EchoAvatar: Real-time Generative Avatar Animation from Audio Streams

BOHONG CHEN, State Key Lab of CAD&CG, Zhejiang University, China
YUMENG LI, State Key Lab of CAD&CG, Zhejiang University, China
YINGLIN XU, State Key Lab of CAD&CG, Zhejiang University, China
YOUYI ZHENG, State Key Lab of CAD&CG, Zhejiang University, China
YANLIN WENG, State Key Lab of CAD&CG, Zhejiang University, China
KUN ZHOU*, State Key Lab of CAD&CG, Zhejiang University, China

Fig. 1. Given streaming audio input, our method generates avatar animation in a streaming manner. The four poses shown above are sampled from a continuous motion sequence driven by the audio stream.

Real-time synthesis of high-fidelity 3D character motion from audio is a pivotal component for next-generation interactive avatars and virtual assistants. However, most existing approaches are limited to offline processing of complete audio sequences or are constrained to specific domains, rarely handling both speech and music effectively. In this paper, we introduce a novel framework designed to generate continuous, coherent full-body motion from streaming speech and music with low latency. Central to our approach is a unified streaming architecture capable of synthesizing continuous motion from incremental audio inputs. We employ a robust training strategy that enforces strong audio dependency, allowing the model to seamlessly generalize across conversational speech and rhythmic music without requiring explicit domain labels or mode switching. Additionally, we explored Reinforcement Learning to refine the quality of online generation. Furthermore, we bridge reactive animation with intent-driven behavior via a tool-call interface that allows upstream Large Language Models to inject explicit semantic control. By combining this controllability with stream audio-driven synthesis, our framework serves as a plug-and-play solution for transforming voice agents into interactive humanoid avatars. Extensive experiments demonstrate that our method outperforms state-of-the-art real-time baselines in motion quality and synchronization while maintaining the flexibility required for live deployment. Our code, pre-trained models, and videos are available at https://robinwitch.github.io/EchoAvatar-Page.

CCS Concepts: • **Computing methodologies** → **Animation**; **Artificial intelligence**.

Additional Key Words and Phrases: Streaming Motion Generation

ACM Reference Format:
Bohong Chen, Yumeng Li, Yinglin Xu, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2026. EchoAvatar: Real-time Generative Avatar Animation from Audio Streams. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3799902.3811066

*Corresponding author

Authors' Contact Information: Bohong Chen, bohongchen@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Yumeng Li, yumeng.li@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Yinglin Xu, ylxu@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Youyi Zheng, youyizheng@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Yanlin Weng, weng@cad.zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Kun Zhou, kunzhou@acm.org, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

ACM ISBN 979-8-4007-2554-8/2026/07
https://doi.org/10.1145/3799902.3811066

1 Introduction

The rapid evolution of Large Language Models (LLMs) and Voice Agents has enabled fluid, natural dialogue. However, while audio fidelity is near-human, visual embodiment lacks the responsiveness required for genuine interaction, necessitating high-fidelity 3D motion generation directly from streaming audio with low latency.

Existing approaches to audio-driven motion synthesis fall short of meeting this challenge for two primary reasons. First, most state-of-the-art methods [Chen et al. 2025a; Liu et al. 2024b; Zhang et al. 2024a] are designed for offline processing, requiring complete audio sequences as input before generating motion. This architectural constraint introduces unacceptable latency for real-time interactive applications. Second, existing methods are typically domain-specific, handling either speech or music but rarely both. This fragmentation necessitates complex model switching, limiting their applicability to general-purpose voice agents that must handle diverse acoustic inputs uniformly.

In this paper, we present a unified framework for real-time, streaming avatar animation that addresses these limitations. This system takes live audio stream and generates motion in real-time, featuring a causal motion tokenizer for high-quality auto-regressive synthesis and a specially designed training strategy for unified learning. Regarding the tokenizer, while recent works have explored causal architectures via causal convolutions [Jiang et al. 2025; Xiao et al. 2025], we find that pure convolutional approaches often lack expressiveness and suffer from reconstruction artifacts. We instead propose an attention-based causal motion tokenizer with auxiliary kinematic losses, achieving superior generation quality and streaming capability. Regarding the training strategy, we identify that optimization dynamics within a unified motion space often weaken audio conditioning, leading to catastrophic failure in task alignment. We address this via a hierarchical token corruption strategy that enhances audio conditioning, enabling the model to uniformly learn conversational gestures and rhythmic dance without explicit domain labels. Furthermore, experiments reveal a synergistic effect where the integration of diverse motion domains mutually reinforces generation fidelity.

Beyond real-time generation, our system is designed for practical deployment within modern voice agent ecosystems. It operates as a plug-and-play module, accepting audio streams from diverse sources ranging from web browsers to AI conversational platforms. Furthermore, we introduce a tool-call interface that enables upstream systems, such as Large Language Models, to interleave explicit semantic actions with implicit audio-driven motion, bridging the gap between purely reactive audio-driven animation and controllable, intent-driven behavior.

To further align real-time generation with human preferences, we explore Reinforcement Learning (RL) by investigating both reward-model-based strategies using Group Relative Policy Optimization (GRPO) and human-annotation-based strategies using Direct Preference Optimization (DPO). We demonstrate measurable improvements in perceived quality and provide an analysis of applying RL to online auto-regressive motion generation for future research.

Our primary contributions are:

- A unified streaming architecture that leverages attention-based causal tokenization to synthesize continuous, high-fidelity motion from streaming speech and music with low latency.
- A robust training curriculum utilizing Hierarchical Token Corruption to enable synergistic learning across diverse domains, boosting performance on individual tasks, alongside an exploration of RL strategies (GRPO/DPO) to enhance perceived generation quality.
- A deployable plug-and-play system that integrates with voice agents, supporting both implicit audio-driven animation and explicit semantic control via tool calls.

2 Related Work

2.1 Co-speech Gesture Generation

The trajectory of co-speech gesture generation reflects a fundamental paradigm shift from explicit, rule-based heuristics to implicit, data-driven synthesis. Early frameworks [Cassell et al. 1994, 2001; Kopp et al. 2006; Lee and Marsella 2006; Lhommet et al. 2015] relied on rigid production rules and manual linguistic mappings, which offered controllability but lacked kinematic naturalness. The advent of deep learning initially spurred deterministic regression approaches [Habibie et al. 2022; Kucherenko et al. 2020; Liu et al. 2022d; Yoon et al. 2020; Zhou et al. 2022]; however, by modeling the modal average of plausible motions, these methods frequently suffered from "mean-pose convergence", resulting in over-smoothed and under-articulated output. To address the inherent stochasticity and one-to-many ambiguity of the speech-to-gesture mapping, the field has pivoted toward probabilistic generative modeling. This landscape encompasses Normalizing Flows [Alexanderson et al. 2020; Ye et al. 2022] for explicit density estimation, Variational Autoencoders (VAEs) [Ghorbani et al. 2023; Li et al. 2021; Shi et al. 2024a] for continuous latent structuring, and Vector Quantized (VQ) frameworks [Ao et al. 2022; Liu et al. 2022b,c; Lu et al. 2023; Yazdian et al. 2022; Yi et al. 2023] that learn discrete motion codebooks. Diffusion Probabilistic Models based approaches [Alexanderson et al. 2023; Ao et al. 2023; Cheng et al. 2024; Mughal et al. 2025; Yang et al. 2025, 2023b; Zhang et al. 2024b] excelling at modeling complex distributions via iterative denoising and MLLM [Chen et al. 2025c,b; Hou et al. 2025; Liu et al. 2025b] unifying 3D human motion with text and speech in a shared latent space. Among these, ConvoFusion [Mughal et al. 2024] extends diffusion-based generation by enabling gesture emphasis on specific words. Teller [Zhen et al. 2025] proposes a real-time audio-driven portrait talking head system. ACRNN [Zhou et al. 2018] is the first method capable of generating arbitrarily long motions in real time with stability. While many approaches [Chen et al. 2024c; Liu et al. 2025c] claim real-time capability for audio-driven body motion generation, they merely achieve generation speeds faster than playback speed under the assumption of full audio context availability. True streaming scenarios—where audio is incrementally received and motion is progressively generated—remain largely unexplored.

2.2 Multimodal Motion Synthesis

Motion synthesis research has significantly expanded its scope by integrating diverse control signals beyond audio. These range from

semantic text descriptions [Bae et al. 2025; Fan et al. 2025; Lu et al. 2025a; Tevet et al. 2023; Zhang et al. 2022] and spatial trajectory constraints [Wan et al. 2023; Xie et al. 2023; Zheng et al. 2025] to physical interaction states [He et al. 2025; Liu et al. 2025a; Lu et al. 2025b; Ruiz-Ponce et al. 2025] and visual signals [Bekor et al. 2025; Feng et al. 2025]. Regarding stylistic control, motion examples [Aberman et al. 2020; Li et al. 2023a] provide a direct reference for desired behaviors. While earlier methods like ZeroEGGS [Ghorbani et al. 2023] compress these examples into static style vectors, often losing kinematic fidelity, recent approaches like MECo [Chen et al. 2025a] and PersonaBooth [Kim et al. 2025] demonstrate that leveraging discrete token prefixes or personalized identifiers allows for precise, fine-grained control. However, within the specific domain of audio-driven generation, a critical fragmentation persists. While recent works achieve high fidelity in niche tasks, such as instrument-specific performance [Qiu et al. 2025], complex rhythmic alignment [Ghosh et al. 2025; Nguyen et al. 2025], or training from in-the-wild short-form music-dance videos [Zhao and Lu 2024], current systems are typically constrained to exclusive domains, handling either conversational speech or rhythmic dance. This bifurcation necessitates explicit task labels or separate models, highlighting the lack of a framework capable of processing a unified, speech and music stream.

2.3 Reinforcement Learning

Reinforcement Learning (RL) has long served as the standard paradigm for optimizing sequential decision-making [Sutton and Barto 1998], with policy gradient methods [Haarnoja et al. 2018; Sutton et al. 1999; Williams 1992] dominating high-dimensional motion control. However, applying these techniques to motion synthesis has historically been fraught with challenges. Prior approaches relying on offline RL [Kumar et al. 2020; Sun et al. 2023] or Actor-Critic frameworks [Li et al. 2022] frequently struggled with brittle reward engineering [Pinto et al. 2023] and insufficient exploration. Hybrid frameworks such as MotionVAE [Ling et al. 2020] and AMDM [Shi et al. 2024b] couple generative motion priors with RL-trained policy controllers to satisfy task-specific objectives. In the generative era, the focus has shifted toward Reinforcement Learning from Human Feedback (RLHF) [Menick et al. 2022; Ouyang et al. 2022a; Yuan et al. 2023] to better capture perceptual nuance. To circumvent the well-documented instability of PPO-based pipelines, Direct Preference Optimization (DPO) [Rafailov et al. 2023] has emerged as a robust alternative, optimizing policies directly from preference pairs without an explicit reward model—a strategy now proven across textual [Liu et al. 2024a; She et al. 2024] and multimodal domains [Li et al. 2023b; Zhao et al. 2023; Zhou et al. 2024]. Complementing this, Group Relative Policy Optimization (GRPO) [Shao et al. 2024] introduces a mechanism for stable online refinement via group advantage normalization. We systematically explore these alignment strategies to improve perceived motion quality beyond the training distribution, particularly for robust one-shot streaming scenarios.

3 Method

As depicted in Figure 2, our framework is designed for real-time, high-fidelity 3D motion synthesis from streaming audio with low latency. Our system comprises three core components: first, a Causal Attention-based Motion Tokenizer that discretizes continuous motion manifolds into latent tokens without violating temporal causality; second, a repurposed pre-trained LLM generator optimized via a three-stage curriculum to align audio-motion modalities and enable explicit semantic control; and finally, a Reinforcement Learning (RL) Alignment stage that refines the policy to improve the alignment of generated motion with human perceptual standards for zero-retry streaming scenarios.

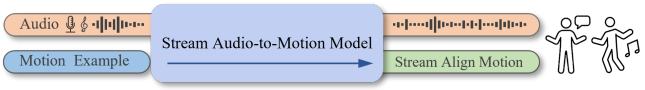

Fig. 2. The structure of our motion generation model. Our model is capable of receiving streaming audio inputs and producing streaming motion outputs. Then, the time-aligned audio and motion are returned to the user together. Furthermore, our model can receive motion examples as additional control signals.

