InterGPT: Autoregressive Spatio-Temporal Modeling for Online Human Reaction Generation

Xiaolong Ma
Nanjing University of Science and Technology
Nanjing, China
xiaolong.ma@njust.edu.cn

Zicheng Jiao
Nanjing University of Science and Technology
Nanjing, China
jzc6915@njust.edu.cn

Yunlian Sun
Nanjing University of Science and Technology
Nanjing, China
yunlian.sun@njust.edu.cn

Hongwen Zhang
Beijing Normal University
Beijing, China
zhanghongwen@bnu.edu.cn

Jinhui Tang
Nanjing Forestry University
Nanjing, China
tangjh@njfu.edu.cn

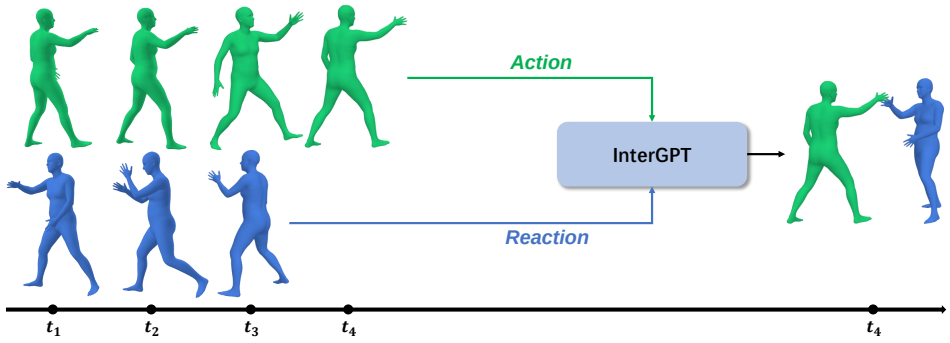


Figure 1: InterGPT generates reaction conditioned on action in an autoregressive way.

Abstract

Human reaction generation can be understood as a dynamic process where one person acts and the other reacts, inherently requiring the generation of the current reaction based on information from past time steps. However, existing approaches tend to overlook the exploration of reaction history as well as spatial dependence of body parts, during motion generation. In this paper, we propose InterGPT, a novel two-stage framework for generating human reactions with high realism and faithful adherence to real-world dynamics. In the first stage, we employ a Vector Quantized Variational Autoencoder (VQ-VAE) to discretize and reconstruct motions, thereby obtaining spatio-temporal latent representations of motions. In VQ-VAE, a graph-based encoder is designed to capture the skeletal dependence, which adopts hierarchical skeleton pooling, adaptive graph convolution and spatio-temporal modeling together to learn compact latent vectors. These latent vectors are temporally compressed and quantized to facilitate efficient autoregressive modeling. In the second stage, we adopt a spatio-temporal autoregressive architecture to generate reactions, conditioned on both action sequence up to the current time step and previously generated reaction sequence. A spatio-temporal block is further designed to model body-part dependence within each time step while preserving temporal causality, enabling structured online generation. Extensive experimental results on multiple datasets demonstrate that InterGPT achieves substantial improvements over previous methods.

CCS Concepts

• **Computing methodologies** → **Procedural animation**.

Keywords

Human Reaction Generation; Spatio-Temporal Modeling; Autoregressive Generation; Motion Generation

ACM Reference Format:
Xiaolong Ma, Zicheng Jiao, Yunlian Sun, Hongwen Zhang, and Jinhui Tang. 2026. InterGPT: Autoregressive Spatio-Temporal Modeling for Online Human Reaction Generation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836451

1 Introduction

Human reaction generation plays a vital role in animation, virtual reality, and robotics, emerging as a significant research direction in computer vision. The core characteristic of human reaction generation lies in the action-reaction causality between agents. In contrast to single-person motion generation, achieving high-fidelity reaction generation is significantly more challenging. The increased

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836451

Methods	FID ↓	Diversity →	MM Dist ↓	AITS ↓	Training Time of Model ↓	Causal Online
ReGenNet	11.445	7.137	3.860	0.43	55.2 hours	✗
InterGPT	**2.039**	**7.781**	**3.845**	**0.27**	**45 hours**	✓

Table 1: A comparative analysis is conducted between our model and the state-of-the-art model, ReGenNet [55], on the human interaction dataset, InterHuman [25]. Causal Online indicates that the generation of reactions is determined jointly by the action sequence up to the current time step and previously generated reactions. Red indicates the best result.

difficulty stems from the need to simultaneously model the action and the reaction, along with the intricate interactions between them. Furthermore, reaction generation requires capturing both global temporal dynamics and the fine-grained spatial dependence of body parts. Therefore, a key challenge lies in learning a motion representation that is both compact and structured while preserving temporal dynamics and spatial dependence.

