# ARMS: Anchor–Relational Motion Streaming for Seamless Solo-Social Motion Transitions

Huakun Liu[1], Qing Yu[2], Kent Fujiwara[2], Hideaki Uchiyama[1], and Kiyoshi Kiyokawa[1]

[1] Nara Institute of Science and Technology, Ikoma, Japan
[2] LY Corporation, Tokyo, Japan
{liu.huakun.li0,hideaki.uchiyama,kiyo}@is.naist.jp,
{yu.qing,kent.fujiwara}@lycorp.co.jp

**Abstract.** Generating temporally continuous and socially coherent human motion from text remains a fundamental challenge, particularly in realistic streams where people act alone, enter interactions, and later disengage. Most existing methods generate fixed-length motion clips under static agent configurations, which makes them brittle to solo–social transitions and unsuitable for incremental generation over long horizons. We propose ARMS, an Anchor–Relational Motion Streaming framework that unifies solo motion and human–human interaction within a single causal generative process. ARMS introduces a dynamics-asymmetric representation that decouples per-person temporal evolution from inter-person alignment via a partner-referenced relative-translation term, enabling seamless switching of social coupling without sacrificing long-horizon stability or spatial consistency between agents. On top of a causal latent space, a causal relational diffusion model progressively refines motion segment by segment using only past context, capturing both intra-person temporal dependencies and inter-person relations. Mode-aware relational gating activates or masks cross-agent connections, allowing the same model to support both solo and interaction generation. Experiments show that ARMS improves transition smoothness and social coherence compared to interaction-centric baselines, while also achieving competitive results on human–human interaction benchmarks.

**Keywords:** Human motion generation · Human–human interaction · Autoregressive models

## 1 Introduction

Text-driven human motion generation aims to synthesize human motion trajectories from natural language descriptions. Most existing models treat motions as short and self-contained clips rather than continuous and socially coherent behaviors [36, 41]. However, in realistic settings, individuals may act alone, engage with others, or transition between solo and interactive behaviors [5, 8, 20]. The ability to generate such dynamic long-horizon behaviors is essential for applications including animation, virtual agents, robotics, and immersive environments.

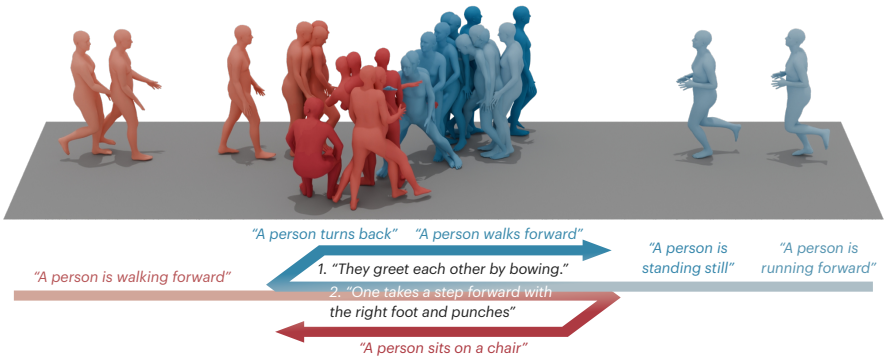


**Fig. 1:** ARMS generates long-horizon motion streams from evolving textual instructions, enabling seamless transitions between solo behavior and human–human interaction within a single generative process. In this example, two individuals approach each other, engage in multiple interactions (e.g., greeting and fighting), and then disengage to continue with independent actions, without resetting or reinitializing generation.

Despite recent progress in human–human interaction generation, existing methods are typically designed to produce fixed-length motion clips in a single pass [41]. This design conflicts with realistic social behavior, where interactions emerge, evolve, and dissolve without fixed temporal boundaries. Moreover, most approaches are tailored to a fixed agent configuration, modeling either solo motion or interaction using separate frameworks [5,8,20]. As a result, they struggle to support solo–social transitions in a temporally incremental setting, often causing boundary discontinuities or relational drift over long horizons.

Generating socially coherent motion in an incremental setting is challenging because the model must preserve temporal coherence for each individual while maintaining consistent inter-person relationships over extended horizons [6, 19]. Supporting streams that evolve between solo and two-person interaction further requires the model to accommodate changes in active agents and social coupling within a unified generative process.

To address these challenges, we propose **ARMS** (**A**nchor–**R**elational **M**otion **S**treaming), a causal autoregressive diffusion framework for temporally incremental human motion and interaction synthesis under changing one-/two-person social configurations. ARMS is built on a novel *dynamics-asymmetric representation* that unifies solo and interactive motions. The key is a partner-referenced relative-translation term that controls when inter-person coupling is represented. The term is inactive for solo motion and for the Anchor stream, but instantiated as the partner-relative displacement for the Relational stream during interaction. This design enables unbounded causal extension while avoiding global-coordinate drift, and preserves stable inter-person geometry when interactions emerge.

To enable streaming generation, motion sequences are first encoded into a temporally compressed latent space. Crucially, the same causal temporal autoencoder encodes all individual motions, both in solo scenes and as components of

interactions, under the proposed Anchor–Relational representation. Building on a causal autoregressive diffusion backbone, we introduce *mode-aware relational gating masks* that realize appropriate causal relational attention in both solo and interaction modes. Dynamic history conditioning further improves long-horizon consistency by attending to variable-length past context. Our contributions are:

- We formulate temporally incremental motion synthesis under *one-/two-person social configuration changes*, unifying solo behavior and human–human interaction within a single streaming process that supports seamless solo–social transitions.
- We propose a *dynamics-asymmetric* Anchor–Relational motion representation with a partner-referenced relative-translation term, enabling configuration-agnostic learning from both solo and interaction data while using a shared causal temporal autoencoder for individual motion encoding.
- Building on an autoregressive diffusion backbone, we instantiate the denoiser as a causal relational transformer and introduce *mode-aware relational gating masks* with dynamic history conditioning, enabling incremental generation while preserving temporal coherence and stable inter-person geometry.

## 2 Related Work

Human motion generation has been widely studied under various conditioning modalities, including action labels [15, 30], audio signals [3, 9, 12, 18, 21], scene context [8, 20, 25, 43, 54], and natural language [11, 13, 14, 17, 31, 44, 56]. We focus on text conditioning for its flexibility and accessibility [8, 27, 56].

**Human–Human Interaction Generation.** Early motion generation primarily focused on single-person dynamics without social interactions. Building upon a pretrained single-person diffusion model, ComMDM [38] extends it to two-person generation by learning a lightweight communication module, while RIG [42] and InterGen [19] introduce cooperative diffusion denoisers with two shared-weight Transformer branches. To enhance interaction diversity, in2IN [33] and MixerMDM [34] augment interaction modeling with additional single-person motion data, while Text2Interact [47] introduces a scalable synthesis-by-composition pipeline with fine-grained word-level conditioning. TIMotion [46] explicitly structures interaction dynamics along a temporal sequence to better capture evolving inter-person dependencies. Beyond continuous diffusion, InterMask [16] formulates interactions in a discrete 2D VQ space with masked modeling, and SocialGen [50] aligns motion and language tokens to extend interaction modeling toward multi-human social scenarios.