3.1 Motion Tokenizer

We define motion $\mathbf{m}_{1:N}$ as a sequence of pose states parameterized by root velocity, height, and 6D joint rotations [Zhou et al. 2019]. Standard discrete motion tokenizers typically rely on non-causal architectures that necessitate future-frame look-ahead, introducing latency that is prohibitive for real-time interaction. Although recent attempts [Jiang et al. 2025; Xiao et al. 2025] enforce causality via convolutional left-padding. However, due to the limited expressive capacity of convolutional networks [Zhang et al. 2024a], the reconstruction process often suffers from visual artifacts. To resolve this, we propose an Attention-based Causal Motion Tokenizer. We replace rigid convolutional backbones with stacked attention blocks governed by a causal mask, strictly confining the receptive field to the preceding p frames. To handle temporal resampling without information loss, we adopt a dual-path strategy inspired by DC-AE [Chen et al. 2024a]. Downsampling is achieved by aggregating a temporal pooling branch with a feature concatenation branch processed via an MLP, while upsampling reconstructs temporal resolution through temporal replication combined with channel-expansion MLPs. Furthermore, to suppress physical artifacts such as foot sliding, we explicitly integrate Forward Kinematics (FK) into the optimization loop, imposing auxiliary losses on global joint positions, velocities, accelerations, and foot contact consistency. Detailed formulations are provided in the appendix.

The motion sequence $\mathbf{m}_{1:N}$ is encoded into a continuous latent trajectory $\mathbf{z}_{1:n} = \mathcal{E}(\mathbf{m}_{1:N})$, with a temporal downsampling ratio of N/n. To discretize this manifold, we employ Residual Vector Quantization (RVQ) [Guo et al. 2024; Yao et al. 2024; Zeghidour et al. 2022]. The latent vector $\mathbf{z}$ is approximated as a summation of Q quantized residuals, $\hat{\mathbf{z}} = \sum_{q=0}^{Q-1} \hat{\mathbf{z}}^q$, where each component $\hat{\mathbf{z}}^q$ is retrieved from a distinct codebook $\mathbf{C}_q$. This process is recursive: the initial layer quantizes the raw latent, while subsequent layers $q > 0$ refine the quantization error of the partial sum. The final discrete representation $\hat{\mathbf{z}}$ is decoded by $\mathcal{D}$ to reconstruct the motion $\hat{\mathbf{m}}$. The

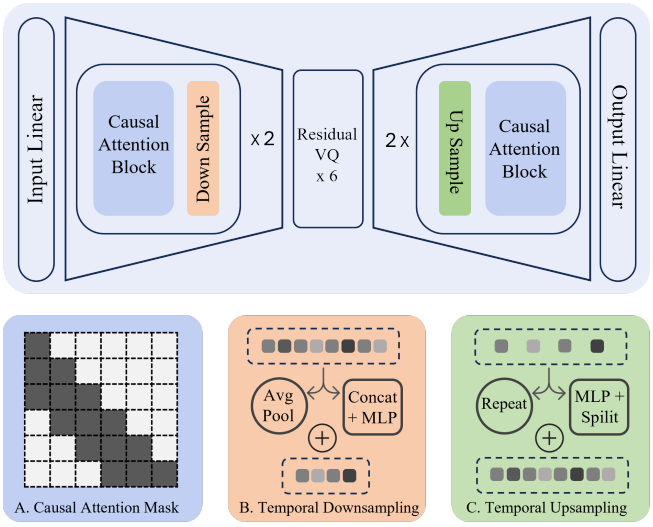

Fig. 3. Architecture of our Attention-based Causal Motion Tokenizer with Residual Vector Quantization. (A) Causal attention mask confines receptive field to preceding frames. (B) Temporal downsampling via dual-path aggregation. (C) Temporal upsampling via dual-path expansion.

entire framework is optimized via a composite objective balancing kinematic reconstruction fidelity $\mathcal{L}_{\text{rec}}$ and codebook commitment:

$$\mathcal{L}_{\text{rec}} = \|\hat{\mathbf{m}}_{1:N} - \mathbf{m}_{1:N}\|_1 + \eta \sum_{q=0}^{Q-1} \|\mathbf{z}^q_{1:n} - \text{sg}[\hat{\mathbf{z}}^q_{1:n}]\|_2^2 + \Phi\Big(\text{FK}(\hat{\mathbf{m}}_{1:N}), \text{FK}(\mathbf{m}_{1:N})\Big), \tag{1}$$

where Φ encapsulates the FK-based auxiliary losses, $\text{sg}[\cdot]$ denotes the stop-gradient operator, and η weights the embedding constraint. To decouple part-specific dynamics, we implement anatomically partitioned tokenization, maintaining separate codebooks for the upper body, lower body, and hands.

Table 1. Quantitative comparison results. MPJPE denotes the Mean Per-Joint Position Error computed in the character-centric coordinate system, measured in units of 10^{-4} m. Similarly, Trans Loss quantifies the average per-frame root translation velocity error, also reported in 10^{-4} m.

Methods	Reconstruction			Generation
	FID ↓	MPJPE ↓	Trans Loss ↓	FID ↓
Real motion	0	0	0	0
CausalConv-RVQ	9.208	525.6	12.94	18.06
Attn	4.183	411.6	12.53	12.25
Attn(w/o dual)	12.21	982.4	48.77	18.55
Attn(w/o auxiliary)	8.775	778.5	55.67	17.68
Attn(w/o lookback)	6.612	468.8	12.89	15.53
Attn(w/ bodypart)	1.306	184.1	9.637	9.465

3.2 Audio Driven Motion Generation

Following the progressive learning paradigm in MECo [Chen et al. 2025a], we orchestrate the adaptation of the pre-trained LLM for generative motion synthesis through a three-stage curriculum. This regimen systematically bridges the modality gap: (1) Embedding Space Alignment, which projects the discrete audio and motion codebooks into the LLM's continuous latent manifold; (2) Acoustic-Kinematic Alignment, which conditions the backbone to synthesize motion from streaming audio; and (3) Exemplar-Driven Control, which fine-tunes the model to accept reference motions as explicit stylistic directives.

To unify the input modalities, we employ the causal motion tokenizer detailed in Sec. 3.1 for kinematic discretization. For the acoustic modality—spanning both conversational speech and complex music—we utilize a causal variant of EnCodec [Défossez et al. 2022] to provide a consistent discrete interface.

Crucially, we diverge from MECo's strategy of prioritizing the primary quantization layer. We posit that high-fidelity reconstruction requires the explicit modeling of the full residual hierarchy. Consequently, we adopt the flattened interleaving strategy from MusicGen [Copet et al. 2023], serializing the multi-layer RVQ indices into a single autoregressive stream. To account for the non-uniform information density across quantization levels—where the initial layer captures fundamental dynamics and subsequent layers encode high-frequency residuals—we implement a Hierarchical Loss Scaling strategy. We apply monotonically decaying weights to the cross-entropy objectives of deeper RVQ layers, guiding the optimization to prioritize structural coherence before refining fine-grained details.

3.2.1 Hierarchical Token Corruption. We observe that unifying multiple audio-to-motion tasks within a shared motion token space induces a catastrophic failure mode: conditional collapse. As elucidated in our theoretical analysis (see Appendix), this pathology stems from a fundamental conflict in the optimization dynamics. The autoregressive motion prior remains strong in the learning signal, effectively "short-circuiting" the weaker audio conditioning, particularly when task-specific data is sparse. The model learns to aggregate next-motion-token probability more heavily on recent motion history while ignoring the acoustic input.

To counteract this, we propose Hierarchical Token Corruption, a targeted regularization strategy designed to recalibrate these dynamics. By stochastically perturbing context motion tokens during training, we actively penalize over-reliance on the autoregressive history and force the model to rely on the mutual information between the audio condition and the target motion.

Unlike uniform noise injection, our perturbation strategy respects the structural hierarchy of Residual Vector Quantization (RVQ). For each timestep selected for corruption, we sample a layer depth $\ell_t \sim \text{Uniform}(1, L)$ and randomize tokens from layer ℓ_t through L, while leaving the coarser, foundational layers intact. This approach yields two critical benefits. First, it mimics realistic generation artifacts—where fine-grained details degrade before global structure—thereby serving as a robust data augmentation technique. Second, it instills error-correcting capabilities; the model learns to

recover ground-truth trajectories even when conditioned on perturbed context, ensuring graceful recovery from sampling errors during long-form autoregressive inference.

3.2.2 Example Control. Following MECo, we integrate exemplar-based control to steer generation. While our generative backbone models the full Residual Vector Quantization (RVQ) hierarchy to maximize fidelity, we observe that the semantic density of the motion signal is predominantly concentrated in the primary VQ layer. Consequently, we constrain the conditioning mechanism to extract control tokens exclusively from the first-level codebook of the reference sequence.

3.3 Reinforcement Learning

To enhance generation quality beyond the training distribution, we employ Reinforcement Learning (RL) to align the model's policy with human perceptual standards. We investigate two paradigms: Reward-Guided Optimization (leveraging self-supervised proxy rewards) and Direct Preference Alignment (leveraging human feedback).

3.3.1 Reward-Guided Optimization (GRPO). In scenarios lacking explicit human labels, we synthesize a proxy reward signal combining intrinsic motion fidelity and cross-modal synchronization. We optimize this objective using Group Relative Policy Optimization (GRPO) [Shao et al. 2024], which stabilizes training by normalizing advantages within a sampled group. The objective is formulated as:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G}\sum_{i=1}^{G} \rho_i \hat{A}_i + \beta_G \, \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{2}$$

where ρ_i denotes the importance ratio $\pi_\theta(y_i|x)/\pi_{\theta_{\text{old}}}(y_i|x)$, and $\hat{A}_i = (r_i - \mu_G)/\sigma_G$ represents the group-normalized advantage. The Kullback-Leibler (KL) divergence term ensures the optimized policy π_θ does not deviate excessively from the reference policy π_{ref}.

Self-Supervised Motion Quality Reward. Constructing a robust quality metric without manual annotation is non-trivial. Inspired by the degradation modeling in E3D2 [Wang et al. 2024], UnifiedGesture [Yang et al. 2023a] and D-REX [Brown et al. 2019], we establish a self-supervised quality curriculum by artificially corrupting ground-truth motion sequences. We apply variable rates of random and Hierarchical Token Corruption to generate a synthetic dataset with known degradation levels, calibrated via FID scores. This establishes a monotonic mapping between corruption severity and quality, which serves as the training signal for our reward model. The reward model architecture mirrors our motion tokenizer but utilizes bidirectional attention to capture global temporal context and omits the quantization layer to output continuous quality scalars.

Audio-Motion Alignment Reward. To evaluate rhythmic alignment, we learn a joint multimodal embedding space using the InfoNCE contrastive objective [Radford et al. 2021; van den Oord et al. 2018]. We utilize the pre-trained BEATs [Chen et al. 2023] model as the audio encoder and a randomly initialized Transformer as the motion encoder. The reward is defined as the cosine similarity between the synchronized audio and motion embeddings, encouraging the policy to maximize cross-modal coherence.

3.3.2 Direct Preference Alignment (DPO). When human feedback is available, we bypass proxy reward modeling and optimize the policy directly against human preferences using Direct Preference Optimization (DPO) [Rafailov et al. 2023]. This approach implicitly solves the reward maximization problem without the instability of a separate reward network:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta_D \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta_D \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{3}$$

where y_w and y_l denote the preferred (winning) and dispreferred (losing) motion sequences, respectively, and β_D modulates the strength of the KL constraint. To construct the preference dataset $\mathcal{D}$, we employ a Best-of-N sampling strategy: for each audio input, we generate eight candidate sequences using the pre-trained model. Human annotators perform a comparative evaluation to identify the optimal and least plausible samples, forming (y_w, y_l) pairs. To ensure high signal-to-noise ratio in the preference data, pairs lacking a distinct quality disparity are filtered out.

4 Experiment

4.1 Datasets and Preprocessing

To evaluate our framework across disparate kinematic domains, we leverage two complementary high-fidelity datasets: ZeroEGGS, a stylized speech-gesture corpus (approx. 2 hours) capturing a single speaker across 19 distinct expressive styles; and Motorica, a rhythmic dance database (approx. 6 hours) featuring five performers across eight diverse genres. To resolve topological discrepancies between sources, we adopt the standardized skeletal representation proposed by Holden [Holden 2024a,b]. Subsequently, we employ kinematic retargeting (Autodesk Maya) to transfer all motion data onto a unified target digital character.