Recent work in this field mostly employs deep generative models to generate the reaction, achieving promising progress. For instance, the Role-Aware model [43] employs a cross-attention module to bridge two Transformers [48] for interaction modeling. ReGenNet [55] fuses action and reaction via a linear layer before feeding into a diffusion network to predict the reaction. MARRS [50] uses a compact MLP as noise predictor to estimate the probability distribution over motion tokens. Despite these successes, there are still notable limitations. On one hand, most existing reaction generation approaches [4, 55] neglect the causal nature of human-human interaction. Specifically, they condition reactions solely on past action information while ignoring the previously generated reaction sequence. On the other hand, many approaches [25, 38] directly model motion as temporal sequences of joints, without explicitly modeling the spatial structure of the human body. Models such as InterFormer [4] introduce separate temporal and spatial attention mechanisms. However, they operate on dense continuous features and lack an explicit structured abstraction, which is essential for efficient spatio-temporal interaction modeling.

In this paper, we present **InterGPT**, a novel two-stage framework built upon autoregressive token prediction in a discrete spatio-temporal space. In the first stage, we employ Vector Quantized Variational Autoencoder (VQ-VAE) [47] to transform continuous motions to compact spatio-temporal representations. For spatial compression, unlike InterMask [18] which relies on direct spatial downsampling, our encoder leverages an adaptive graph-based architecture to explicitly capture skeletal dependence. The architecture adopts a learnable body-part pooling module to aggregate joints into a set of structural tokens. These features are then compressed along the temporal dimension, leading to a compact and structured representation for subsequent autoregressive generation. In the second stage, we leverage an Autoregressive Spatio-Temporal Transformer (AST-Transformer) to generate the spatio-temporal token sequence of reaction, conditioned on the action token sequence. Our AST-Transformer is an encoder-decoder architecture, which consists of an action encoder and a reaction decoder. The action encoder encodes the action sequence up to the current time step into a conditional representation. The reaction decoder then causally generates the reaction sequence conditioned on this representation. In AST-Transformer, we introduce a spatio-temporal block to model body-part dependence within each time step while preserving temporal causality. The spatio-temporal block adopts both spatial attention and temporal attention. The spatial attention models spatial dependence while the temporal attention captures temporal causality. As shown in Table 1, InterGPT achieves superior performance over the current state-of-the-art method, ReGenNet [55]. The evaluation metrics include Frechet Inception Distance (FID), Diversity, Multimodal Distance (MM Dist), Average Inference Time per Sentence (AITS) in seconds, and Training Time in hours. Notably, the causal online setting indicates that reaction generation is determined jointly by the action sequence up to the current time step and previously generated reactions. This alignment with real-world interaction scenarios demonstrates that our method is better suited for reaction generation tasks. Our main contributions are summarized as follows:

- We propose a spatio-temporal motion VQ-VAE framework with structured body-part tokenization. Spatially, it captures skeletal dependence by modeling body-part relationships. Temporally, it performs downsampling along the time dimension, ultimately constructing compact structural tokens.
- We introduce InterGPT, a novel autoregressive reaction generation framework. It dynamically generates the reaction at the current time step by conditioning on both the action sequence up to the current time step and the previously generated reaction sequence.
- Extensive experiments on InterHuman [25] and InterX [53] benchmarks demonstrate the effectiveness of our approach for reaction generation.