Despite these advances, many interaction models remain optimized for fixed-length synthesis from a static prompt, and do not support incremental extension. Moreover, most interaction methods assume a fixed two-person setting and do not address solo–social mode switching in evolving streams, where agents may approach, interact, and disengage without restarting generation. This leaves open the problem of unified streaming motion synthesis that preserves both per-person temporal coherence and stable inter-person geometry across transitions.

**Long-horizon Motion Generation.** Recent works extend motion generation beyond isolated short clips toward long-horizon or sequential synthesis [4,32,43, 48,52,53,57]. Several methods perform autoregressive diffusion for real-time control, generating motion incrementally from past states, including CAMDM [10], A-MDM [39], and PRIMAL [52]. CLoSD [43] and DART [53] formulate motion as autoregressive segments, while InfiniDreamer [57] explores arbitrarily long motion synthesis via segment refinement and score distillation. MotionStreamer [48] operates in a continuous causal latent space with a temporally causal autoencoder and strictly causal masking. Most long-horizon autoregressive methods are developed for single-person motion generation [13]. Recent interaction extensions such as Interact2Ar [35] and HINT [22] extend diffusion-based interaction models to the autoregressive setting via sliding-window generation. Interact2Ar incorporates a memory mechanism to reuse past motion features, while HINT adopts hierarchical conditioning in a canonicalized latent space, which can introduce discontinuities across segments.

Overall, existing approaches either generate human–human interactions within fixed-length clips, or extend motion synthesis to long horizons primarily for single-person scenarios. Autoregressive interaction extensions typically still assume a fixed agent configuration and focus on maintaining continuity within that regime. Consequently, streaming generation under changing one-/two-person social configurations remains largely unexplored.

**Interactive Motion Representation.** Motion representation is critical for both motion quality and long-horizon stability in human motion generation. Most single-person motion generation methods adopt canonicalized representations derived from HumanML3D [13], expressing motion in a root-centered coordinate system with frame-to-frame relative updates. Such incremental formulations simplify learning and naturally support temporally extensible generation [43,48]. Several interaction models extend this paradigm by introducing initial relative transformations between agents [38,50], but accumulated incremental updates can introduce spatial drift that disrupts inter-person alignment [19]. To avoid this issue, InterGen [19] models motion in global coordinates to explicitly preserve spatial relations between agents, which improves alignment but constrains motion generation to the coordinate distributions observed during training and limits their compatibility with open-ended or temporally extensible synthesis. HINT [22] mitigates drift by generating short motion segments in a canonicalized latent space, but segment-wise canonicalization introduces discontinuities that cause jittery transitions. DuetGen [12] instead represents two-person motion as a unified long feature sequence that jointly encodes canonicalized motion and relational features, but is limited to fixed two-person scenarios and increases dimensionality.

These designs highlight a fundamental tension in interaction representation: maintaining stable inter-person spatial relations while enabling temporally extensible generation. A shared representation that supports causal encoding and autoregressive rollout, preserves interaction geometry, and enables seamless solo–social transitions within a unified streaming process remains underexplored.

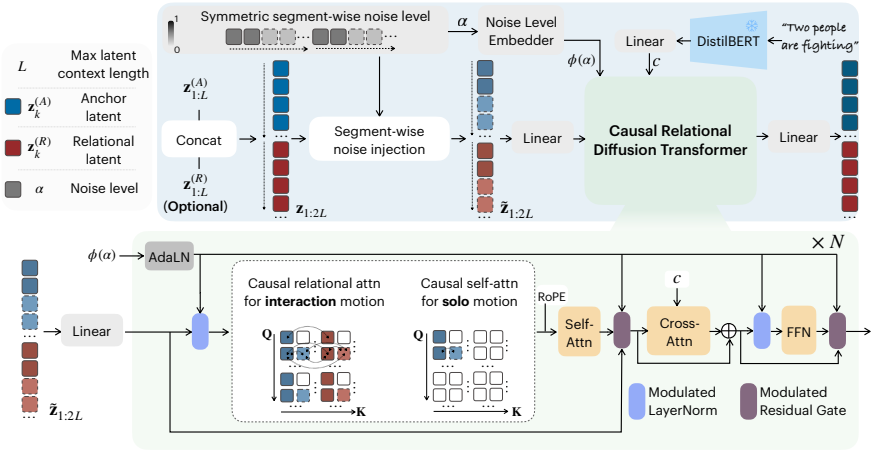


**Fig. 2:** We encode motion into causal latent streams for the Anchor and Relational agents, concatenate them, and apply symmetric segment-wise noising. A text-conditioned Causal Relational Diffusion Transformer denoises the sequence using segment-wise causal attention and mode-aware relational gating masks, enabling cross-agent visibility for interaction and masking it for solo motion.

## 3 Method

### 3.1 Problem Formulation

As shown in Fig. 1, we address the task of streaming motion generation under changing one-/two-person social configurations, where textual instructions arrive incrementally, with one text remaining active until replaced by a new one. Let $\mathcal{T} = \{\tau_1, \tau_2, \dots\}$ denote the ordered instruction stream, and $\beta(t)$ the index of the instruction active at timestep $t$, inducing change points that partition the timeline into variable-length intervals. At each timestep $t$, the model generates the motion states of all active agents:

$$\{\mathbf{x}_t^{(i)}\}_{i=1}^{N_t}, \tag{1}$$

where $N_t \in \{1, 2\}$ denotes the number of active agents at time $t$. We denote the generated motion history before $t$ as

$$\mathcal{H}_{<t} = \{\mathbf{x}_{t'}^{(i)} \mid t' < t, ; i = 1, \dots, N_{t'}\}, \tag{2}$$

where the number of active agents in the history may vary over time. The objective is therefore to learn a causal generative model that produces motion incrementally,

$$p\big(\{\mathbf{x}_t^{(i)}\}_{i=1}^{N_t} \mid \mathcal{H}_{<t}, \tau_{\beta(t)}, N_t\big), \tag{3}$$

while preserving long-horizon temporal coherence within each individual and maintaining relational consistency whenever interactions are present.

### 3.2 Motion Representation with Asymmetric Dynamics

Streaming motion generation commonly adopts canonicalized incremental features that enable temporally extensible synthesis, but these representations can accumulate spatial drift over long horizons. In contrast, interaction models often operate in global coordinates to preserve inter-person alignment, at the cost of restricting generation to coordinate distributions observed during training. To balance these trade-offs, inspired by [12, 19], we adopt a dynamics-asymmetric representation that separates incremental temporal dynamics from relational dynamics within a shared motion state. The key is a partner-referenced relative-translation term that controls when inter-person coupling is represented: it is set to zero for solo motion (and for the Anchor stream), and instantiated as the relative displacement to the partner for the Relational stream during interaction. This design preserves temporally extensible causal rollout while maintaining stable inter-person spatial geometry whenever interactions are present.