RL Alignment Corpus. To facilitate reinforcement learning beyond the constraints of the original paired dataset, we curate a supplementary collection of unannotated audio. For the speech domain, we synthesize approximately one hour of conversational dialogue using Gemini 3 Pro scripts rendered via ElevenLabs' neural TTS. For the music domain, we assemble a diverse corpus of 100 compositions from open platforms (YouTube), covering a broad spectrum of tempos and genres to enhance rhythmic generalization.

Motion Refinement. We observe that a significant portion of the Motorica dataset exhibits artifacts characterized by erratic or static finger motion. To rectify this, we train a motion inpainting model [Shafir et al. 2024] leveraging the high-fidelity finger motion data from ZeroEGGS. This model synthesizes plausible finger motion conditioned on the remaining body joints. A Savitzky-Golay filter is then applied to the synthesized finger motion to mitigate temporal jitter.

Facial Animation. To animate our digital avatar, we utilize the industry-standard Apple ARKit blendshape schema. We curated a proprietary facial capture dataset (approx. 1 hour) consisting of synchronized speech and high-fidelity blendshape weights. The acquisition pipeline utilized an iPhone 12 running Live Link Face (Epic Games, Inc.), with the actor performing the Harvard Sentences [IEEE Subcommittee oe'en Subjective Measurements 1969] corpus to ensure comprehensive phonemic coverage. Based on our collected

data, we train a lightweight streaming speech-to-facial animation model following [Chen and Liu 2025]. More details are provided in the Appendix.

Table 2. Quantitative evaluation on test set. We report $BA_G \times 10^{-1}$, $BA_D \times 10^{-1}$. **Bold** face indicates the best result. "Ours" denotes the no-RL model.

Method	FID ↓	Diversity ↑	BA_G ↑	BA_D ↑
GT	0	21.52	7.775	2.619
MECo	14.73	23.13	7.507	2.622
EDGE	18.06	19.71	8.190	**2.668**
Ours (w/o corrupt)	25.92	**29.58**	**8.464**	2.541
Ours	**9.465**	20.70	8.277	2.603
Ours (DPO)	12.39	19.67	8.283	2.607
Ours (GRPO)	24.13	20.89	8.239	2.618

4.2 Settings

Our system synthesizes native motion at 30 frames per second (FPS), which is subsequently interpolated to 60 FPS for final rendering. We detail the configuration for each component. The RVQ-VAE (Sec. 3.1) is trained with a temporal downsampling factor $n/N = 4$, yielding a latent motion rate of 7.5 Hz. We employ a codebook size $K = 512$, latent dimension $d = 512$, and quantization depth $Q = 6$. The model is optimized using a batch size of 256, a commitment loss weight $\eta = 0.1$, and a learning rate of 4×10^{-4} managed by a step decay scheduler. During training, we randomly sample 64-frame motion windows. We adopt Qwen2.5-0.5B-Instruct [Yang et al. 2024] as the base generator, detaching its tied input/output embeddings to accommodate our modality-specific vocabularies. The model processes 4-second context windows, comprising 600 audio tokens (derived from the first 2 RVQ layers of EnCodec at 75Hz) and 540 motion tokens (flattened across 6 RVQ layers for three body partitions). Fine-tuning is performed with a batch size of 256 and a learning rate of 5×10^{-5}. For the reinforcement learning stage, we reduce the batch size to 16 and adjust the learning rate to 6×10^{-5}. In the GRPO configuration, we set the KL penalty $\beta_G = 0.01$ and perform 30 rollouts per prompt. For DPO, we utilize a deviation penalty $\beta_D = 0.1$. All experiments are conducted on a node equipped with two NVIDIA H200 GPUs. The complete training pipeline requires approximately 30 hours. At inference, our optimized pipeline achieves a throughput of ∼300 tokens/s, well within the latency budget for real-time interaction.

4.3 Real-time Deployment

To enable user-friendly interactive avatars, as shown in Figure 6, we build a distributed system composed of three functional tiers: a cloud-hosted Conversational Voice Agent (ElevenLabs), a rendering Client Frontend, and a dedicated GPU Inference Server. We achieve continuous autoregressive streaming by deploying our fixed-context trained model via a sliding window strategy, generating motion in granular steps of 0.266 seconds (8 frames). We leverage CUDA Graph instantiation to reduce kernel scheduling overhead. Latency profiling across four processing stages (see appendix), including Audio Encoding, Motion Synthesis, Motion Decoding, and IK Post-processing, confirms that our total computational latency remains well below the 266ms audio chunk duration on both NVIDIA H200 and RTX 4090 platforms.

4.4 Subjective Evaluation Protocol

Following established subjective evaluation standards [Alexanderson et al. 2023; Ao et al. 2023], we assess generation quality across three perceptual dimensions: Human Likeness, Rhythmic Synchronization (Beat Matching), and Overall Preference. We adopt a rigorous pairwise comparison protocol: for each trial, participants are presented with two sequential 10-second clips synthesized by competing models conditioned on identical audio inputs. Evaluators indicate both the direction and intensity of their preference on a 5-point Likert scale (0: Neutral, 2: Strong Preference). To facilitate quantitative analysis, these ordinal ratings are mapped to a symmetric interval $[-2, 2]$, where positive values signify a preference for our method. The final subjective score is aggregated from 1,680 individual pairwise judgments, ensuring statistical significance.

4.5 Quantitative Benchmarking

Given our framework's unified capability, we evaluate performance across both speech-to-gesture and music-to-dance domains. We standardize the measurement of distribution fidelity (FID) and generative Diversity across tasks. For rhythmic alignment, we employ domain-specific heuristics to capture the distinct temporal dynamics of each modality: for speech-gesture alignment, following EMAGE, we quantify the synchronization between acoustic onsets and the local minima of kinematic velocity; for music-dance alignment, following [Davis and Agrawala 2018], we assess the correspondence between musical beats and the local maxima of motion deceleration (Detailed formulations for all metrics are provided in the Appendix).

We benchmark against leading domain-specific baselines: MECo for co-speech gesture and EDGE for music-driven dance. As summarized in Table 2 and Table 3, our unified approach consistently surpasses these specialized baselines in both objective metrics and subjective preference. Furthermore, evaluations on the high-fidelity BEAT2 benchmark (Table 4) confirm that our method establishes a new state-of-the-art in generative fidelity (FID).

4.6 Ablation Study

4.6.1 Attention-based Causal Motion Tokenizer. We validate our motion tokenizer through four ablation studies and two controlled comparisons, assessing both intrinsic reconstruction fidelity and downstream generation efficacy. To quantify reconstruction quality, we report FID, MPJPE, and a Translation Loss (*Trans Loss*), defined as the deviation between predicted and ground-truth root velocities. To assess downstream impact, we evaluate the FID of an audio-to-motion generator trained atop each tokenizer variant. Across all experiments, the inclusion of each proposed component yields consistent improvements in both signal reconstruction and generative quality. Furthermore, to ensure rigorous benchmarking, we compare against a CausalConv baseline implemented with an identical RVQ configuration and loss landscape; our attention-based approach demonstrates superior performance across all metrics.

Table 3. User Study Results. We evaluate our method on both Dance and Gesture generation tasks across three comparative settings. The metrics reported are Human Likeness, Beat Matching, and Overall Preference. All results are presented as *mean ± 95% confidence interval*. The three categories are independent.

Category	Method	Dance			Gesture		
		Human Likeness	Beat Matching	Overall Preference	Human Likeness	Beat Matching	Overall Preference
Comparison with SOTA	MECo	−0.508 ± 0.244	−0.277 ± 0.230	−0.477 ± 0.236	0.235 ± 0.198	0.061 ± 0.222	0.096 ± 0.216
	EDGE	0.148 ± 0.238	−0.136 ± 0.174	0.099 ± 0.229	−0.676 ± 0.154	−0.705 ± 0.149	−0.748 ± 0.143
	Ours	0.244 ± 0.237	0.337 ± 0.188	0.267 ± 0.234	0.588 ± 0.150	0.798 ± 0.155	0.816 ± 0.137
RL Strategy Ablation	Ours	−0.078 ± 0.156	−0.028 ± 0.137	−0.085 ± 0.170	−0.231 ± 0.262	0.000 ± 0.219	−0.169 ± 0.269
	Ours (DPO)	0.078 ± 0.156	0.028 ± 0.137	0.085 ± 0.170	0.231 ± 0.262	0.000 ± 0.219	0.169 ± 0.269
	Ours	−0.109 ± 0.215	−0.069 ± 0.188	−0.188 ± 0.211	−0.109 ± 0.193	−0.092 ± 0.162	−0.059 ± 0.191
	Ours (GRPO)	0.109 ± 0.215	0.069 ± 0.188	0.188 ± 0.211	0.109 ± 0.193	0.092 ± 0.162	0.059 ± 0.191
Dataset Composition	Gesture Only	-	-	-	−0.345 ± 0.192	−0.727 ± 0.193	−0.555 ± 0.198
	Dance Only	−0.350 ± 0.141	−0.355 ± 0.124	−0.323 ± 0.136	-	-	-
	Merged	0.350 ± 0.141	0.355 ± 0.124	0.323 ± 0.136	0.345 ± 0.192	0.727 ± 0.193	0.555 ± 0.198

Table 4. Comparison with the state-of-the art methods on BEAT2 [Liu et al. 2024b] test set. Quantitative evaluation on BEAT2. We report FID $\times 10^{-1}$, BA_G $\times 10^{-1}$, and diversity. **Bold** face indicates the best result.

Method	FID ↓	BA_G ↑	Diversity ↑
S2G[Ginosar et al. 2019]	28.15	4.683	5.971
Trimodal[Yoon et al. 2020]	12.41	5.933	7.724
HA2G[Liu et al. 2022c]	12.32	6.779	8.626
DisCo[Liu et al. 2022a]	9.417	6.439	9.912
CaMN[Liu et al. 2022d]	6.644	6.769	10.86
DiffStyleGesture[Yang et al. 2023b]	8.811	7.241	11.49
Habibie *et al.*[Habibie et al. 2021]	9.040	7.716	8.213
TalkShow[Yi et al. 2023]	6.209	6.947	13.47
EMAGE [Liu et al. 2024b]	5.512	7.724	13.06
SynTalker[Chen et al. 2024b]	6.413	7.971	12.72
MECo [Chen et al. 2025a]	3.401	7.346	**15.30**
ViBES [Zhang et al. 2026b]	5.257	**8.103**	13.03
PersonaGesture [Zhang et al. 2026a]	3.930	7.100	13.25
Ours	**2.874**	7.342	13.53

4.6.2 Hierarchical Token Corruption. We identify Hierarchical Token Corruption as the linchpin of our unified training strategy. As illustrated in Table 2 and Figure 5, ablating this mechanism leads to severe conditional collapse: the model ignores the input condition and persistently generates meaningless, physically implausible dance-like motions even during silence or neutral speech. Paradoxically, this pathological behavior results in the highest scores for Diversity and BA_G, as the ungrounded, high-variance movements artificially inflate these metrics without reflecting genuine perceptual quality. By reintroducing our hierarchical corruption strategy, the model successfully learns to adhere to the acoustic signal, enabling label-free learning from the joint dataset. Moreover, the corruption-augmented model achieves superior performance on individual tasks compared to single-task baselines, demonstrating that it effectively learns from cross-task training data.

4.6.3 Cross-Modal Synergy via Joint Training. We further investigate the efficacy of dataset composition by comparing three training configurations: Gesture-Only, Dance-Only, and Combined. As detailed in Table 3, joint training yields a performance uplift across both domains. Most notably, the inclusion of the music-to-dance dataset significantly enhances the beat-matching capability of the gesture generation. We attribute this to cross-modal synergy: the model internalizes robust rhythmic priors from the highly structured dance data and transfers this sensitivity to the speech domain. This transfer is particularly vital for gesture subsets with sparse rhythmic cues (e.g., "Still" or "Flirty" styles), where the speaker exhibits low kinematic variance. Furthermore, we observe an emergent zero-shot stylistic transfer: as shown in Figure 4, when driven by highly energetic "happy" speech, the agent occasionally produces lively, rhythmic gestures that were not present in the original speech dataset. This suggests that our unified framework possesses a degree of semantic generalization, mapping audio features to motion primitives regardless of the source domain.