2 Related Work

2.1 Human Motion Generation

Human motion generation aims to generate human-like motion conditioned on diverse modalities, including text [5, 12, 13, 17, 24, 32, 33, 40, 45, 58, 60, 63, 65, 66, 68], audio [1–3, 9, 27–30, 35, 36, 41, 52, 56, 57, 62, 67, 69], or music [11, 34, 59, 64]. Among these modalities, text has been widely explored due to its flexibility and expressiveness. Models such as CLIP [37] are utilized to extract semantic features from textual descriptions. Early work such as Action2Motion [14] employs variational autoencoder (VAE) [6] to generate human motions conditioned on action category labels. However, such discrete labels are limited in expressiveness and fail to capture diverse motion patterns. Benefiting from advances in motion capture technology [46, 61], large-scale motion datasets can now be efficiently collected, enabling effective learning of motion generation models. T2M [13] introduces the large-scale text-motion dataset HumanML3D and trains a network to predict

motion length from text. Recently, many methods based on diffusion models [5, 22, 40, 45, 66] have demonstrated great quality. T2M-GPT [65] adopts VQ-VAE with 1D downsampling to generate motion tokens sequentially. While effective, its reliance on unidirectional context limits global dependence modeling. Subsequently, MoMask [12] and MMM [33] introduce conditional masked modeling, enabling bidirectional context learning and improving inference efficiency while supporting motion editing. Building upon this paradigm, BAMM [32] and BAD [17] further enhance modeling with hybrid masking strategies, addressing the limitation of fixed-length generation in conventional masked approaches.

2.2 Human-Human Interaction Generation

Unlike single human motion generation, which primarily models an agent's internal state, human-human interaction (HHI) critically depends on the mutual status. ComMDM [39] generates each agent's motion independently via separate MDM [45] branches, then uses a communication module to impose inter-agent coupling. Subsequently, InterGen [25] introduces the InterHuman dataset, a human interaction dataset rich in both text and motion modalities, and proposes a diffusion-based approach for realistic two-person motion generation. Furthermore, in2IN [38] enhances the dataset by employing a Large Language Model (LLM) to extend the InterHuman dataset with detailed individual descriptions. In order to extend HHI to multi-person scenarios, methods like ActFormer [54] and FreeMotion [7] adopt the autoregressive approach to model interactions involving more than two individuals. DyadicMamba [44] and InterMamba [51] leverage the efficient Mamba architecture as their backbone, achieving superior performance in arbitrary length. Recently, HiTMM [21] utilizes RVQ-VAE to encode two-person interactive motions into multi-layer discrete tokens. Two dedicated Transformers are then designed to hierarchically model base and residual tokens on a unified shared timeline. InterMask [18] employs VQ-VAE to encode motion sequences into 2D discrete motion token maps, and then utilizes a generative masked modeling framework to synthesize tokens of both individuals simultaneously. However, its spatial downsampling operates directly without explicitly capturing the dependence between different body parts.

2.3 Reactive Motion Generation

While HHI has gained substantial attention, reaction generation remains comparatively under-explored. InterFormer [4] is the first work to tackle reaction generation, using separate temporal and spatial attention modules to model temporal cues and spatial dependence. Inspired by [23], [43] introduces two learnable parameters to enable the network to automatically distinguish between the active and passive roles in an interaction. Each participant is modeled using a Transformer architecture, and their interactions are captured through a cross-attention module. Subsequently, ReGenNet [55] introduces a diffusion-based model equipped with a distance-based interaction loss to generate reactions in an online manner, where the actor's future motion is unavailable to the reactor. ARFlow [20] further extends ReGenNet [55] by leveraging a Flow Matching [26] architecture to construct direct transport paths between the action and reaction distributions, thereby reducing reliance on handcrafted conditional mechanisms. Beyond motion-only conditioning, E-React [70] explores emotion-driven reaction generation by incorporating an emotion prior to learn implicit affective cues from motion sequences with limited emotion labels. Meanwhile, Human-X [19] introduces an autoregressive interaction diffusion planner and an actor-aware reinforcement learning tracking policy to achieve real-time interactive motion synthesis. Furthermore, [42] employs an LLM-based model to infer action intentions and guide motion prediction. Recently, ReMoS [10] and MARRS [50] incorporate hand-level modeling to capture fine-grained contact interactions in close-proximity scenarios.