**Per-frame state definition.** We adopt a coordinate system with axes defined as $x$-left, $y$-up, and $z$-forward. At each timestep $t$, the motion state of an agent is defined as

$$\mathbf{x}_t = \left[\mathbf{r}_t^{yaw}, \mathbf{v}_t^{xz}, \mathbf{\Delta}_t^{xz}, \mathbf{p}_t^{local}, \dot{\mathbf{p}}_t^{local}, \mathbf{q}_t, \mathbf{c}_t\right], \tag{4}$$

where $\mathbf{r}_t^{yaw} = [\sin(\theta_t), \cos(\theta_t)] \in \mathbb{R}^2$ encodes the root yaw orientation to stabilize long-horizon integration and avoid angular discontinuities. The root linear velocity $\mathbf{v}_t^{xz} \in \mathbb{R}^2$ and the relational displacement $\mathbf{\Delta}_t^{xz} \in \mathbb{R}^2$ are defined on the world-frame $xz$-plane. Joint positions $\mathbf{p}_t^{local} \in \mathbb{R}^{J\times 3}$ and velocities $\dot{\mathbf{p}}_t^{local} \in \mathbb{R}^{J\times 3}$ are expressed in the root-local frame. Joint rotations $\mathbf{q}_t \in \mathbb{R}^{J\times 6}$ adopt the continuous 6D rotation representation [55] for all $J$ non-root joints. Foot-contact indicators $\mathbf{c}_t \in \mathbb{R}^4$ encode binary contact states for four predefined foot joints.

**Asymmetric motion dynamics.** Although both agents share the same state formulation, the two motion dynamics are instantiated asymmetrically during generation. We designate one agent as the *Anchor Agent (A)*, which models canonical temporal dynamics, and the other as the *Relational Agent (R)*, which explicitly encodes relational alignment. For the Anchor branch, the relational displacement is fixed:

$$\mathbf{\Delta}_t^{(A)} = \mathbf{0}, \tag{5}$$

which ensures that the Anchor motion is modeled independently of inter-agent spatial relations and remains compatible with solo motion generation. The global root position $\mathbf{P}_t^{(A)}$ is obtained via integration of $\mathbf{v}_t^{xz}$ over time. The Relational branch encodes interaction dynamics. It stores the root translation offset $\mathbf{\Delta}_t^{(R)}$ with respect to the Anchor agent. Its global root position is reconstructed as

$$\mathbf{P}_t^{(R)} = \mathbf{P}_t^{(A)} + \mathbf{\Delta}_t^{(R)}. \tag{6}$$

Restricting global integration to a single branch prevents independent integration errors that cause drifting global trajectories and unstable inter-agent alignment. Both agents retain identical state space, which enables architectural symmetry and parameter sharing. Notably, when the relational displacement is inactive, the

Anchor branch alone reduces to canonical incremental motion modeling, making the formulation naturally compatible with solo motion generation.

### 3.3 Causal Temporal AutoEncoder

Long-horizon autoregressive generation in high-dimensional motion space is computationally costly and prone to drift. Following the causal compression paradigm introduced in [48], we adopt a *causal variational autoencoder* to compress motion into a temporally causal latent space, where each latent depends only on past frames and supports incremental decoding during streaming generation. The model is a temporal variational autoencoder composed of 1D causal convolution and residual blocks in both encoder and decoder. Given a motion sequence $\mathbf{x}_{t\,t=1}^{T}$ for a single agent, the encoder produces temporally downsampled Gaussian parameters $\boldsymbol{\mu}_k, \boldsymbol{\sigma}^2_{k\,k=1}^{T/l}$, where $l$ is the temporal downsampling factor. Latent variables $\mathbf{z}_k \in \mathbb{R}^d$ are sampled via reparameterization, forming a continuous causal latent sequence $\mathbf{z}_{1:T/l}$. For latent timestep $k$, we denote by $\mathbf{z}_{\leq k}$ the causal latent prefix available up to that timestep. The decoder reconstructs motion from the causal latent prefix $\mathbf{z}_{\leq k}$.

For interaction modeling, the same autoencoder is applied independently to each agent branch with shared weights. The motion of each individual is encoded identically whether it appears in a solo sequence or as part of an interaction under our Anchor–Relational representation. This yields a unified latent space while preserving independent temporal encoding for different agents.

### 3.4 Causal Relational Diffusion

Let $\mathbf{z}_k^{(A)} \in \mathbb{R}^d$ and $\mathbf{z}_k^{(R)} \in \mathbb{R}^d$ denote the causal latents at latent timestep $k$ for the Anchor and Relational agent, respectively. For a window of length $L$ latent steps per agent, where $L$ denotes the maximum context size, we form a single ordered latent stream by concatenating the two agent streams into a single sequence:

$$\mathbf{z}_{1:2L} = \left[\mathbf{z}_{1:L}^{(A)}\,;\,\mathbf{z}_{1:L}^{(R)}\right]. \tag{7}$$

**Segment-wise causal noising in latent space.** Following diffusion forcing [7], we adopt a temporally structured noise schedule such that earlier latent segments are assigned lower noise levels than later ones. This creates a progressive certainty gradient along the temporal axis, encouraging committed refinement of past frames while preserving flexibility for future ones. Instead of assigning independent noise levels for every latent timestep, we partition the latent timeline into contiguous non-overlapping segments of size $S$, where $S$ denotes the number of latent steps in each segment. All latent steps within the same segment share a common noise level. Let $s(k) = \lceil k/S \rceil$ denote the segment index for timestep $k$, and let $\alpha_s \in [0, 1]$ denote the noise level associated with segment $s$. The noised latent at timestep $k$ is constructed as

$$\tilde{\mathbf{z}}_k = \alpha_{s(k)}\,\mathbf{z}_k + (1 - \alpha_{s(k)})\,\boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{8}$$

Segment-wise noise sharing enforces joint refinement over short temporal segments rather than isolated latent steps. This increases the effective temporal context at each diffusion stage and improves local motion consistency. For the two-agent latent stream $\mathbf{z}_{1:2L}$, segments corresponding to the same timestep of each agent share identical noise levels. This symmetric assignment ensures that the Anchor and Relational branches are refined at the same temporal stage, which stabilizes interaction dynamics during diffusion.

**Relational diffusion denoiser.** The denoiser is built upon the causal diffusion architecture proposed in [51]. As shown in Fig. 2, the noised latent stream $\tilde{\mathbf{z}}_{1:2L}$ is processed by a stack of transformer layers that jointly model temporal dynamics and inter-agent relations under a segment-wise causal attention mechanism. Noise levels are first embedded and injected into each transformer layer through timestep-conditioned adaptive layer normalization (AdaLN) [29]. Self-attention is then applied with rotary positional encoding (RoPE) [40] along the temporal dimension to preserve ordering under incremental extension. To enforce temporal causality and relational reasoning, we apply a segment-wise causal relational attention mask to the self-attention logits.