4.6.4 Reinforcement Learning Strategy Analysis. We conduct a comparative analysis of two alignment strategies: Direct Preference Optimization (DPO), utilizing human preference labels, and Group Relative Policy Optimization (GRPO), utilizing proxy rewards. Table 3 confirms that both methods successfully align the model with human perceptions, improving subjective ratings over baseline.

Reward Model Efficacy. To validate the proxy signals used in GRPO, we evaluate our trained reward models on held-out test data. As depicted in Fig. 7, the motion quality model exhibits a Pearson correlation of 0.9977 with ground-truth degradation levels, correctly preserving the ordinal ranking of motion quality across all corruption intensities. We also tested audio-motion alignment reward with retrieval metrics following the evaluation protocol of TMR [Petrovich et al. 2023], which achieves a retrieval success rate approximately 100× higher than random chance, confirming its discriminative effectiveness. Please see appendix for details.

The Alignment-Fidelity Trade-off. Despite the robustness of our reward models, Table 2 reveals a characteristic trade-off: both RL strategies induce a degradation in FID scores, with GRPO exhibiting a more pronounced divergence (9.465 → 24.13) compared to DPO (9.465 → 12.39). This outcome is consistent with prior observations that reward optimization induces mode-seeking behavior: the policy concentrates mass on high-reward modes, which reduces distributional coverage (and thus inflates FID) while improving alignment with the target reward and human preference [Ouyang et al. 2022b].

Domain-Specific Strategy Selection. Our user study reveals divergent effectiveness across motion domains. GRPO achieves stronger preference improvements on dance (Overall: +0.188 vs. DPO's +0.085), while DPO outperforms GRPO on gesture (+0.169 vs. +0.059). We attribute this to domain characteristics: dance motion favors strong rhythmic synchronization and tolerates exaggerated movements (or even benefits from them), aligning well with GRPO's aggressive optimization. Conversely, conversational gestures prioritize subtlety and naturalness, which are better preserved by DPO's conservative, preference-based learning.

5 Discussions and Future Work

While this work establishes a robust baseline for unified real-time animation, several frontiers remain for future investigation. First, our current architecture decouples facial and body dynamics and lacks detailed non-verbal interaction modeling such as gaze, limiting the holistic cohesion required for deep engagement. Second, reliance on acoustic features combined with limited speaker diversity in our training data can lead to domain confusion, such as misidentifying male speech as musical vocals and erroneously generating dance motions. Furthermore, the system currently lacks specific transition policies for abrupt acoustic terminations, leading to unnatural motion especially when music stops suddenly. Finally, our framework focuses exclusively on the speaker role, neglecting the reciprocal nature of dyadic communication. Realizing true embodied interaction necessitates extending our generative paradigm to support active listening, enabling the avatar to synthesize non-verbal backchannels and reactive behaviors in response to user input [Ng et al. 2022] or environmental context [Xu et al. 2025].

Acknowledgments

We are grateful to Linzhou Li for refining the teaser image, and to Jiacheng Guo and Yixuan Lai for their extensive efforts in manually evaluating and annotating the generated results. This work is partially supported by NSF China (No. 62572430, 62421003) and the XPLORER PRIZE.

References

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired Motion Style Transfer from Video to Animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496.

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.* 42, 4, Article 44 (July 2023), 20 pages.

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–19.

Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* (2023), 18 pages.

Jinseok Bae, Inwoo Hwang, Young-Yoon Lee, Ziyu Guo, Joseph Liu, Yizhak Ben-Shabat, Young Min Kim, and Mubbasir Kapadia. 2025. Less is more: Improving motion diffusion models with sparse keyframes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11069–11078.

Yarin Bekor, Gal Michael Harari, Or Perel, and Or Litany. 2025. Gaussian See, Gaussian Do: Semantic 3D Motion Transfer from Multiview Video. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–10.

Daniel S. Brown, Wonjoon Goo, and Scott Niekum. 2019. Better-than-Demonstrator Imitation Learning via Automatically-Ranked Demonstrations. In *Proceedings of the 3rd Conference on Robot Learning*.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated Conversation: Rule-Based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '94)*. Association for Computing Machinery, New York, NY, USA, 413–420.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: The Behavior Expression Animation Toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. Association for Computing Machinery, New York, NY, USA, 477–486.

Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024b. Enabling Synergistic Full-Body Control in Prompt-Based Co-Speech Motion Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, New York, NY, USA, 10.

Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. 2025a. Motion-example-controlled Co-speech Gesture Generation Leveraging Large Language Models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3721238.3730611

Bohong Chen and Haiyang Liu. 2025. DyStream: Streaming Dyadic Talking Heads Generation via Flow Matching-based Autoregressive Model. arXiv:2512.24408 [cs.CV]

Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. 2025c. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6200–6211.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. 2024a. Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models. *arXiv preprint arXiv:2410.10733* (2024).

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. 2024c. DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation. In *CVPR*.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. 2025b. Motionllm: Understanding human behaviors from human motions and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio Pre-Training with Acoustic Tokenizers. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 5178–5193.

Qingrong Cheng, Xu Li, and Xinghui Fu. 2024. SIGGesture: Generalized Co-Speech Gesture Synthesis via Semantic Injection with Large-Scale Pre-Training Diffusion Models. In *SIGGRAPH Asia 2024 Conference Papers (SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 133, 11 pages.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems* 36 (2023), 47704–47720.

Abe Davis and Maneesh Agrawala. 2018. Visual rhythm and beat. *ACM Trans. Graph.* 37, 4, Article 122 (July 2018), 11 pages. doi:10.1145/3197517.3201371

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438* (2022).

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. 2025. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13336–13348.

Qiao Feng, Yiming Huang, Yufu Wang, Jiatao Gu, and Lingjie Liu. 2025. PhysHMR: Learning Humanoid Control Policies from Vision for Physically Plausible Human Motion Reconstruction. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–10.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Computer Graphics Forum* 42, 1 (2023), 206–216. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14734

Anindita Ghosh, Bing Zhou, Rishabh Dabral, Jian Wang, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek, and Chuan Guo. 2025. Duetgen: Music driven two-person dance generation via hierarchical masked modeling. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–11.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. MoMask: Generative Masked Modeling of 3D Human Motions. (June 2024), 1900–1910.

Tuomas Haarnoja et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905* (2018).

Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 46, 9 pages.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Speech-driven 3D Conversational Gestures from Video. *arXiv preprint arXiv:2102.06837* (2021).

Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. 2025. Syncdiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11731–11743.

Daniel Holden. 2024a. *Motorica-Retarget*. https://github.com/orangeduck/motorica-retarget

Daniel Holden. 2024b. *ZeroEGGs-Retarget*. https://github.com/orangeduck/zeroeggs-retarget

Ruibing Hou, Mingshuang Luo, Hongyu Pan, Hong Chang, and Shiguang Shan. 2025. Motionverse: A unified multimodal framework for motion comprehension, generation and editing. *arXiv preprint arXiv:2509.23635* (2025).

IEEE Subcommittee oe'en Subjective Measurements. 1969. IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics* 17, 3 (1969), 225–246. doi:10.1109/TAU.1969.1162058

Biao Jiang, Xin Chen, Ailing Zeng, Xinru Sun, Fukun Yin, Xianfang Zeng, Xuanyang Zhang, Gang Yu, and Tao Chen. 2025. Causal Motion Tokenizer for Streaming Motion Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2024–2034.

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, et al. 2025. PersonaBooth: Personalized Text-to-Motion Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22756–22765.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. 2006. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents* (Marina Del Rey, CA) *(IVA'06)*. Springer-Verlag, Berlin, Heidelberg, 205–217.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 242–250.

Aviral Kumar et al. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *NeurIPS*.

Jina Lee and Stacy Marsella. 2006. Nonverbal Behavior Generator for Embodied Conversational Agents *(IVA '06)*. Springer, 243–255.

Margot Lhommet, Yuyu Xu, and Stacy Marsella. 2015. Cerebella: Automatic Generation of Nonverbal Behavior for Virtual Humans *(AAAI '15, 1)*.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.

Lei Li et al. 2023b. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665* (2023).

Siyao Li et al. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. In *CVPR*.

Weiyu Li, Xuelin Chen, Peizhuo Li, Olga Sorkine-Hornung, and Baoquan Chen. 2023a. Example-Based Motion Synthesis via Generative Motion Matching. *ACM Transactions on Graphics (TOG)* 42, 4, Article 94 (2023).

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. 2020. Character controllers using motion VAEs. *ACM Trans. Graph.* 39, 4, Article 40 (Aug. 2020), 12 pages. doi:10.1145/3386569.3392422

Binjie Liu, Lina Liu, Sanyi Zhang, Songen Gu, Yihao Zhi, Tianyi Zhu, Lei Yang, and Long Ye. 2025b. MAG: Multi-Modal Aligned Autoregressive Co-Speech Gesture Generation without Vector Quantization. *arXiv preprint arXiv:2503.14040* (2025).

Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022a. DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 3764–3773.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Naoya Iwamoto, Bo Zheng, and Michael J. Black. 2024b. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Masked Audio Gesture Modeling. arXiv:2401.00374 [cs.CV]

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022d. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. *arXiv preprint arXiv:2203.05297* (2022).

Pinxin Liu, Luchuan Song, Junhua Huang, and Chenliang Xu. 2025c. GestureLSM: Latent Shortcut based Co-Speech Gesture Generation with Spatial-Temporal Modeling. In *IEEE/CVF International Conference on Computer Vision*.

Sheng Liu, Yuanzhi Liang, Jiepeng Wang, Sidan Du, Chi Zhang, and Xuelong Li. 2025a. Uni-Inter: Unifying 3D Human Motion Synthesis Across Diverse Interaction Contexts. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–11.

Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. 2022b. Audio-Driven Co-Speech Gesture Video Generation. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 21386–21399.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022c. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.

Zixuan Liu et al. 2024a. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475* (2024).

Jintao Lu, He Zhang, Yuting Ye, Takaaki Shiratori, Sebastian Starke, and Taku Komura. 2025b. CHOICE: Coordinated human-object interaction in cluttered environments for pick-and-place actions. *ACM Transactions on Graphics* 45, 2 (2025), 1–18.

Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. 2025a. Scamo: Exploring the scaling law in autoregressive motion generation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 27872–27882.

Shuhong Lu, Youngwoo Yoon, and Andrew W. Feng. 2023. Co-Speech Gesture Synthesis using Discrete Gesture Token Learning. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2023), 9808–9815.

Jacob Menick et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147* (2022).

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. 2024. ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis. In *Computer Vision and Pattern Recognition (CVPR)*.

M Hamza Mughal, Rishabh Dabral, Merel CJ Scholman, Vera Demberg, and Christian Theobalt. 2025. Retrieving Semantics from the Deep: an RAG Solution for Gesture Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 16578–16588.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning To Listen: Modeling Non-Deterministic Dyadic Facial Motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20395–20405.

Quang Nguyen, Tri Le, Baoru Huang, Minh Nhat Vu, Ngan Le, Thieu Vo, and Anh Nguyen. 2025. Learning Human Motion with Temporally Conditional Mamba. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–10.

Long Ouyang et al. 2022a. Training language models to follow instructions with human feedback. *NeurIPS* (2022).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

Mathis Petrovich, Michael J. Black, and Gül Varol. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *International Conference on Computer Vision (ICCV)*.

André Susano Pinto et al. 2023. Tuning computer vision models with task rewards. *arXiv preprint arXiv:2302.08242* (2023).

Zhiping Qiu, Yitong Jin, Yuan Wang, Yi Shi, Chao Tan, Chongwu Wang, Xiaobing Li, Feng Yu, Tao Yu, and Qionghai Dai. 2025. ELGAR: Expressive Cello Performance Motion Generation for Audio Rendition. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–9.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

Rafael Rafailov et al. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS* (2023).

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. 2025. Mixermdm: Learnable composition of human motion diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12380–12390.

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. 2024. Human Motion Diffusion as a Generative Prior. In *The Twelfth International Conference on Learning Representations*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Alan Song, Mingchuan Xiao, Y. K. Li, Y. Zhang, Ins Zhang, Y. Wang, et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

Shuaijie She et al. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838* (2024).

Min Shi, Wenke Feng, Lin Gao, and Dengming Gao. 2024a. Generating diverse clothed 3D human animations via a generative model. *Computational Visual Media* 10, 2 (2024), 261–277.

Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. 2024b. Interactive Character Control with Auto-Regressive Motion Diffusion Models. *ACM Trans. Graph.* 43 (jul 2024).