3 Method

The objective of our work is to generate a reaction $\mathbf{m}_x$ conditioned on an action $\mathbf{m}_y$. Both motion sequences are represented as tensors in $\mathbb{R}^{N\times J\times D}$, corresponding to N poses, J joints, and D-dimensional joint features. Our methodology consists of two stages. In the first stage, we introduce a Structured Motion VQ-VAE to encode motion sequences into compact discrete spatio-temporal tokens that capture skeletal dependence and body-part's temporal dynamics, as detailed in section 3.1. In the second stage, we introduce an encoder-decoder Transformer architecture that autoregressively generates reactions conditioned on the action sequence, which is described in section 3.2. The inference process is described in section 3.3.

3.1 Structured Motion VQ-VAE

In order to model temporal dynamics and spatial dependence, we introduce Structured Motion VQ-VAE, which encodes skeleton sequences into compact discrete spatio-temporal tokens. The overall architecture is shown in Figure 2(a). The encoder alternates adaptive graph convolution (AGC) and hierarchical skeleton pooling for spatial encoding, followed by a spatio-temporal modeling module and two temporal downsampling blocks to obtain compact latent features for vector quantization. The decoder mirrors the encoder by performing temporal upsampling followed by hierarchical graph reconstruction with AGC and skeleton unpooling.

Hierarchical Skeleton Pooling. Inspired by SALAD [16] and InterMoE [49], we adopt a pooling strategy to progressively aggregate skeletal information. We first aggregate original joints into intermediate body parts, and then further compress these representations into final body parts, leading to three levels of representations. Each pooling operation is implemented as a fixed weighted mapping derived from the kinematic tree, preserving the coarse anatomical topology while reducing spatial redundancy. This hierarchical abstraction produces compact and structured body-part representations, progressively processing skeletal representations from the original joint level to intermediate and final body-part representations for efficient multi-scale modeling of human motion, as illustrated in Figure 2(b).

Adaptive Graph Convolution. The hierarchical representation naturally forms graph structures at multiple levels, including the original joint graph as well as the intermediate and final body-part graphs. To model spatial dependence across these hierarchical graph representations, we employ graph convolution. Notably, relying solely on a fixed skeletal graph may limit the model's ability to capture spatial dependence. To address this limitation, we adopt an adaptive graph convolution layer, which augments the predefined