Let $q, r \in \{1, \ldots, 2L\}$ index query and key positions in the concatenated latent stream. For any position $u$, we denote by $a(u) \in \{A, R\}$ its agent identity and by $k(u) \in \{1, \ldots, L\}$ its per-agent latent timestep. Timesteps are grouped into segments of size $S$ with segment index

$$s(u) = \left\lceil \frac{k(u)}{S} \right\rceil . \tag{9}$$

The relational attention mask is defined as

$$M_{qr} = \begin{cases} 0, & s(r) \le s(q) \ \wedge \ \mathit{\Gamma}\big(a(q), a(r)\big) = 1, \\ -\infty, & \text{otherwise}. \end{cases} \tag{10}$$

This formulation ensures that each query attends only to its own, past segments, and optionally cross-agent segments, preventing future leakage. $\mathit{\Gamma}(\cdot, \cdot)$ is a mode gate that controls cross-agent visibility and is defined as

$$\mathit{\Gamma}(a(q), a(r)) = \begin{cases} 1, & \text{interaction mode}, \\ 1[a(q) = a(r) = A], & \text{solo mode}. \end{cases} \tag{11}$$

In solo mode, only the Anchor tokens remain visible, while Relational tokens are ignored. After the self-attention, text instructions are encoded using a frozen DistilBERT encoder [37], producing token embeddings that are linearly projected to the model dimension and incorporated via cross-attention in every transformer layer. This design allows latent vectors to condition on the active textual instruction while maintaining temporal causality in the motion stream. The overall denoiser predicts velocity-like residuals [24] for all latent steps jointly:

$$\mathbf{v}_\theta = \mathcal{F}_\theta(\tilde{\mathbf{z}}_{1:2L}, \phi(s), \mathbf{c}, M) , \tag{12}$$

where $\phi(s)$ denotes embedded segment noise levels, $\mathbf{c}$ the text features, and $M$ the mode-aware relational gating mask.

**Inference.** At inference time, we generate motions by progressively refining latent segments under a temporally offset noise schedule [7, 51]. Let $K$ denote the number of diffusion refinement iterations and $\delta$ the refinement offset between adjacent segments. Segment $i+1$ begins denoising $\delta$ refinement steps after segment $i$. Under this schedule, earlier segments become deterministic sooner while later segments remain partially noisy and continue to be refined. This overlapping denoising avoids hard segment boundaries and improves efficiency compared to strictly sequential autoregressive refinement. For continuous generation, we adopt variable-length history conditioning during inference. Under the same prompt, the full context window $L$ is preserved to maintain long-range temporal coherence. When the textual instruction changes, we retain only a short history $H$ of previously generated latents while replacing the conditioning text embedding $\mathbf{c}$ for future segments. This reduced history provides sufficient motion context for transition while allowing the model to adapt smoothly to the new instruction. Agent configurations can also change dynamically during inference. When generating solo motion, the Relational branch is masked through $\Gamma(\cdot, \cdot)$, reducing the model to a single-agent generator. When transitioning to interaction, the Relational branch is activated and initialized either from noise or from encoded relative dynamics derived from existing motion. Unless otherwise stated, we initialize the Relational branch from noise at interaction onset. Details of the exact inference procedure are provided in the supplementary material.

## 4 Experiments

### 4.1 Experimental Setup

**Dataset.** We use HumanML3D [13] for single-person motion and InterHuman [19] for two-person interactions. We adopt a 272-dimensional representation variant of the HumanML3D dataset used in [48] for better compatibility with InterHuman. It contains 26,846 motion sequences derived from the AMASS [23] dataset, each paired with three textual descriptions. InterHuman consists of 7,779 two-person interaction sequences with 23,337 textual descriptions. Since HumanML3D and InterHuman adopt different motion representations, we convert both datasets into our shared Anchor–Relational representation for training. This conversion is deterministic and invertible. For evaluation, generated motions are converted back to the native representation of each dataset before computing metrics, strictly following standard evaluation protocols. We train a unified model jointly on HumanML3D and InterHuman for cross-scenario streaming motion generation, while training dataset-specific models on InterHuman for fair comparison in interaction generation evaluation.

Following prior work [16], we additionally report results on InterX [49] to assess cross-skeleton generalization. InterX follows the SMPL-X [28] representation and contains 11,388 motion sequences, each paired with three textual descrip-

**Table 1: Quantitative evaluation** on the **InterHuman** test sets. Values are reported with 95% confidence intervals. The arrow → indicates that performance closer to ground truth is preferred. Best and second-best are **bold** and underlined.

| Method | R Precision↑ | | | FID↓ | MM Dist↓ | Diversity→ |
|---|---|---|---|---|---|---|
| | Top 1 | Top 2 | Top 3 | | | |
| Ground Truth | $0.452^{\pm.008}$ | $0.610^{\pm.009}$ | $0.701^{\pm.008}$ | $0.273^{\pm.007}$ | $3.755^{\pm.008}$ | $7.948^{\pm.064}$ |
| T2M [13] | $0.238^{\pm.012}$ | $0.325^{\pm.010}$ | $0.464^{\pm.014}$ | $13.769^{\pm.072}$ | $5.731^{\pm.013}$ | $7.046^{\pm.022}$ |
| MDM [45] | $0.153^{\pm.012}$ | $0.339^{\pm.012}$ | $0.339^{\pm.012}$ | $9.167^{\pm.056}$ | $7.125^{\pm.018}$ | $7.602^{\pm.045}$ |
| ComMDM [38] | $0.223^{\pm.009}$ | $0.334^{\pm.008}$ | $0.466^{\pm.010}$ | $7.069^{\pm.054}$ | $6.212^{\pm.021}$ | $7.244^{\pm.038}$ |
| RIG [42] | $0.285^{\pm.010}$ | $0.409^{\pm.014}$ | $0.521^{\pm.013}$ | $6.775^{\pm.069}$ | $4.876^{\pm.018}$ | $7.311^{\pm.043}$ |
| InterGen [19] | $0.371^{\pm.010}$ | $0.515^{\pm.012}$ | $0.624^{\pm.010}$ | $5.918^{\pm.079}$ | $5.108^{\pm.014}$ | $7.387^{\pm.029}$ |
| MoMat–MoGen [5] | $0.449^{\pm.004}$ | $0.591^{\pm.003}$ | $0.666^{\pm.004}$ | $5.674^{\pm.085}$ | $3.790^{\pm.001}$ | $8.021^{\pm.350}$ |
| in2IN [33] | $0.425^{\pm.008}$ | $0.576^{\pm.008}$ | $0.662^{\pm.009}$ | $5.535^{\pm.120}$ | $3.803^{\pm.002}$ | $7.953^{\pm.047}$ |
| InterMask [16] | $0.449^{\pm.004}$ | $0.599^{\pm.005}$ | $0.683^{\pm.004}$ | $5.154^{\pm.061}$ | $3.790^{\pm.002}$ | $\mathbf{7.944}^{\pm.033}$ |
| Text2Interact [47] | $0.483^{\pm.005}$ | $0.638^{\pm.005}$ | $0.717^{\pm.005}$ | $5.191^{\pm.055}$ | $3.778^{\pm.001}$ | $7.900^{\pm.030}$ |
| HINT [22] | $0.432^{\pm.004}$ | $0.587^{\pm.004}$ | $0.672^{\pm.004}$ | $\mathbf{3.100}^{\pm.035}$ | $3.796^{\pm.001}$ | $7.898^{\pm.023}$ |
| TIMotion [46] | $\underline{0.501}^{\pm.005}$ | $\underline{0.656}^{\pm.006}$ | $\underline{0.734}^{\pm.006}$ | $4.702^{\pm.069}$ | $\underline{3.769}^{\pm.001}$ | $\underline{7.943}^{\pm.034}$ |
| Ours (full-window generation) | $\mathbf{0.529}^{\pm.008}$ | $\mathbf{0.690}^{\pm.008}$ | $\mathbf{0.764}^{\pm.004}$ | $\underline{4}.436^{\pm.069}$ | $\mathbf{3.763}^{\pm.002}$ | $8.089^{\pm.059}$ |
| Ours (streaming generation) | $0.492^{\pm.004}$ | $0.647^{\pm.005}$ | $0.723^{\pm.005}$ | $4.444^{\pm.068}$ | $3.778^{\pm.001}$ | $8.017^{\pm.026}$ |

tions. To ensure compatibility with our model, we convert the InterX motions into the same anchor–relational representation used for training.