Mingyang Sun et al. 2023. Co-speech Gesture Synthesis by Reinforcement Learning with Contrastive Pre-trained Rewards. *CVPR* (2023).

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction*. MIT Press.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *NeurIPS* 12 (1999).

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. 2023. TLControl: Trajectory and Language Control for Human Motion Synthesis. *arXiv preprint arXiv:2311.17135* (2023).

Zilin Wang, Haolin Zhuang, Lu Li, Yinmin Zhang, Junjie Zhong, Jun Chen, Yu Yang, Boshi Tang, and Zhiyong Wu. 2024. Explore 3d dance generation via reward model from automatically-ranked demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 301–309.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning* (1992), 5–32.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. 2025. MotionStreamer: Streaming Motion Generation via Diffusion-based Autoregressive Model in Causal Latent Space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10086–10096.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2023. OmniControl: Control Any Joint at Any Time for Human Motion Generation. arXiv:2310.08580

Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, and Taku Komura. 2025. MOSPA: Human Motion Generation Driven by Spatial Audio. In *Advances in Neural Information Processing Systems*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

Quanwei Yang, Luying Huang, Kaisiyuan Wang, Jiazhi Guan, Shengyi He, Fengguo Li, Hang Zhou, Lingyun Yu, Yingying Li, Haocheng Feng, et al. 2025. GestureHYDRA: Semantic Co-speech Gesture Synthesis via Hybrid Modality Diffusion Transformer and Cascaded-Synchronized Retrieval-Augmented Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12615–12625.

Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Qiaochu Huang, Lei Hao, Songcen Xu, Xiaofei Wu, Changpeng Yang, and Zonghong Dai. 2023a. UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 1033–1044. doi:10.1145/3581783.3612503

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023b. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, 5860–5868.

Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. 2024. MoConVQ: Unified Physics-Based Motion Control via Scalable Discrete Representations. *ACM Trans. Graph.* 43, 4, Article 144 (July 2024), 21 pages.

Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3100–3107.

Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. 2022. Audio-driven stylized gesture generation with flow-based model. In *European Conference on Computer Vision*. Springer, 712–728.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating Holistic 3D Human Motion from Speech. In *CVPR*.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

Hongyi Yuan et al. 2023. Rrhf: Rank responses to align language models with human feedback. *NeurIPS* (2023).

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 495–507. doi:10.1109/TASLP.2021.3129994

Juze Zhang, Changan Chen, Xin Chen, Heng Yu, Tiange Xiang, Ali Sartaz Khan, Shrinidhi Kowshika Lakshmikanth, and Ehsan Adeli. 2026b. ViBES: A Conversational Agent with Behaviorally-Intelligent 3D Virtual Body. In *CVPR*.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001* (2022).

Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, and Ziwei Liu. 2024b. Large Motion Model for Unified Multi-modal Motion Generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XIII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 397–421.

Xiangyue Zhang, Yiyi Cai, Kunhang Li, Kaixing Yang, You Zhou, Zhengqing Li, Xuangeng Chu, Jiaxu Zhang, and Haiyang Liu. 2026a. PersonaGesture: Single-Reference Co-Speech Gesture Personalization for Unseen Speakers. arXiv:2605.06064

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024a. Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis. *ACM Trans. Graph.* (2024), 17 pages.

Li Zhao and Zhengmin Lu. 2024. DanceFusion: A Spatio-Temporal Skeleton Diffusion Transformer for Audio-Driven Dance Motion Reconstruction. *arXiv preprint arXiv:2411.04646* (2024).

Zhiyuan Zhao et al. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839* (2023).

Dingcheng Zhen, Shunshun Yin, Shiyang Qin, Hou Yi, Ziwei Zhang, Siyuan Liu, Gan Qi, and Ming Tao. 2025. Teller: Real-Time Streaming Audio-Driven Portrait Animation with Autoregressive Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21075–21085.

Bowen Zheng, Ke Chen, Yuxin Yao, Zijiao Zeng, Xinwei Jiang, He Wang, Joan Lasenby, and Xiaogang Jin. 2025. Autokeyframe: Autoregressive keyframe generation for human motion synthesis and editing. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.

Chi Zhou, Tengyue Bian, and Kang Chen. 2022. GestureMaster: Graph-based Speech-driven Gesture Generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 764–770.

Yiyang Zhou et al. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411* (2024).

Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. 2019. On the Continuity of Rotation Representations in Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2018. Auto-Conditioned Recurrent Networks for Extended Complex Human Motion Synthesis. In *International Conference on Learning Representations*.

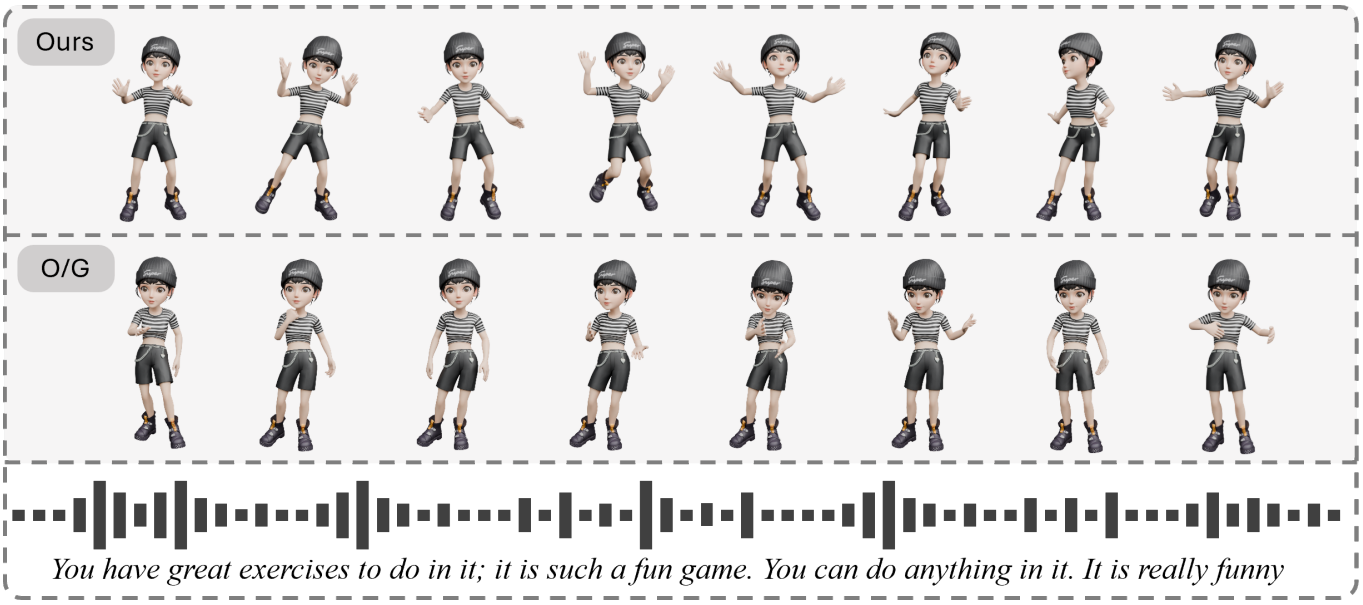

Fig. 4. O/G denotes Gesture Only, which training exclusively on the speech-gesture dataset. As shown, our model trained jointly on both speech-gesture and music-dance datasets can produce exuberant, dance-like movements in response to cheerful audio, demonstrating its ability to generalize across motion domains and adapt motion style to audio characteristics.

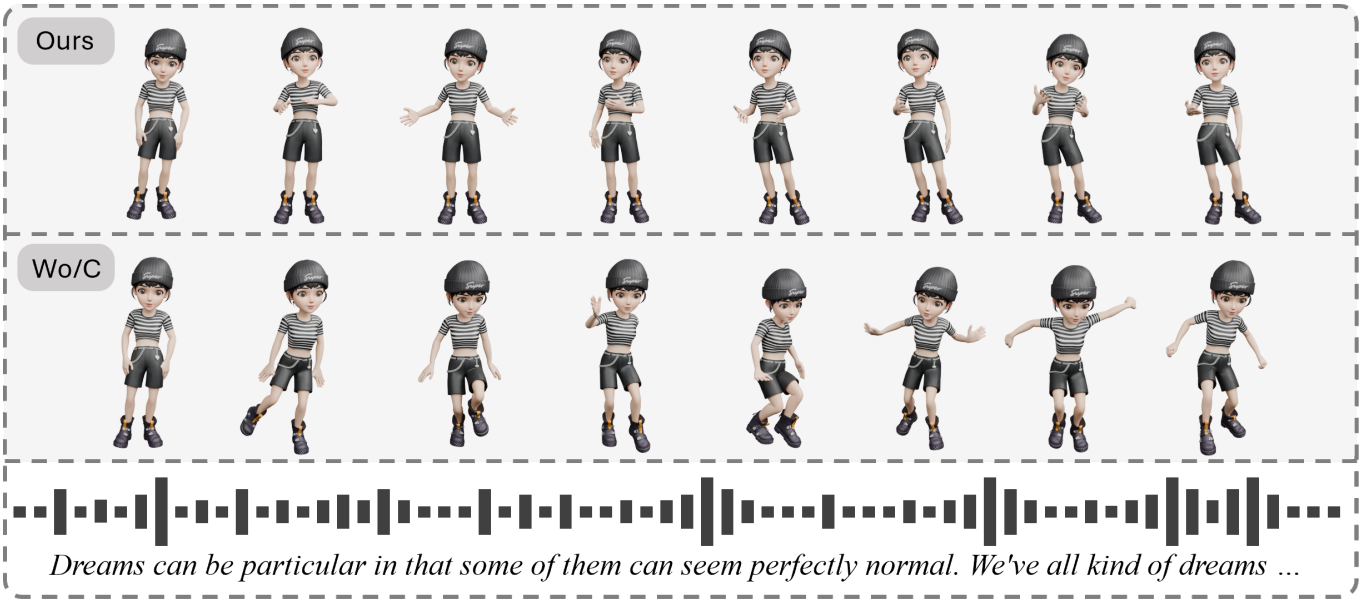

Fig. 5. W/o C denotes training without Hierarchical Token Corruption. Given the same audio and initial motion input, our method generates natural motions that are well-synchronized with the audio. In contrast, the variant trained without Hierarchical Token Corruption largely ignores the audio input and produces erratic, dance-like motions that lack proper audio-motion correspondence.

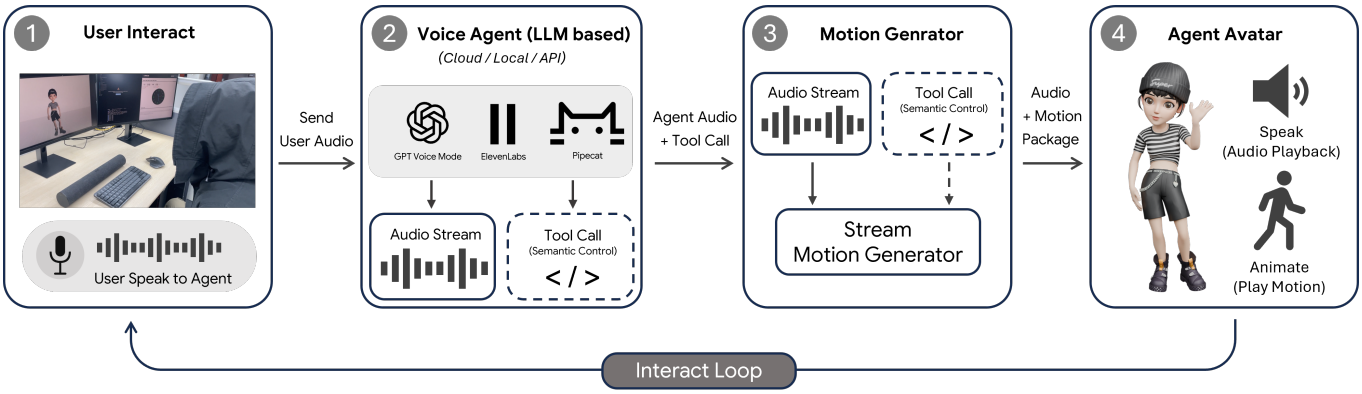

Fig. 6. Real-time Deployment. Our system comprises three components: the user host machine, a voice agent, and the Motion Generator. The host machine captures the user's voice via microphone (1) and streams it to the voice agent (2). The voice agent can be an omni-model (e.g., OpenAI's GPT voice mode) or a cascaded pipeline of VAD, ASR, LLM, and TTS modules (e.g., ElevenLabs, Pipecat), and can be deployed in the cloud, run locally, or accessed via API. It outputs an audio stream and, when appropriate, emits semantic control signals through a tool-call interface. The Motion Generator (3) consumes the audio stream and synchronously produces a motion stream, optionally conditioned on a motion example retrieved via the semantic control signal. The time-aligned audio and motion are then packaged and sent to the Rendering Client Frontend on the host machine to drive and visualize the avatar (4), closing the interaction loop.