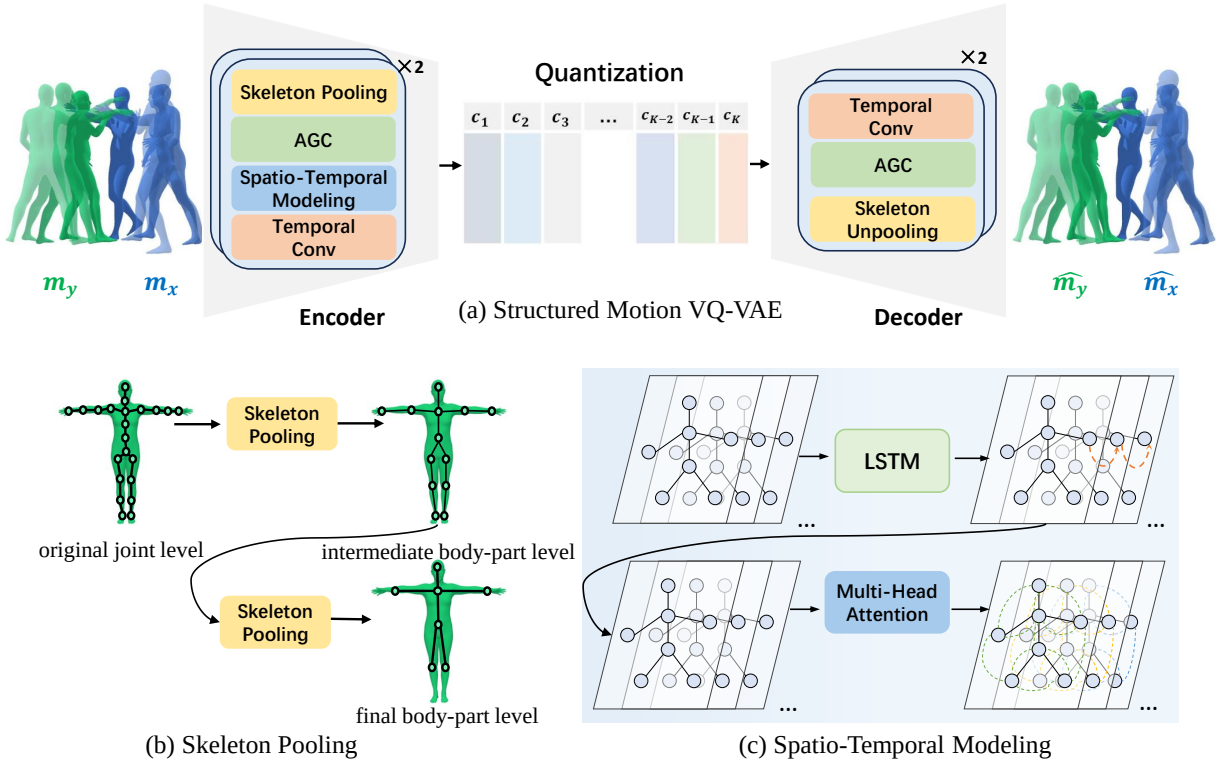


Figure 2: The architecture of Structured Motion VQ-VAE.

skeleton adjacency with learnable connections. Given an initial adjacency matrix A_0, we introduce a learnable matrix B to adaptively refine the graph structure:

$$A = \text{Softmax}\left(\frac{A_0 + \alpha B}{\tau}\right), \tag{1}$$

where α is a learnable scaling factor and τ controls the temperature of the normalization. The resulting normalized adjacency matrix is then used for message passing among different joints/body-parts:

$$U' = \sigma(AUW), \tag{2}$$

where U denotes joint/body-part features, $\sigma(\cdot)$ denotes a non-linear activation function, and W is a learnable projection matrix. This adaptive formulation preserves the previous skeletal topology while enabling the model to autonomously learn spatial dependence.

Spatio-Temporal Modeling. With the final body-part graph (processed by AGC), we further introduce a spatio-temporal modeling module to jointly model temporal dynamics and spatial coordination. As illustrated in Figure 2(c), the temporal operation processes each body-part sequence using a unidirectional LSTM [8], capturing the progression of motion while respecting causal temporal order. The spatial operation instead employs a multi-head attention (MHA) layer to enable coordination among different body parts. By integrating these two complementary operations, the module produces rich spatio-temporal representations that are both compact and expressive.

Tokenization and Sequence Organization. Given the encoded representation $\mathbf{z}_x, \mathbf{z}_y \in \mathbb{R}^{n\times j\times d}$, where n and j denote the temporal length and number of body parts after downsampling from the original motion $\mathbf{m}_x, \mathbf{m}_y \in \mathbb{R}^{N\times J\times D}$. d is the dimension of the latent feature vectors. Each $n \times j$ vector in $\mathbf{z}_x$, $\mathbf{z}_y$ is then quantized by replacing it with the closest entry from a learnable codebook $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K}$, yielding a discrete index map $\mathbf{s} \in \mathbb{Z}^{n\times j}$, where each entry $s_{t,p} \in \{1, ..., K\}$ corresponds to a codebook index. The corresponding embeddings are retrieved from the codebook for reconstruction via the decoder. Before feeding the tokens into the second stage, we organize them in a temporal-first manner to facilitate autoregressive modeling while preserving the spatial structure. Specifically, we rearrange the index map: $\mathbf{s} = \{s_{t,p} \mid t = 1, \ldots, n,\ p = 1, \ldots, j\} = \{s_{1,1}, s_{1,2}, \ldots, s_{1,j}, \ldots, s_{n,1}, s_{n,2}, \ldots, s_{n,j}\}$. This ordering ensures that tokens at time step t are generated before those at $t + 1$, while maintaining the spatial structure of body parts within each time step.

Optimization Goal. To learn a representative codebook, we train VQ-VAE by reconstructing motion sequences $\mathbf{m}$ (including both $\mathbf{m}_x$ and $\mathbf{m}_y$). The standard VQ-VAE objective consists of a reconstruction loss and a commitment loss:

$$\mathcal{L}_{vq} = \|\mathbf{m} - \hat{\mathbf{m}}\|_1 + \beta\|\mathbf{z} - \text{sg}[\mathbf{z_q}]\|_2^2, \tag{3}$$

where $\mathbf{m}$ is the input motion, $\hat{\mathbf{m}}$ is its reconstruction, $\mathbf{z}$ is the encoder output, and $\mathbf{z}_q$ is the quantized latent representation. β is a hyperparameter controlling the commitment loss, and $\text{sg}(\cdot)$ denotes the stop-gradient operator.