**Implementation details.** The causal temporal encoder downsamples the motion sequences by a factor of 4, producing latent tokens with dimension 64. The motion generation model is implemented as a causal relational diffusion transformer with 8 layers, a hidden dimension of 512, 4 attention heads, a maximum context length of $L = 75$, and a segment size of $S = 5$. Text conditions are extracted using a frozen DistilBERT encoder. We train the denoiser with a batch size of 64 and a maximum sequence length of 300 frames, i.e., 75 latent tokens, for 500 epochs. Optimization uses AdamW with learning rate $2 \times 10^{-4}$, betas $(0.9, 0.99)$, and weight decay $1 \times 10^{-5}$. We employ flow matching [1, 2, 24] as the ODE sampler in causal diffusion forcing. During inference, we perform $K = 50$ denoising steps, and the refinement offset $\delta$ is set to 5.

### 4.2 Quantitative Evaluation on Interaction Generation

We evaluate human–human interaction generation capability of ARMS under the standard benchmark setting. We follow evaluation protocols in text-to-motion generation using the pretrained text–motion retrieval evaluator adopted in prior work [16, 19]. We report: (1) R-Precision, the retrieval accuracy of the ground-truth motion given a text query; (2) MM Dist, the average embedding distance between generated motions and their paired captions; (3) FID, the Fréchet distance between feature distributions of generated and real motions; and (4) Diversity, the average pairwise distance among generated motions. All methods are evaluated under identical settings as prior work for fair comparison.

**Table 2: Quantitative evaluation** on the **InterX** test sets. Values are reported with 95% confidence intervals. The arrow → indicates that performance closer to ground truth is preferred. Best and second-best results are **bold** and underlined.

| Method | R Precision↑ | | | FID↓ | MM Dist↓ | Diversity→ |
|---|---|---|---|---|---|---|
| | Top 1 | Top 2 | Top 3 | | | |
| Ground Truth | $0.429^{\pm.004}$ | $0.626^{\pm.003}$ | $0.736^{\pm.003}$ | $0.002^{\pm.0002}$ | $3.536^{\pm.013}$ | $9.734^{\pm.078}$ |
| T2M [13] | $0.184^{\pm.010}$ | $0.298^{\pm.006}$ | $0.396^{\pm.005}$ | $5.481^{\pm.382}$ | $9.576^{\pm.006}$ | $2.771^{\pm.151}$ |
| MDM [45] | $0.203^{\pm.009}$ | $0.329^{\pm.007}$ | $0.426^{\pm.005}$ | $23.701^{\pm.057}$ | $9.548^{\pm.014}$ | $5.856^{\pm.077}$ |
| ComMDM [38] | $0.090^{\pm.002}$ | $0.165^{\pm.004}$ | $0.236^{\pm.004}$ | $29.266^{\pm.067}$ | $6.870^{\pm.017}$ | $4.734^{\pm.067}$ |
| InterGen [19] | $0.207^{\pm.004}$ | $0.335^{\pm.005}$ | $0.429^{\pm.005}$ | $5.207^{\pm.216}$ | $9.580^{\pm.011}$ | $7.788^{\pm.208}$ |
| InterMask [16] | $0.403^{\pm.005}$ | $0.595^{\pm.004}$ | $0.705^{\pm.005}$ | $0.399^{\pm.013}$ | $3.705^{\pm.017}$ | $9.046^{\pm.073}$ |
| TIMotion [46] | $0.412^{\pm.004}$ | $0.601^{\pm.004}$ | $0.714^{\pm.003}$ | $0.385^{\pm.022}$ | $3.706^{\pm.015}$ | $\underline{9.191}^{\pm.092}$ |
| Interact2Ar [35] | $\underline{0.441}^{\pm.00}$ | $\underline{0.631}^{\pm.00}$ | $\underline{0.737}^{\pm.00}$ | $\mathbf{0.148}^{\pm.01}$ | $\underline{3.581}^{\pm.01}$ | $9.147^{\pm.06}$ |
| HINT [22] | $0.386^{\pm.005}$ | $0.572^{\pm.004}$ | $0.682^{\pm.003}$ | $\underline{0.278}^{\pm.012}$ | $4.007^{\pm.016}$ | $8.886^{\pm.066}$ |
| Ours (streaming generation) | $\mathbf{0.479}^{\pm.005}$ | $\mathbf{0.679}^{\pm.005}$ | $\mathbf{0.780}^{\pm.003}$ | $0.279^{\pm.010}$ | $\mathbf{3.405}^{\pm.016}$ | $\mathbf{9.399}^{\pm.064}$ |

We evaluate our model under two inference regimes. *Full-window generation* performs diffusion over the entire motion window in a single denoising process. *Streaming generation* adopts our causal segment-wise refinement schedule and incrementally synthesizes motions. Table 1 reports quantitative results on the InterHuman test set. Our *full-window generation* variant achieves the best overall performance among all compared methods, indicating superior text–motion alignment and realism. In the *streaming generation* setting, ARMS maintains strong quality (FID 4.44, MM Dist 3.78), with lower R-Precision due to compounding uncertainty over long incremental rollouts. Importantly, streaming generation enables long-duration motion synthesis beyond fixed-length clips while maintaining comparable realism and alignment quality.

Table 2 presents results on InterX. Our streaming generation model outperforms other models on all metrics. Notably, the improved R precision and diversity demonstrates the strong text–motion alignment and rich motion variation. These results indicate that our formulation generalizes effectively across different skeletal definitions, without being specialized to a specific joint configuration.