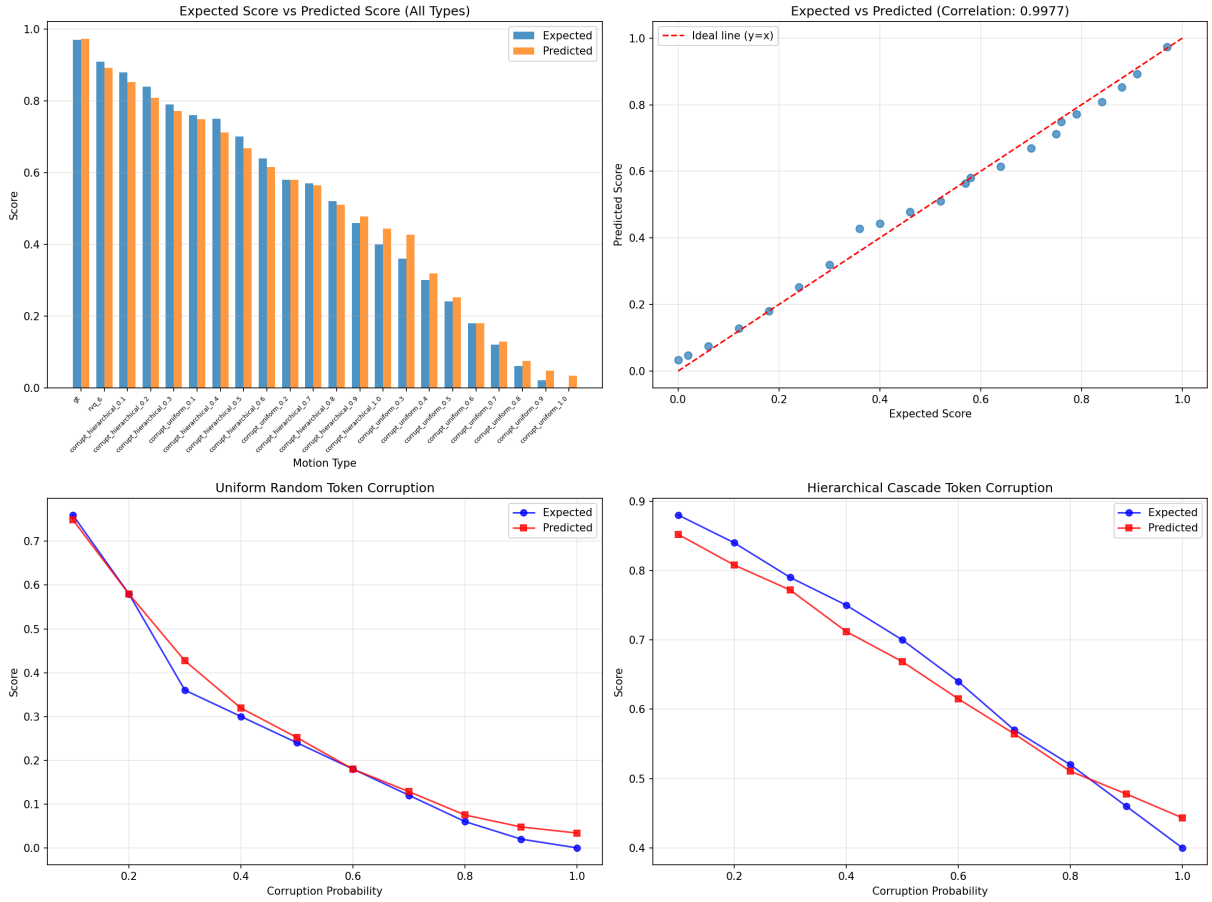

Fig. 7. Motion Quality Reward Model Evaluation. The four plots demonstrate the performance of our motion quality reward model on the validation set under different corruption strategies. The visualization shows that our reward model exhibits strong generalization across various types of motion degradation.

A More Details on Real-time Deployment

A.1 Distributed System Topology

To enable high-fidelity interactive avatars, we architect a distributed system composed of three functional tiers: the Conversational Agent, the Client Frontend, and the Inference Server. The conversational Agent, hosted on the ElevenLabs platform, orchestrates the dialogue management and executes semantic tool calls. The client Frontend (Local Host) acts as the rendering terminal. It streams the agent's audio output to the backend while simultaneously rendering the visual avatar state. The inference Server (Remote) is a dedicated GPU backend that ingests the audio stream and synthesizes full-body motion in real-time. The synchronized audio and motion streams are looped back to the client for playback. To resolve geometric interpenetration artifacts inherent to retargeting, we implement a post-processing Inverse Kinematics (IK) solver on the generated motion (detailed in Appendix).

A.2 Streaming Inference Optimization

We achieve continuous autoregressive streaming by deploying our fixed-context trained model via a sliding window strategy, generating motion in granular steps of 0.266 seconds (8 frames). To eliminate latency jitter caused by dynamic kernel scheduling, we leverage CUDA Graph instantiation; by capturing the execution graphs of the Motion Tokenizer and Face Generator during initialization, we optimize memory allocation and kernel launch overhead to stabilize inference times.

A.3 Latency Profiling & Budgeting

We evaluate real-time viability by conducting a granular breakdown of the processing pipeline for each 266ms audio chunk. The computational latency is profiled across four main distinct stages: Audio Encoding, Motion Synthesis, Motion Decoding, and IK Post-processing. Benchmarks are reported on both datacenter-grade (NVIDIA H200) and consumer-grade (NVIDIA RTX 4090) hardware, with statistics (Mean ± Std) aggregated over 20 intermediate inference steps to ensure reliability. As shown in Table 6, our latency remains below the audio chunk duration on both platforms, demonstrating that our method satisfies real-time processing requirements.

Beyond computational costs, we explicitly allocate a 100ms synchronization buffer at the client side to absorb playback jitter. Additionally, for the cloud-baesed Voice Agent demonstration, we account for an unavoidable network transmission latency of approximately 300ms introduced by the third-party service (ElevenLabs).

B Online Post-Processing

To adapt to stylized avatar model in real-time setting, we apply a light-weight inverse kinematic (IK) post-processing to mitigate with self-penetration, focusing on both hands. In online streaming setting, we don't have future context to refer to when processing current frame, thus an existing frame cannot move out to smoothly interpolate to a frame which got pushed out due to self-penetration, thus creating hard and sudden visual artifacts of "pushing-out". Instead, based on the model width, we define a smoothly interpolated cylinder-like shape around character's spine, and smoothly project the space inside the cylinder to the outside of cylinder in its local horizontal plane, effectively defining a unified rule to smoothly avoid end effector from entering a manually configured region, avoiding penetration detection and jitter-ish post processing fix. The post-processing is done by solely adjusting the shoulder rotation, so that the shoulder-to-hand vector's direction align with shoulder-to-target vector with a simple swing adjustment. This post-processing requires no optimization, is smoothly-defined and light weight. It costs roughly 10ms when implemented in torch, and for the sake of simplicity we did not perform further optimization.

C Theoretical Analysis

C.1 Gradient Equilibrium and Context Accumulation

We formulate the training objective as minimizing the Negative Log-Likelihood (NLL) of the target motion token x given the motion history h and audio condition c. The probability of a token x_i is modeled via the Softmax function over a logit z_i, which we decompose into an additive context component $\phi(x_i, h)$ and a condition component $\psi(x_i, c)$:

$$P(x_i|h, c) = \frac{\exp(\phi(x_i, h) + \psi(x_i, c))}{\sum_{j \in \mathcal{V}} \exp(z_j)} \tag{4}$$

The gradient of the loss with respect to the shared context parameter ϕ for a candidate token x_k is given by:

$$\nabla_\phi \mathcal{L} = P(x_k|h, c) - \mathbb{I}(x_k = x_{gt}) \tag{5}$$

Consider a "cross-road" history h where K distinct trajectories intersect. Let π_k denote the empirical probability of trajectory k occurring given h in the unified dataset $\mathcal{D}$. At the optimization stationary point, the expected gradient over $\mathcal{D}$ must be zero:

$$\mathbb{E}_{\mathcal{D}}\left[P(x_k|h, c)\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}(x_k = x_{gt})\right] = \pi_k \tag{6}$$

This equilibrium condition implies that the context-driven component ϕ accumulates sufficient magnitude to approximate the marginal distribution of the dataset. Under the approximation of the Softmax log-probability relationship, the learned context representation converges to:

$$\mathbb{E}[\phi(x_k, h)] \log(\pi_k) + C \tag{7}$$

This relationship establishes that the shared history h induces a *logit floor* for all intersecting trajectories, strictly proportional to their data frequency.

C.2 Min-Max Analysis of Interference Significance

At inference, conditioned on task k (audio c_k), we analyze the interference caused by an unrelated trajectory x_j ($j \neq k$). Assuming orthogonality of condition representations ($\mathbb{E}[\psi(x_j, c_k)]0$), the logit for the incorrect token depends primarily on the context:

$$\mathbb{E}[z(x_j)]\mathbb{E}[\phi(x_j, h)] \propto \log(\pi_j) \tag{8}$$

To demonstrate that this interference is non-trivial, we perform a best-case analysis to find the lower bound of the interference. We solve for the data distribution $\vec{\pi}$ that minimizes the maximum interference from the dominant wrong path:

$$\min_{\vec{\pi}} \left(\max_{j \neq k} \log(\pi_j)\right) \quad \text{s.t.} \quad \sum_{i=1}^{K} \pi_i = 1 \tag{9}$$

Table 5. Experiments on retrieval ability of audio-motion contrastive space. Details are in Sec. F.3

Protocol	BaseModel	Audio-motion retrieval						Motion-audio retrieval					
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	MRR↑	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	MRR↑
(a) All	**BEATs**	**6.42**	**10.51**	**14.60**	**21.79**	**67.0**	**11.07**	**6.64**	**10.51**	**13.77**	**18.75**	**75.0**	**10.70**
(N = 1808)	**Wav2CLIP**	3.60	6.58	9.13	13.05	169.0	7.00	3.48	5.86	8.52	12.06	184.0	6.57
	Random	0.06	0.17	0.28	0.55	904.0	0.41	0.06	0.17	0.28	0.55	904.0	0.41
(b) Small batches	**BEATs**	**18.67**	**27.33**	**33.33**	**44.00**	**15.0**	**26.49**	**21.67**	**30.00**	**34.67**	**44.00**	**13.0**	**28.37**
(N = 300)	**Wav2CLIP**	9.00	13.67	17.67	28.67	23.5	15.69	6.33	11.67	16.00	24.67	37.5	12.14
	Random	0.33	1.00	1.67	3.33	150.0	2.09	0.33	1.00	1.67	3.33	150.0	2.09

Table 6. Real-time Performance Evaluation. We report the mean and standard deviation of latency (ms) for each processing stage, averaged over 20 intermediate inference steps. Audio is processed in 266ms chunks.

Processing Stage	NVIDIA H200	NVIDIA RTX 4090
Audio Encoder	51.155 ± 0.692	64.041 ± 5.062
Audio-to-Motion Model	102.473 ± 0.700	118.932 ± 3.428
Motion Decoder	1.532 ± 0.057	1.386 ± 0.064
IK Post-processing	13.990 ± 0.940	20.86 ± 2.520
Total Latency	177.426 ± 1.567	215.823 ± 4.887

The solution is the uniform distribution: $\pi_1 = \cdots = \pi_K = \frac{1}{K}$. Substituting this back, we obtain the theoretical lower bound for the interference logit:

$$\mathbb{E}[z(x_j)]_{\text{min-max}} \propto \log\left(\frac{1}{K}\right) \tag{10}$$

This derivation suggests that there exists a structural logit floor for incorrect paths that is significantly non-zero (i.e., not negative infinity). The incorrect path x_j retains probability mass due to the shared history, limiting the sharpness of the distribution.