Following InterMask [18], we further incorporate geometric constraints to enhance motion fidelity, including velocity loss, bone

length loss, and foot contact loss. These losses are computed as:

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{i=1}^{N} \|(m_{i+1} - m_i) - (\hat{m}_{i+1} - \hat{m}_i)\|_1, \tag{4}$$

$$\mathcal{L}_{\text{fc}} = \frac{1}{N-1}\sum_{i=1}^{N} \|(\hat{m}_{i+1} - \hat{m}_i) \cdot f_i\|_1, \tag{5}$$

$$\mathcal{L}_{\text{bl}} = \frac{1}{N-1}\sum_{i=1}^{N} \|B(m_i) - B(\hat{m}_i)\|_1. \tag{6}$$

Here, m_i and $\hat{m}_i$ denote the ground-truth pose and its reconstruction at time step i, respectively, and N is the total number of frames in the sequence. $f_i \in \{0, 1\}$ indicates the foot contact state, and B computes the bone lengths defined by the kinematic tree of the human body.

The overall training objective for VQ-VAE is formulated as:

$$\mathcal{L}_{\text{vqvae}} = \mathcal{L}_{vq} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{fc}}\mathcal{L}_{\text{fc}} + \lambda_{\text{bl}}\mathcal{L}_{\text{bl}}, \tag{7}$$

where λ_{vel}, λ_{fc}, and λ_{bl} are hyperparameters balancing the contribution of each geometric loss.

3.2 AST-Transformer

For human reaction generation, we argue that reactions are influenced not only by the action sequence up to the current time step but also by the sequence of previously generated reactions. Therefore, we propose a novel reaction generation module conditioned on both action and generated reaction. Given an action sequence represented as discrete tokens $\mathbf{y} \in \mathbb{Z}^{n\times j}$, our goal is to generate the corresponding reaction sequence $\mathbf{x} \in \mathbb{Z}^{n\times j}$ in an online manner. Each token $y_{t,p}/x_{t,p}$ represents the codebook index of the p-th body part at time step t. We formulate this task as a conditional autoregressive sequence modeling problem, where the reaction sequence is generated token by token under strict temporal causality:

$$p(\mathbf{x} \mid \mathbf{y}) = \prod_{t=1}^{n} p(\mathbf{x}_{t,:} \mid \mathbf{y}_{\le t,:}, \mathbf{x}_{<t,:}). \tag{8}$$

After obtaining $\mathbf{x}$, we project it into the continuous embedding space via the codebook $\mathbf{C}$, then input it into the decoder for the final reconstruction. This formulation ensures the prediction at step t depends solely on past reactions and observed actions to t, preventing future leakage. For efficiency, body-part tokens within each time step are generated in parallel. As illustrated in Figure 3(a), we adopt an encoder-decoder Transformer, where the action encoder processes the action sequence and the reaction decoder autoregressively generates the reaction.

Spatio-Temporal Block. To capture both intra-frame spatial coordination and inter-frame temporal dynamics, we design a spatio-temporal block that decouples spatial and temporal attention while preserving the structured (n, j) layout. The discrete reaction and action tokens are first projected into continuous latent embeddings $\mathbf{h} \in \mathbb{R}^{n\times j\times \widetilde{d}}$ via an embedding layer. Specifically, each discrete index $s_{t,p}$ is mapped to a $\widetilde{d}$-dimensional vector, where $\widetilde{d}$ denotes the hidden dimension. As shown in Figure 3(b), each block processes these representations through two specialized MHA modules. First, we apply **spatial attention** independently for each time step to allow body parts to interact within the same time step:

$$\mathbf{h}^{\text{spat}}_{t,:} = \mathbf{h}_{t,:} + \text{MHA}(Q = \mathbf{h}_{t,:}, K = \mathbf{h}_{t,:}, V = \mathbf{h}_{t,:}), \tag{9}$$

where $\mathbf{h}_t \in \mathbb{R}^{j\times\widetilde{d}}$ denotes the latent features at time step $t \in \{1, \cdots, n\}$. Next, we model the **temporal dynamics** independently for each body part $p \in \{1, \cdots, j\}$ along the time dimension:

$$\mathbf{h}^{\text{temp}}_{:,p} = \mathbf{h}_{:,p} + \text{CausalMHA}(Q = \mathbf{h}_{:,p}, K = \mathbf{h}_{:,p}, V = \mathbf{h}_{:,p}), \tag{10}$$

where $\mathbf{h}_{:,p} \in \mathbb{R}^{n\times\widetilde{d}}$ denotes the sequence of features for the p-th body part across all n time steps. The causal attention mechanism is defined as:

$$\text{CausalMHA}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V, \tag{11}$$

where d_k denotes the feature dimension of each attention head, and $M \in \mathbb{R}^{n\times n}$ is the causal mask matrix defined as:

$$M_{ij} = \begin{cases} 0, & i \ge j, \\ -\infty, & i < j \end{cases}. \tag{12}$$

This mask ensures the temporal causality. Each attention module is preceded by LayerNorm and followed by a feed-forward network (FFN) with residual connections. By decoupling spatial and temporal attention, the spatio-temporal block effectively captures both spatial dependence and temporal dynamics.

Conditional Autoregressive Generation. To enable conditional generation, we incorporate a cross-attention mechanism into the decoder, where reaction tokens serve as queries and action tokens provide keys and values. Given the decoder hidden states $\mathbf{h}_x \in \mathbb{R}^{n\times j\times\widetilde{d}}$ and the encoded action features $\mathbf{h}_y \in \mathbb{R}^{n\times j\times\widetilde{d}}$, the cross-attention operation is defined as:

$$\text{CrossAttn}(\mathbf{h}_x, \mathbf{h}_y) = \text{MHA}(Q = \mathbf{h}_x, K = \mathbf{h}_y, V = \mathbf{h}_y). \tag{13}$$

As shown in Figure 3(c), to preserve temporal causality for online generation, we further impose the causal mask matrix M in Eq.(12) on the cross attention, as described previously. This mask ensures that each reaction at time step t can only attend to action tokens up to time t, preventing any future information leakage. With this design, the model generates reaction tokens in a fully autoregressive manner while dynamically conditioning on the observed action sequence, strictly adhering to temporal causality.

Parallel Token Prediction. Benefiting from the structured spatio-temporal modeling capability of the spatio-temporal block, we adopt a parallel token prediction strategy within each time step. Rather than generating tokens sequentially across body parts, the model predicts all j body-part tokens at time step t simultaneously ($x_t = \{x_{t,1}, \ldots, x_{t,j}\}$, where $x_{t,p}$ denotes the discrete index for the p-th body part). This design leverages the spatial coordination already captured by the spatio-temporal block, enabling consistent motion generation across body parts while substantially improving inference efficiency. The model thus maintains structured coherence without sacrificing the parallelism required for real-time generation.

Training Strategy. We employ classifier-free guidance (CFG) [15] during training. Rather than completely disabling the action condition (i.e., using an empty condition), we randomly drop out several tokens from the action sequence [52]. To mitigate the exposure bias in autoregressive generation, we adopt a scheduled sampling strategy during training. Specifically, at each time step, ground-truth

