a Gaussian reconstruction objective, KL regularization, and an additional root-position reconstruction term. The encoder and decoder are then fixed during generator training. Motion channels are standardized using training-set statistics; decoded channels are converted back to physical coordinates before source geometry is computed. The initial GT latent encodes four motion frames.

**Acoustic encoding and feature availability.** Audio is sampled at 48 kHz. Each channel is converted to 80-dimensional log-mel features using a 2048-sample Hann window, hop 1600, and non-centered windows. The shared per-ear causal convolutional stack has 64 channels and dilation rates $(1, 2, 4)$; the subsequent two-layer causal CNN has 128 channels and dilations $(1, 2)$. Both use kernel size three and left padding. A layer-normalized projection, four-frame mean pooling, and a final linear projection produce acoustic tokens of width 768.

The final analysis window contributing to block $k$ ends at sample $(4k + 3)1600 + 2048$. Thus its audio token becomes available at

$$b_k = \frac{(4k + 3)1600 + 2048}{48000}, \tag{11}$$

which is 9.33 ms beyond the nominal four-frame boundary $4(k + 1)/30$. This specifies the feature delay required by the factorization in Eq. (2). Appendix C reports the subsequent motion-generation computation, starting from precomputed acoustic tokens.

**Spatial encoding.** Spatial descriptors are normalized channel-wise using training-set statistics. The spatial MLP has widths $11 \to 768 \to 768$ with SiLU activation. Direction and distance increments are measured per latent step, rather than divided by elapsed time. Both increments are zero for the first block. Distance denominators are clamped from below at $10^{-8}$ for direction normalization and $10^{-6}$ for inverse distance and normalized approach. For SAM, we recover world-space source coordinates as the coordinatewise temporal median of $\mathbf{R}_t^{\mathrm{GT}} \boldsymbol{\ell}_t^{\mathrm{GT}} + \mathbf{p}_t^{\mathrm{GT}}$. At evaluation, the corresponding egocentric coordinates are updated from the generated root.

**Guidance and sampling.** For flow matching, we sample $\tau = \mathrm{sigmoid}(\xi)$ with $\xi \sim \mathcal{N}(0, 1)$. During training, the current acoustic and spatial conditions are jointly dropped with probability 0.1 by supplying zero-normalized inputs; the reaction-style embedding is retained. At inference, the conditional and null branches share the motion history and cached prefix, but differ in the current acoustic and spatial tokens. The guided velocity is

$$v_\omega = v_{\mathrm{null}} + \omega(v_{\mathrm{cond}} - v_{\mathrm{null}}), \tag{12}$$

where $\omega$ is the guidance scale; $v_{\mathrm{cond}}$ and $v_{\mathrm{null}}$ denote the respective branch predictions. Only the conditional branch updates the persistent key–value cache. This guidance rule preserves past conditional states in both branches. Evaluation uses ten uniform Euler steps per generated latent. Training rollouts also use ten steps and unguided sampling.

# F TRAINING SEQUENCE CONSTRUCTION

**Two-forward training.** Let $\widehat{\mathbf{z}}_i$ denote a detached first-pass denoising estimate. For full replacement, the two sequence passes are

$$[\mathbf{z}_0^{\mathrm{GT}}, \mathbf{z}_1^{\mathrm{GT}}, \mathbf{z}_2^{\mathrm{GT}}, \ldots] \xrightarrow{\text{first pass}} [\widehat{\mathbf{z}}_1, \widehat{\mathbf{z}}_2, \widehat{\mathbf{z}}_3, \ldots],$$
$$[\mathbf{z}_0^{\mathrm{GT}}, \widehat{\mathbf{z}}_1, \widehat{\mathbf{z}}_2, \ldots] \xrightarrow{\text{second pass}} \text{next-latent losses against GT.}$$

Audio and spatial tokens are omitted for clarity. The first motion latent is retained, while first-pass predictions replace later inputs at corresponding positions. With updated SSL, the replaced sequence is decoded before the second pass to construct its spatial conditions. The second pass is causally masked, but its outputs do not become later inputs within that pass.

For normalized training progress $u \in [0, 1]$, replacement schedules are

$$p_{0\to1}(u) = \tfrac{1}{2}(1 - \cos \pi u), \quad p_{1\to0}(u) = \tfrac{1}{2}(1 + \cos \pi u), \quad p_{\mathrm{constant}}(u) = 1. \tag{13}$$

A subset of input positions is randomly selected according to the replacement fraction. These schedules alter motion-input replacement rather than switching between spatial-conditioning protocols.

**Rehearsal initialization and recursion.** For each sequence with $K$ valid latents, the start $s$ is sampled uniformly from $\{1, \ldots, K - H - 1\}$, with $H = 16$. The rollout seed is

$$\overline{\mathbf{z}}_s = \mathrm{sg}\left[\mathbf{u}_\tau^{(s)} + (1 - \tau) v_\theta(\mathbf{u}_\tau^{(s)}, \tau; \mathbf{h}_{s-1}^{\mathrm{GT}})\right], \tag{14}$$

where $\mathbf{u}_\tau^{(s)}$ is an interpolation between noise and the GT target $\mathbf{z}_s^{\mathrm{GT}}$, and $\mathrm{sg}$ denotes stop-gradient. Subsequent latents are sampled from fresh Gaussian noise and appended recursively, as in Eq. (6). The decoder receives the pre-window GT prefix followed by these generated latents. With online spatial updates, the decoded root determines the next spatial descriptor after each append; the GT-SSL variant retains recorded conditions. Each transition is supervised with the next GT latent, independently of the sampled latent used to extend the history.

The acoustic features are cropped at frame $rs$ and re-encoded. Acoustic convolutional context and the transformer's key–value cache are reset at this boundary, while the decoder retains its GT prefix. Each rollout contains 16 transitions, producing 64 frames. Gradients are stopped through the sampled latents and decoded geometry, but retained through transformer states and conditioning encoders in the supervised objective.

**Rollout-only ablations.** Both variants optimize only the rollout loss and retain the stop-gradient initialization in Eq. (14); the teacher-forced prediction pass supplies the seed but contributes no training loss. Rollout Only uses the same independently sampled starts and reset acoustic and transformer contexts as rehearsal. Rollout Only with History instead builds a transformer cache from the GT motion and conditioning prefix before $s$, and takes acoustic conditions from the full causal encoding. Gradients from the rollout loss propagate through this prefix cache. Both variants use $H = 16$ and ten sampling steps per latent.