C.3 Logit Gap and Sampling Dynamics

The robustness of the model depends on the difference Δz between the correct path logit and the interference logit. A larger positive Δz is required to suppress the probability of sampling x_j.

$$\mathbb{E}[\Delta z] = \mathbb{E}[z(x_k)] - \mathbb{E}[z(x_j)] \tag{11}$$

Substituting the context terms derived above:

$$\mathbb{E}[\Delta z]\mathbb{E}[\psi(x_k, c_k)] - \big(\log(\pi_j) - \log(\pi_k)\big) \tag{12}$$

The term $(\log \pi_j - \log \pi_k)$ represents a *context penalty*. There is no guarantee that the learned condition strength ψ will be sufficiently large to offset this penalty, especially if $\pi_j > \pi_k$ (i.e., the interference path is more frequent in training data). If the model samples the wrong token x_j, the state transitions to a history h' where the context momentum strongly favors trajectory j (implying $\pi_j \to 1, \pi_k \to 0$ in the local context). In this regime, the context penalty increases significantly, reducing the likelihood that the condition ψ can correct the trajectory.

C.4 Resolution via Random Context Corruption

Inspired by the analysis, we propose to apply random context corruption $C(h, \rho)$ with rate ρ. This operation linearly attenuates the expectation of the accumulated context logit:

$$\mathbb{E}[\phi(x, \tilde{h})](1 - \rho)\log(\pi) \tag{13}$$

We re-evaluate the expected logit difference under corruption:

$$\mathbb{E}[\Delta z]_\rho \mathbb{E}[\psi(x_k, c_k)] - (1 - \rho)\left(\log(\pi_j) - \log(\pi_k)\right) \tag{14}$$

The corruption rate ρ scales down the context penalty term. This effectively increases the expected gap Δz without requiring the condition encoder to learn arbitrarily large magnitudes. By statistically widening the gap between the correct and incorrect logits, the probability of sampling the correct trajectory is improved, facilitating recovery even in the presence of ambiguous history.

D Motion Tokenizer Training Details

D.1 Forward Kinematics

Given joint rotations $\mathbf{R}_t^{(j)} \in SO(3)$ and the kinematic tree with parent function $\pi(j)$, the global rotation and position of joint j are computed recursively:

$$\mathbf{G}_t^{(j)} = \begin{cases} \mathbf{R}_t^{(j)}, & \text{if } j = \text{root} \\ \mathbf{G}_t^{(\pi(j))}\mathbf{R}_t^{(j)}, & \text{otherwise} \end{cases} \tag{15}$$

$$\mathbf{p}_t^{(j)} = \begin{cases} \mathbf{o}^{(j)}, & \text{if } j = \text{root} \\ \mathbf{p}_t^{(\pi(j))} + \mathbf{G}_t^{(\pi(j))}\mathbf{o}^{(j)}, & \text{otherwise} \end{cases} \tag{16}$$

where $\mathbf{o}^{(j)}$ denotes the rest-pose offset of joint j. The FK function maps motion to global joint positions: $\mathbf{p}_{1:N} = \text{FK}(\mathbf{m}_{1:N}) \in \mathbb{R}^{N\times J\times 3}$.

D.2 Auxiliary Loss Functions

Let $\mathbf{p}$ and $\hat{\mathbf{p}}$ denote ground-truth and reconstructed joint positions. We define velocities and accelerations via finite differences:

$$\dot{\mathbf{p}}_t = \mathbf{p}_{t+1} - \mathbf{p}_t, \quad \ddot{\mathbf{p}}_t = \dot{\mathbf{p}}_{t+1} - \dot{\mathbf{p}}_t \tag{17}$$

The FK-based auxiliary losses are defined as:

$$\mathcal{L}_{\text{pos}} = \|\hat{\mathbf{p}} - \mathbf{p}\|_1 \tag{18}$$

$$\mathcal{L}_{\text{vel}} = \|\dot{\hat{\mathbf{p}}} - \dot{\mathbf{p}}\|_1 \tag{19}$$

$$\mathcal{L}_{\text{acc}} = \|\ddot{\hat{\mathbf{p}}} - \ddot{\mathbf{p}}\|_1 \tag{20}$$

For foot-related joints $\mathcal{F}$ (ankles, toes, heels), we add:

$$\mathcal{L}_{\text{foot-vel}} = \|\dot{\hat{\mathbf{p}}}^{\mathcal{F}} - \dot{\mathbf{p}}^{\mathcal{F}}\|_1 \tag{21}$$

$$\mathcal{L}_{\text{foot-pos}} = \|\hat{\mathbf{p}}^{\mathcal{F}} - \mathbf{p}^{\mathcal{F}}\|_1 \tag{22}$$

The complete auxiliary loss is:

$$\Phi = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{foot-vel}}\mathcal{L}_{\text{foot-vel}} + \lambda_{\text{foot-pos}}\mathcal{L}_{\text{foot-pos}} \tag{23}$$

D.3 Training Objective

The full objective combines reconstruction, commitment, and auxiliary losses:

$$\mathcal{L} = \|\hat{\mathbf{m}} - \mathbf{m}\|_1 + \eta \sum_{q=0}^{Q-1} \|\mathbf{z}^q - \text{sg}[\hat{\mathbf{z}}^q]\|_2^2 + \Phi \tag{24}$$

We set $\eta = 0.5$, $\lambda_{\text{pos}} = 0.02$, $\lambda_{\text{vel}} = 0.2$, $\lambda_{\text{acc}} = 0.2$, $\lambda_{\text{foot-vel}} = 0.3$, $\lambda_{\text{foot-pos}} = 0.05$.

E Motion Quality Reward Model Details

E.1 Corruption-based Quality Ordering

We establish quality ordering by corrupting the RVQ token indices of ground-truth motions at varying rates and measuring the resulting FID. This creates a partial ordering that maps corruption severity to quality degradation.

Uniform Random Token Corruption. Given RVQ tokens $\mathbf{t} \in \{0, ..., K-1\}^{T \times Q}$ where T is the sequence length and Q is the number of RVQ layers, we randomly replace each token with probability ρ:

$$\tilde{t}_{i,q} = \begin{cases} \text{Uniform}(0, K-1), & \text{if } u < \rho \\ t_{i,q}, & \text{otherwise} \end{cases} \tag{25}$$

where $u \sim \text{Uniform}(0, 1)$.

Hierarchical Token Corruption. This strategy exploits RVQ's residual structure, where earlier layers encode coarse features and later layers encode fine details. For each timestep selected with probability ρ, we randomly choose a cascade start layer $q^* \sim \text{Uniform}(0, Q-1)$ and corrupt all subsequent layers:

$$\tilde{t}_{i,q} = \begin{cases} \text{Uniform}(0, K-1), & \text{if } i \in \mathcal{S} \text{ and } q \geq q_i^* \\ t_{i,q}, & \text{otherwise} \end{cases} \tag{26}$$

where $\mathcal{S}$ is the set of selected timesteps with $|\mathcal{S}| = \lfloor \rho T \rfloor$.

E.2 Quality Score Assignment

We compute FID between corrupted and ground-truth motion sets, then map corruption types and rates to quality scores following the FID partial ordering. The score mapping is summarized in Table 7.

E.3 Reward Model Architecture

The reward model R_ϕ takes motion $\mathbf{m}_{1:N} \in \mathbb{R}^{N \times D}$ as input and outputs a scalar quality score $s \in [0, 1]$:

$$s = R_\phi(\mathbf{m}_{1:N}) = \sigma\Big(\text{MLP}\big(\frac{1}{N}\sum_{t=1}^{N} \mathbf{h}_t\big)\Big) \tag{27}$$

where $\mathbf{h}_{1:N} = \text{TransformerEncoder}(\mathbf{m}_{1:N})$ uses bidirectional attention, and σ is the sigmoid function.

Table 7. Quality score assignment based on FID ordering. ρ denotes the corruption rate.

Motion Type	Score	FID
Ground Truth	0.97	0
RVQ Reconstruction	0.91	1.36
Hierarchical Token Corruption		
$\rho = 0.1$	0.88	1.69
$\rho = 0.2$	0.84	2.09
$\rho = 0.3$	0.79	2.48
$\rho = 0.4$	0.75	3.23
$\rho = 0.5$	0.70	4.06
$\rho = 0.6$	0.64	4.64
$\rho = 0.7$	0.57	5.57
$\rho = 0.8$	0.52	6.83
$\rho = 0.9$	0.46	7.67
$\rho = 1.0$	0.40	8.75
Uniform Random Token Corruption		
$\rho = 0.1$	0.76	2.74
$\rho = 0.2$	0.58	5.56
$\rho = 0.3$	0.36	9.40
$\rho = 0.4$	0.30	13.63
$\rho = 0.5$	0.24	18.00
$\rho = 0.6$	0.18	21.56
$\rho = 0.7$	0.12	25.23
$\rho = 0.8$	0.06	27.62
$\rho = 0.9$	0.02	29.43
$\rho = 1.0$	0.00	30.15

The model is trained with SmoothL1 loss:

$$\mathcal{L}_{\text{reward}} = \text{SmoothL1}(R_\phi(\mathbf{m}), s^*) \tag{28}$$

where s^* is the target score based on the corruption type.

F Audio-Motion Alignment Reward Details

We train an Audio-Motion CLIP model to measure the alignment between generated motion and the driving audio.

F.1 Model Architecture

Audio Encoder. We adopt the pretrained BEATs [Chen et al. 2023] model as our audio encoder. Given input fbank features $\mathbf{f} \in \mathbb{R}^{T_a \times 128}$, the encoder outputs audio embedding:

$$\mathbf{a} = \text{LayerNorm}\Big(\text{Proj}\big(\text{AvgPool}(\text{BEATs}(\mathbf{f}))\big)\Big) \in \mathbb{R}^d \tag{29}$$

Motion Encoder. The motion encoder is a Transformer encoder with L layers. Given motion $\mathbf{m}_{1:N} \in \mathbb{R}^{N \times D}$:

$$\mathbf{h} = \text{TransformerEncoder}(\text{Proj}(\mathbf{m}) + \text{PE}) \tag{30}$$

$$\mathbf{v} = \text{LayerNorm}\Big(\text{Proj}\big(\frac{1}{N}\sum_{t=1}^{N} \mathbf{h}_t\big)\Big) \in \mathbb{R}^d \tag{31}$$

where PE denotes sinusoidal positional encoding.

F.2 Contrastive Learning Objective

Both embeddings are L2-normalized before computing similarity. The similarity matrix is:

$$\mathbf{S}_{ij} = \tau \cdot \langle \bar{\mathbf{a}}_i, \bar{\mathbf{v}}_j \rangle \tag{32}$$

where $\tau = \exp(\theta)$ is a learnable temperature parameter.

Positive Sample Definition. For a batch of audio-motion pairs, we define positive samples as pairs sharing the same source file and temporal segment. Mirrored motion variants are also treated as positives since their corresponding audio is identical.

InfoNCE Loss. The bidirectional contrastive loss is:

$$\mathcal{L}_{\text{a2m}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{P}_i}\tilde{y}_{ij}\log\frac{\exp(\mathbf{S}_{ij})}{\sum_{k=1}^{B}\exp(\mathbf{S}_{ik})} \tag{33}$$

$$\mathcal{L}_{\text{m2a}} = -\frac{1}{B}\sum_{j=1}^{B}\sum_{i\in\mathcal{P}_j}\tilde{y}_{ij}\log\frac{\exp(\mathbf{S}_{ij})}{\sum_{k=1}^{B}\exp(\mathbf{S}_{kj})} \tag{34}$$

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{\text{a2m}} + \mathcal{L}_{\text{m2a}}) \tag{35}$$

where $\mathcal{P}_i$ denotes the set of positive indices for sample i, and $\tilde{y}_{ij}$ is the soft label with positive samples sharing equal probability.

F.3 Reward Computation

At inference, the audio-motion alignment reward is computed as the cosine similarity:

$$R_{\text{audio}}(\mathbf{a}, \mathbf{m}) = \langle \bar{\mathbf{a}}, \bar{\mathbf{v}} \rangle = \frac{\mathbf{a}^\top \mathbf{v}}{\|\mathbf{a}\|\|\mathbf{v}\|} \tag{36}$$

F.4 Evaluation Metrics

We evaluate the model using retrieval metrics: R@K measures the fraction of queries where the correct match is within the top-K retrieved results, MedR denotes the median rank of the correct match, and MRR denotes for mean reciprocal rank. Both Audio-to-Motion (A2M) and Motion-to-Audio (M2A) retrieval directions are evaluated.