### 4.3 Motion Streaming: Qualitative Evaluation

We present qualitative results on long-duration interaction synthesis and cross-scenario generalization. Fig. 3 compares our method with the adapted InterMask. In extended sequences, InterMask often exhibits unnatural spatial shifts during interaction transition. In contrast, our method achieves smooth transitions between sub-actions, maintains consistent relative positioning, and preserves semantically aligned interaction evolution over time, demonstrating improved stability in long-duration interaction synthesis. Fig. 4 and Fig. 5 further illustrate results of three cross-agent scenarios: (1) a solo motion transitions into a human–human interaction; (2) an interaction evolves into another; and (3) an interaction resolves back into a single-person motion. In all cases, motions

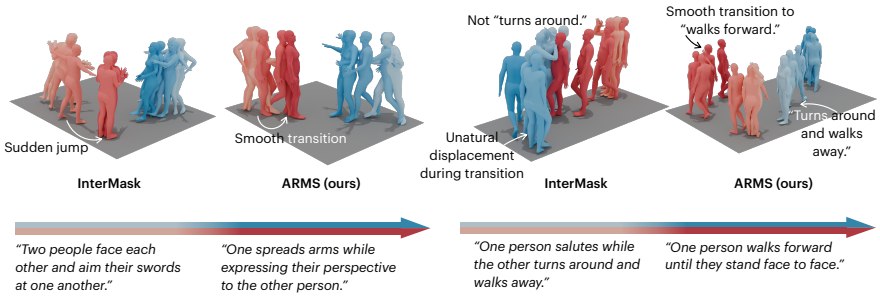


**Fig. 3:** Qualitative comparison of long-duration interaction synthesis. Compared to InterMask, our method generates smoother temporal transitions, avoids abrupt motion discontinuities, and preserves coherent interaction dynamics across extended sequences.

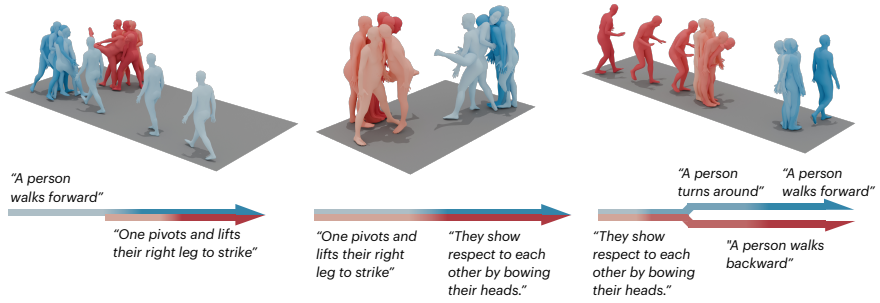


**Fig. 4:** Qualitative results of streaming motion generation. From left to right, we show: (1) solo motion transitioning into human–human interaction, (2) interaction evolving into a different interaction, and (3) interaction resolving back into solo motion. All sequences are generated in a streaming manner without resetting the latent state.

are generated continuously under incremental text updates. The model produces smooth transitions between different scenarios within a single framework, while maintaining temporal coherence and spatial consistency.

### 4.4 Motion Streaming: Quantitative Evaluation

We adopt the jerk-based metrics introduced in FlowMDM [4]. Jerk is defined as the time derivative of acceleration, which is sensitive to abrupt motion changes. Peak Jerk (PJ) measures the maximum instantaneous jerk magnitude across all joints, capturing extreme discontinuities, while Area Under the Jerk (AUJ) accumulates deviations from the average jerk over time, reflecting persistent smoothness irregularities. For transition evaluation, PJ and AUJ are computed within a 2-second temporal window centered at the behavior transition point. We evaluate two long-horizon streaming scenarios: (1) continuous interaction generation and (2) transitions between solo motion and interaction. Since no

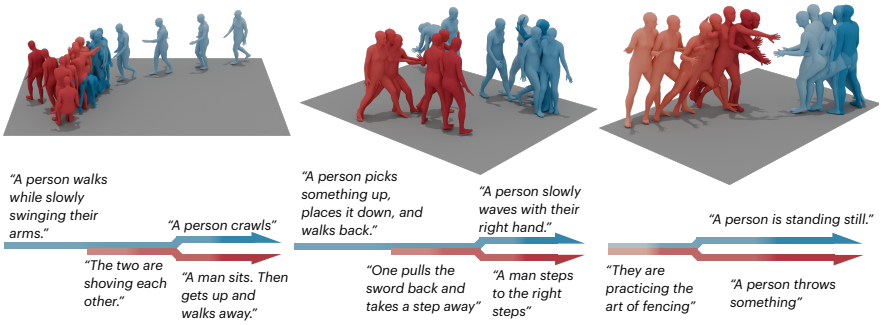


**Fig. 5:** Qualitative results of dynamic social motion streaming generation.

**Table 3:** Quantitative comparison on continuous multi-segment streaming generation.

| Method | Solo ↔ Interaction Transition | | Interaction ↔ Interaction Subsequence | | Interaction ↔ Interaction Transition | |
|---|---|---|---|---|---|---|
| | PJ → | AUJ↓ | R@Top3↑ | FID↓ | PJ → | AUJ↓ |
| Ground truth | $0.046^{\pm.000}$ | $0.672^{\pm.000}$ | $0.643^{\pm.015}$ | $8.344^{\pm.524}$ | $0.074^{\pm.000}$ | $0.584^{\pm.000}$ |
| InterMask (Inpainting) [16] | – | – | $\mathbf{0.631}^{\pm.012}$ | $\mathbf{28.554}^{\pm.329}$ | $0.269^{\pm.002}$ | $22.562^{\pm.057}$ |
| InterMask (Adapted) [16] | $3.131^{\pm.046}$ | $7.444^{\pm.093}$ | $0.557^{\pm.017}$ | $32.307^{\pm.562}$ | $0.265^{\pm.003}$ | $22.440^{\pm.043}$ |
| Ours (Streaming generation) | $\mathbf{0.077}^{\pm.001}$ | $\mathbf{1.070}^{\pm.005}$ | $0.509^{\pm.016}$ | $30.620^{\pm.561}$ | $\mathbf{0.071}^{\pm.001}$ | $\mathbf{2.925}^{\pm.010}$ |

standardized benchmark exists for these tasks, we construct evaluation sequences from the InterHuman and HumanML3D test sets.

For interaction streaming, we randomly sample 64 groups of interaction descriptions and generate sequences composed of eight consecutive interaction segments using streaming inference. To analyze both transition quality and semantic fidelity, we report jerk-based transition metrics together with standard motion generation metrics computed on the individual interaction subsequences. For comparison, we adapt the state-of-the-art method InterMask [16] to support history-conditioned generation by feeding previously generated frames as unmasked prefix tokens. As shown in Table 3, InterMask with inpainting achieves the best R@Top3 and FID on the interaction subsequences. However, we observe that this behavior arises because the model effectively starts a new motion conditioned on the next prompt rather than generating a smooth transition from the preceding motion. In contrast, our streaming model explicitly enforces temporal continuity across segments. As a result, the generated motion gradually adapts to satisfy the new textual condition, which slightly reduces semantic alignment on the isolated subsequences but produces substantially smoother long-horizon motion. In addition, the initial interaction state of each sequence in our generation is conditioned on the previously generated sequence, whereas InterMask initializes each segment from a newly generated starting pose that may exhibit foot sliding. These results show that our method significantly reduces motion