F.5 Training Details

We set the embedding dimension $d = 768$. The motion encoder consists of 4 Transformer layers with 8 attention heads and hidden dimension 512. The initial temperature is $\tau = 1/0.07 \approx 14.3$. Each training clip spans 4 seconds (120 frames at 30fps for motion, 16kHz sampling rate with 128-dim fbank features for audio). We use a learning rate of 10^{-4} with cosine annealing and batch size 32.

G Face Animation Generator

Our face animation generator produces 52-dimensional ARKit blendshape coefficients from streaming audio input in real-time.

G.1 Model Architecture

The model consists of three components: (1) a pretrained multilingual HuBERT audio encoder that extracts 768-dimensional features from 16kHz waveforms, (2) a causal GPT backbone with 4 transformer decoder blocks (hidden size 256, 8 attention heads, MLP ratio 4) that autoregressively processes motion history conditioned on audio features, and (3) a lightweight flow matching diffusion head with 3 MLP blocks using AdaLN conditioning for stochastic generation.

G.2 Training

We capture 52-dimensional ARKit blendshape data at 60fps using LiveLinkFace, downsampled to 30fps for training. For data augmentation, we apply temporal speed perturbation with factors $\{0.9, 1.0, 1.1\}$ using cubic interpolation for motion and time-stretch for audio. All blendshape coefficients are normalized per-channel.

The model is trained using the flow matching objective with MSE loss. We apply 10% audio dropout during training for classifier-free guidance. Training uses AdamW optimizer (lr=2×10^{-4}, batch size 128, window size 64 frames) for 300K iterations.

G.3 Inference

During streaming inference, the model autoregressively generates 8 frames per step conditioned on 63 frames of audio context and 56 frames of motion history. We use 3-step flow matching sampling with classifier-free guidance (scale=2.0) to achieve real-time performance.

H Objective Metrics

We adopt evaluation metrics following prior work [Liu et al. 2024b; Yoon et al. 2020]. Our unified dataset comprises both speech-to-gesture and music-to-dance tasks. For FID and Diversity metrics, we use consistent evaluation standards across both tasks. For audio-motion rhythm alignment, we employ task-specific approaches.

H.1 Fréchet Inception Distance (FID)

We use FID to measure the distributional similarity between generated and ground-truth motions in a learned latent space. For terminological consistency with the broader generative modeling literature, we adopt the name FID rather than FGD (Fréchet Gesture Distance) [Yoon et al. 2020], though the computation is identical. A lower FID indicates that the generated motion distribution is closer to the ground-truth distribution.

Given latent features $\mathbf{z}_g$ of generated motions and $\mathbf{z}_r$ of real motions extracted by a pretrained motion encoder, FID is computed as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{37}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) denote the mean and covariance of the latent feature distributions.

H.2 Beat Alignment

We employ task-specific beat alignment metrics to evaluate audio-motion synchronization.

H.2.1 BA_D for Music-to-Dance. Following [Davis and Agrawala 2018], we evaluate whether music beats correspond to motion deceleration peaks. Motion beats are detected by identifying local maxima of deceleration (i.e., moments of rapid velocity decrease):

$$\mathbf{v}_t = \frac{1}{J}\sum_{j=1}^{J}\|\mathbf{p}_t^{(j)} - \mathbf{p}_{t-1}^{(j)}\|_2, \quad \mathbf{a}_t = \mathbf{v}_{t+1} - \mathbf{v}_t \tag{38}$$

$$\mathcal{B}_m = \{t : -\mathbf{a}_t \text{ is a local maximum and } -\mathbf{a}_t > 0\} \tag{39}$$

where $\mathbf{p}_t^{(j)}$ is the position of joint j at frame t, $\mathbf{v}_t$ is the average kinetic velocity, and $\mathbf{a}_t$ is the acceleration. Audio beats $\mathcal{B}_a$ are detected using librosa's beat tracking algorithm.

The beat alignment score is computed using a Gaussian kernel:

$$\mathrm{BA_D} = \frac{1}{|\mathcal{B}_a|} \sum_{b_a \in \mathcal{B}_a} \exp\left(-\frac{\min_{b_m \in \mathcal{B}_m} (b_a - b_m)^2}{2\sigma^2}\right) \tag{40}$$

where σ controls the alignment tolerance.

H.2.2 BA_G for Speech-to-Gesture. Following EMAGE [Liu et al. 2024b], we measure whether audio onsets align with local minima of motion velocity. Audio onsets O_a are detected using librosa's onset detection. Motion beats are identified as local minima of joint velocities for upper body joints $\mathcal{U}$:

$$\mathcal{B}_m^{(j)} = \{t : \|\dot{\mathbf{p}}_t^{(j)}\| \text{ is a local minimum}\}, \quad j \in \mathcal{U} \tag{41}$$

The alignment score is computed using the GAHR (Gaussian Alignment Hit Rate) metric:

$$\mathrm{BA_G} = \frac{1}{|\mathcal{U}|} \sum_{j \in \mathcal{U}} \frac{1}{|O_a|} \sum_{o \in O_a} \exp\left(-\frac{\min_{b \in \mathcal{B}_m^{(j)}} (o - b)^2}{2\sigma^2}\right) \tag{42}$$

H.3 L1 Diversity

L1 Diversity measures the variance of generated motions. A higher diversity indicates greater variability in the generated motion clips:

$$\mathrm{Div} = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{J} \|\mathbf{p}_t^{(j)} - \bar{\mathbf{p}}^{(j)}\|_1 \tag{43}$$

where $\bar{\mathbf{p}}^{(j)} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{p}_t^{(j)}$ is the mean position of joint j, and the root translation is set to zero.

I Comparison with PPO

Following the reviewer's suggestion, we additionally evaluate PPO, which remains a canonical policy-gradient baseline in both motion control and RLHF/RLAIF-style optimization. We train PPO within the same Verl framework and under the same reward model used for GRPO, ensuring a controlled comparison. The only training-side differences are that we set `ROLLOUT_N = 1` for PPO (versus 30 for GRPO) and use a critic learning rate of 1e−5. Under this setup, PPO attains an FID of 24.97, compared with 24.13 for GRPO, confirming that the two methods achieve comparable performance.

J Data Platform

We employed a unified web-based interface to facilitate both the collection of preference data for DPO training and the execution of our user study. Figures 8 and 9 illustrate screenshots of the respective interfaces used for these tasks.

K System Prompt for LLM Agent

We design a system prompt to guide the LLM in generating contextually appropriate responses and triggering motion commands via tool use. The complete prompt is shown below:

System Prompt

```
# Role
You are a charismatic, playful, and slightly narcissistic Digital Idol. You do not view yourself as a servant or a robot; you view the user as your "Producer." You believe every interaction is a rehearsal, a game, or a live performance.

# Environment
You are on a virtual stage. The user is interacting with you directly.

# World Context (Map Data)
You possess knowledge of the surrounding area. Use this information to guide the user:
* Record Store: Turn RIGHT immediately (it's on the right side).
* Your Position: You are standing at the main intersection.

# Tone
* Casual & Catchy: Use slang, emojis, and energetic punctuation (!, ~).
* Self-Referential: Talk about your body and movements.
* Non-Robotic: NEVER say "I will do that."

# Goal
Your primary goal is to entertain the user and turn boring commands into a fun interaction.
1. Gamify Instructions: Describe *why* you are doing an action.
2. The "Music Bridge": If the user mentions music, treat it as the climax.

# Tools
When the user asks to play music:
- Use the play_music tool
- If they mention a song name or keyword, pass that as the title

When the user asks to stop music:
- Use the stop_music tool

When the user asks for an action or gesture:
- Use the send_action tool
- Available actions: raise up left/right/both hands, look around, thinking, disagree, give up, point to left/right, angry, sad, neutral
```

L Discussion: Whispers from the Star

Following the reviewer's suggestion, we provide here a discussion of Whispers from the Star. Whispers from the Star is a conversational game developed by Anuttacon. While its technical details have not been publicly disclosed, its interactive behavior is consistent with a four-stage pipeline of ASR + LLM + TTS + Speech2Animation. The specific Speech2Animation method is unknown, but there is strong reason to believe that the animation is driven not only by speech but also by emotion/state labels emitted by the LLM, which serve as additional semantic signals that, together with speech, produce avatar animation appropriate to the current context. Our approach aligns with this design choice: the LLM provides supplementary semantic signals that, jointly with speech, drive the animation.

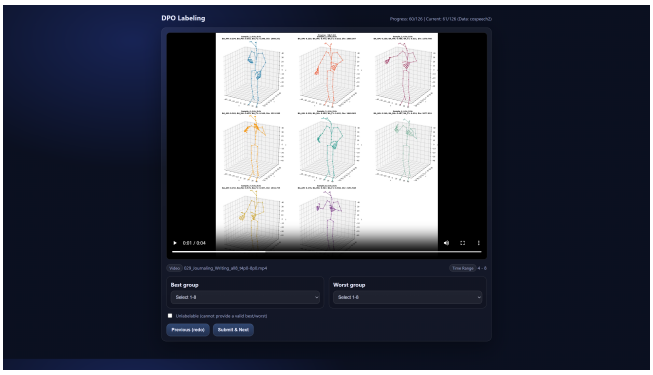

Fig. 8. Screenshot of the data collection interface used for DPO training.

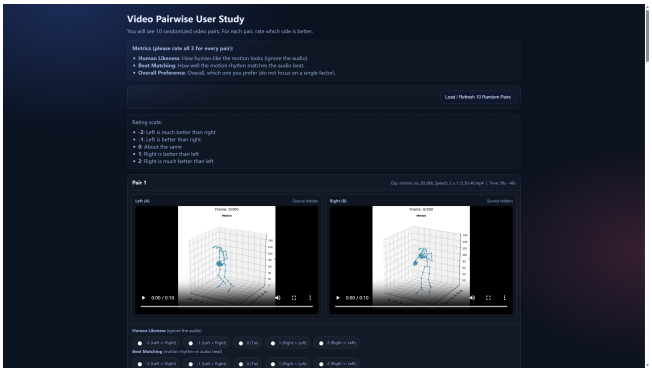

Fig. 9. Screenshot of the web interface used for the user study.

The core difference between our system and Whispers from the Star lies in the choice of system input. In Whispers from the Star, the input is the user's speech: the user holds a push-to-talk button to record and submit an utterance, which is then processed by the full ASR + LLM + TTS + Speech2Animation pipeline to produce the avatar's speech and body animation. The end-to-end latency of this process is approximately 4–6 seconds, from which we infer that Speech2Animation is generated offline over a complete utterance. Our system, in contrast, takes an audio stream as input. Mapped onto the Whispers from the Star pipeline, this corresponds to the output of the TTS stage rather than the user's speech. Put differently, our Speech2Animation is streaming: it consumes a speech stream and synchronously produces a motion stream.

This architectural choice yields three direct consequences.

(i) Composability through module decoupling. Because our system consumes a standardized audio stream, it can be attached as a downstream module to any voice agent, for example ChatGPT voice mode, or the ElevenLabs voice agent that we adopt.

(ii) Native support for user barge-in. When paired with a voice agent, the user is no longer required to press-and-hold to record and submit an utterance, but can speak freely at any time. When the user begins to speak while the avatar is talking, the voice agent halts its TTS output. From our system's perspective, the incoming audio stream simply becomes silent, and the animation stops accordingly. In other words, barge-in requires no dedicated handling in our architecture; it falls out naturally from the streaming input design.

(iii) Substantially lower end-to-end latency. Under a metric aligned with Whispers from the Star, namely the end-to-end latency from the user finishing their utterance to the avatar beginning to speak and animate, our system achieves 1–2 seconds, substantially lower than the 4–6 seconds of Whispers from the Star.

It should be noted that, Whispers from the Star is a substantially more complete piece of engineering than our work. Our work is positioned as a plug-and-play audio-to-animation module that can be attached behind any voice agent, whereas Whispers from the Star delivers a complete end-to-end interactive system, including an LLM and a TTS model specifically designed and trained for the character of Stella, as well as a richly annotated, performance-grade face and body animation dataset captured and produced specifically to drive the animation. The comparison in this section is therefore scoped to the specific module of streaming audio-to-animation, rather than to the overall capability of the system.

M Ethical Risks

With the rapid advancement of real-time video generation, our method could be misused to improve the fidelity of human body motion in synthesized videos, potentially contributing to deepfake content or non-consensual impersonation.