**Table 4:** Ablations on model design and inference hyperparameters.

| Methods | Interaction generation | | Transition | |
|---|---|---|---|---|
| | R@Top3 ↑ | FID↓ | PJ → | AUJ↓ |
| Ground truth | $0.701^{\pm.008}$ | $0.273^{\pm.007}$ | $0.074^{\pm.000}$ | $0.584^{\pm.000}$ |
| w/ InterGen representation | $0.723^{\pm.004}$ | $4.553^{\pm.065}$ | $0.130^{\pm.001}$ | $3.367^{\pm.022}$ |
| w/o autoregressive | $\mathbf{0.764}^{\pm.004}$ | $\mathbf{4.436}^{\pm.069}$ | $0.114^{\pm.006}$ | $3.209^{\pm.024}$ |
| Segment size $S = 1$ | $0.674^{\pm.005}$ | $6.832^{\pm0.115}$ | $0.131^{\pm.003}$ | $3.353^{\pm.079}$ |
| Segment size $S = 3$ | $0.718^{\pm.005}$ | $4.445^{\pm.052}$ | $0.070^{\pm.001}$ | $\mathbf{2.708}^{\pm.011}$ |
| Segment size $S = 10$ | $0.729^{\pm.005}$ | $4.600^{\pm.064}$ | $0.082^{\pm.001}$ | $2.980^{\pm.008}$ |
| Latent dim $d = 32$ | $0.707^{\pm.004}$ | $4.407^{\pm.080}$ | $0.115^{\pm.001}$ | $2.894^{\pm.015}$ |
| Latent dim $d = 128$ | $0.695^{\pm.005}$ | $4.969^{\pm.082}$ | $0.076^{\pm.001}$ | $2.873^{\pm.016}$ |
| Sampling timesteps $K = 10$ | $0.722^{\pm.005}$ | $4.731^{\pm.068}$ | $0.105^{\pm.001}$ | $3.064^{\pm.008}$ |
| Sampling timesteps $K = 20$ | $0.719^{\pm.005}$ | $4.500^{\pm.073}$ | $0.078^{\pm.001}$ | $2.975^{\pm.009}$ |
| Refinement offset $\delta = 1$ | $0.728^{\pm.005}$ | $4.537^{\pm.079}$ | $0.082^{\pm.001}$ | $2.890^{\pm.010}$ |
| Refinement offset $\delta = 3$ | $0.726^{\pm.005}$ | $4.486^{\pm.074}$ | $\mathbf{0.075}^{\pm.001}$ | $2.919^{\pm.011}$ |
| Refinement offset $\delta = 10$ | $0.730^{\pm.005}$ | $4.599^{\pm.064}$ | $0.069^{\pm.001}$ | $2.968^{\pm.010}$ |
| Refinement offset $\delta = 50$ | $0.662^{\pm.004}$ | $8.645^{\pm.110}$ | $0.065^{\pm.001}$ | $3.068^{\pm.013}$ |
| Ours | $0.723^{\pm.008}$ | $4.446^{\pm.116}$ | $0.071^{\pm.001}$ | $2.925^{\pm.010}$ |

discontinuities compared with the InterMask baseline, achieving lower PJ (0.071 vs. 0.265) and AUJ (2.925 vs. 22.440).

For solo–interaction transitions, we randomly construct 64 sequences consisting of a single-person motion, followed by an interaction motion, and another single-person motion. We adapt and retrain InterMask on HumanML3D and InterHuman datasets to support both solo and duo generation by masking the second agent before interaction onset and activating it during the duo phase. As shown in Table 3, the adapted InterMask baseline exhibits large discontinuities at the solo-to-interaction boundary, with PJ/AUJ of 3.131/7.444, while ARMS remains close to ground truth with 0.077/1.070. This gap reflects the difficulty of abruptly synthesizing a globally consistent two-person configuration after suppressing the second agent, whereas ARMS keeps the Anchor stream causally continuous and introduces the second agent through the Relational branch.

### 4.5 Ablation Study

We conduct ablation studies to analyze the contribution of key design components and the sensitivity to critical hyperparameters. Results are summarized in Table 4. Replacing our Anchor–Relational representation with InterGen's non-canonical formulation [19] slightly improves interaction alignment, but significantly degrades transition smoothness, confirming that globally anchored modeling weakens long-horizon stability. Removing the autoregressive streaming mechanism achieves the best single-window interaction metrics, yet increases transition jerk. Setting segment size $S = 1$, i.e., disabling joint latent refinement within

segments, substantially degrades interaction realism and increases jerk, indicating insufficient temporal coupling. Moderate segment sizes achieve the best balance, whereas larger segments slightly reduce transition smoothness, suggesting that coarse refinement weakens local adaptability. Varying latent dimension $d$ shows that smaller latent spaces (e.g., $d = 32$) improve interaction alignment but amplify jerk. Increasing diffusion sampling steps $K$ improves transition stability and realism, while a moderate inter-segment offset achieves the lowest jerk, validating the effectiveness of the temporal refinement schedule. Finally, the refinement offset $\delta$, which controls the temporal delay between denoising adjacent segments during streaming inference, achieves the best trade-off between interaction quality and transition smoothness at a moderate value.

## 5 Conclusion

In this work, we present ARMS, an Anchor–Relational causal diffusion framework for streaming solo–social motion generation. ARMS unifies solo motion and human–human interaction within a single incremental generative process through dynamics-asymmetric motion modeling, decoupling temporal progression from inter-person relational alignment. Our causal relational diffusion model supports temporal refinement, variable-length history conditioning, and seamless transitions across one-/two-person social configurations. Experiments demonstrate strong generation performance and substantially improved transition smoothness and social coherence in long-horizon streaming, suggesting a scalable pathway toward more flexible and realistic socially aware motion synthesis.

**Limitations.** Although ARMS unifies solo and human–human interaction generation within a single streaming framework, several limitations remain. First, the current formulation does not explicitly condition on surrounding agents or environmental context during solo motion generation, so crowd-level and scene-level collision avoidance are not directly addressed. Second, this work focuses on up to two-person interaction scenarios and assumes that social mode is activated within the interaction-distance range covered by the training data. If agents begin far outside this range, solo locomotion should first bring them into a plausible interaction range before activating social coupling. While the relational formulation can be extended to variable numbers of agents through pairwise modeling [22, 26], scaling to multi-agent settings requires explicit reasoning over higher-order social structures. Third, ARMS currently uses hard mode-aware relational gating to preserve strict causal consistency between solo and interaction modes. While effective, this binary switching may still introduce abrupt changes in challenging transition cases. Gradually blending relational attention across mode switches could further improve transition smoothness.

## Acknowledgements

This work was done during an internship at LY Corporation and was also supported by JST BOOST JPMJBS2423.

## References


1. Albergo, M., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. Journal of Machine Learning Research **26**(209), 1–80 (2025) 10
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023) 10
3. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023) 3
4. Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 457–469 (2024) 4, 12
5. Cai, Z., Jiang, J., Qing, Z., Guo, X., Zhang, M., Lin, Z., Mei, H., Wei, C., Wang, R., Yin, W., et al.: Digital life project: Autonomous 3d characters with social intelligence. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 582–592 (2024) 1, 2, 10
6. Cen, Z., Pi, H., Peng, S., Shuai, Q., Shen, Y., Bao, H., Zhou, X., Hu, R.: Ready-to-react: Online reaction policy for two-character interaction generation. In: The Thirteenth International Conference on Learning Representations (2025) 2
7. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024) 7, 9
8. Chen, J., Hu, P., Chang, X., Shi, Z., Kampffmeyer, M., Liang, X.: Sitcom-crafter: A plot-driven human motion generation system in 3d scenes. In: The Thirteenth International Conference on Learning Representations (2025) 1, 2, 3
9. Chen, K., Tan, Z., Lei, J., Zhang, S.H., Guo, Y.C., Zhang, W., Hu, S.M.: Choreomaster: choreography-oriented music-driven dance synthesis. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021) 3
10. Chen, R., Shi, M., Huang, S., Tan, P., Komura, T., Chen, X.: Taming diffusion probabilistic models for character control. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–10 (2024) 4
11. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9760–9770 (2023) 3
12. Ghosh, A., Zhou, B., Dabral, R., Wang, J., Golyanik, V., Theobalt, C., Slusallek, P., Guo, C.: 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. pp. 1–11 (2025) 3, 4, 6
13. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022) 3, 4, 9, 10, 11
14. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022) 3
15. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2021–2029 (2020) 3

16. Javed, M.G., Li, X., et al.: Intermask: 3d human interaction generation via collaborative masked modeling. In: The Thirteenth International Conference on Learning Representations (2025) 3, 9, 10, 11, 13
17. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023) 3
18. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13401–13412 (2021) 3
19. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision **132**(9), 3463–3483 (2024) 2, 3, 4, 6, 9, 10, 11, 14
20. Lim, D., Bae, J., Hwang, I., Lee, S., Lee, H., Kim, Y.M.: Event-driven storytelling with multiple lifelike humans in a 3d scene. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11654–11664 (2025) 1, 2, 3
21. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Zhe, X., Iwamoto, N., Zheng, B., Black, M.J.: Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1144–1154 (2024) 3
22. Liu, M., Di, Y., Wang, G., Qu, Y., Zhu, D., Li, Y., Ji, X.: Hint: Hierarchical interaction modeling for autoregressive multi-human motion generation. arXiv preprint arXiv:2601.20383 (2026) 4, 10, 11, 15
23. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019) 9
24. Meng, Z., Han, Z., Peng, X., Xie, Y., Jiang, H.: Absolute coordinates make motion generation easy. arXiv preprint arXiv:2505.19377 (2025) 8, 10
25. Mullen, J.F., Kothandaraman, D., Bera, A., Manocha, D.: Placing human animations into 3d scenes by learning interaction-and geometry-driven keyframes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 300–310 (2023) 3
26. Ota, S., Yu, Q., Fujiwara, K., Ikehata, S., Sato, I.: Pino: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10676–10685 (2025) 15
27. Ouyang, R., Li, H., Zhang, Z., Wang, X., Zhu, Z., Huang, G., Wang, X.: Motion-r1: Chain-of-thought reasoning and reinforcement learning for human motion generation. arXiv e-prints pp. arXiv–2506 (2025) 3
28. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019) 9
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 8
30. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10985–10995 (2021) 3

31. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European conference on computer vision. pp. 480–497. Springer (2022) 3
32. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Bin Peng, X., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1911–1921 (2024) 4
33. Ruiz-Ponce, P., Barquero, G., Palmero, C., Escalera, S., García-Rodríguez, J.: in2in: Leveraging individual information to generate human interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1941–1951 (2024) 3, 10
34. Ruiz-Ponce, P., Barquero, G., Palmero, C., Escalera, S., García-Rodríguez, J.: Mixermdm: Learnable composition of human motion diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12380–12390 (2025) 3
35. Ruiz-Ponce, P., Escalera, S., García-Rodríguez, J., Deng, J., Potamias, R.A.: Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. arXiv preprint arXiv:2512.19692 (2025) 4, 11
36. Sahili, A.R., Neji, N., Tabia, H.: Text-driven motion generation: Overview, challenges and directions. arXiv preprint arXiv:2505.09379 (2025) 1
37. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019) 8
38. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: The Twelfth International Conference on Learning Representations (2024) 3, 4, 10, 11
39. Shi, Y., Wang, J., Jiang, X., Lin, B., Dai, B., Peng, X.B.: Interactive character control with auto-regressive motion diffusion models. ACM Transactions on Graphics (TOG) **43**(4), 1–14 (2024) 4
40. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 8
41. Sui, K., Ghosh, A., Hwang, I., Zhou, B., Wang, J., Guo, C.: A survey on human interaction motion generation. International Journal of Computer Vision **134**(3), 113 (2026) 1, 2
42. Tanaka, M., Fujiwara, K.: Role-aware interaction generation from textual description. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15999–16009 (2023) 3, 10
43. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: Closd: Closing the loop between simulation and diffusion for multi-task character control. In: The Thirteenth International Conference on Learning Representations (2025) 3, 4
44. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 3
45. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023) 10, 11
46. Wang, Y., Wang, S., Zhang, J., Fan, K., Wu, J., Xue, Z., Liu, Y.: Timotion: Temporal and interactive framework for efficient human-human motion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7169–7178 (2025) 3, 10, 11

47. Wu, Q., Dou, Z., Guo, C., Huang, Y., Feng, Q., Zhou, B., Wang, J., Liu, L.: Text2interact: High-fidelity and diverse text-to-two-person interaction generation. arXiv preprint arXiv:2510.06504 (2025) 3, 10
48. Xiao, L., Lu, S., Pi, H., Fan, K., Pan, L., Zhou, Y., Feng, Z., Zhou, X., Peng, S., Wang, J.: Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10086–10096 (2025) 4, 7, 9
49. Xu, L., Lv, X., Yan, Y., Jin, X., Wu, S., Xu, C., Liu, Y., Zhou, Y., Rao, F., Sheng, X., et al.: Inter-x: Towards versatile human-human interaction analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22260–22271 (2024) 9
50. Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J.C., Adeli, E.: Socialgen: Modeling multi-human social interaction with language models. In: Thirteenth International Conference on 3D Vision (2025) 3, 4
51. Yu, Q., Watanabe, A., Fujiwara, K.: Causal motion diffusion models for autoregressive motion generation. arXiv preprint arXiv:2602.22594 (2026) 8, 9
52. Zhang, Y., Feng, Y., Cseke, A., Saini, N., Bajandas, N., Heron, N., Black, M.J.: Primal: Physically reactive and interactive motor model for avatar learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12725–12736 (2025) 4
53. Zhao, K., Li, G., Tang, S.: Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. In: The Thirteenth International Conference on Learning Representations (2025) 4
54. Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14738–14749 (2023) 3
55. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019) 6
56. Zhu, B., Jiang, B., Wang, S., Tang, S., Chen, T., Luo, L., Zheng, Y., Chen, X.: Motiongpt3: Human motion as a second modality. arXiv preprint arXiv:2506.24086 (2025) 3
57. Zhuo, W., Ma, F., Fan, H.: Infinidreamer: Arbitrarily long human motion generation via segment score distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14688–14698 (2025) 4