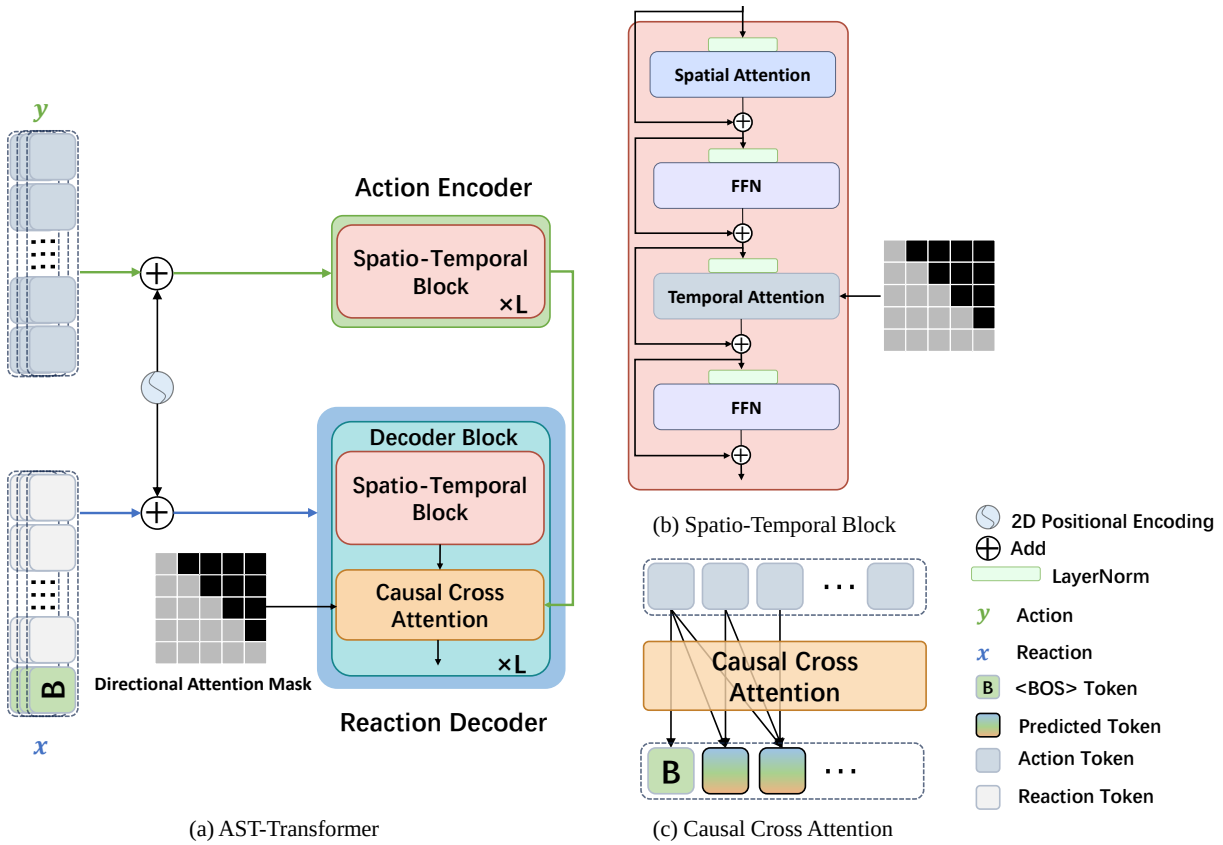


Figure 3: Overview of AST-Transformer. (a) 2D tokens y and x pass through an action encoder of L spatio-temporal blocks and a reaction decoder of L blocks with causal cross attention. (b) Each block includes LayerNorm, spatial attention, temporal attention, and FFN. (c) Causal cross attention prevents future leakage.

Dataset	Model	FID ↓	Diversity →	MM Dist ↓	Online	Causal Online
InterHuman	Ground Truth	$0.273^{\pm.007}$	$7.948^{\pm.064}$	$3.755^{\pm.008}$	-	-
	InterGen [25]	$22.243^{\pm.319}$	$7.310^{\pm.033}$	$3.946^{\pm.002}$	✗	✗
	InterMask [18]	$18.745^{\pm.251}$	$7.258^{\pm.043}$	$3.944^{\pm.002}$	✗	✗
	MDM [45]	$21.998^{\pm.208}$	$7.121^{\pm.026}$	$3.907^{\pm.001}$	✗	✗
	ReGenNet [55]	$11.445^{\pm.134}$	$7.137^{\pm.021}$	$3.860^{\pm.000}$	✓	✗
	InterGPT	$2.039^{\pm.043}$	$7.781^{\pm.030}$	$3.845^{\pm.001}$	✓	✓
InterX	Ground Truth	$0.002^{\pm.000}$	$8.669^{\pm.055}$	$4.005^{\pm.013}$	-	-
	InterMask [18]	$0.696^{\pm.013}$	$8.404^{\pm.067}$	$5.181^{\pm.016}$	✗	✗
	MDM [45]	$0.543^{\pm.014}$	$8.287^{\pm.056}$	$4.925^{\pm.014}$	✗	✗
	ReGenNet [55]	$0.165^{\pm.003}$	$8.718^{\pm.054}$	$4.986^{\pm.000}$	✓	✗
	InterGPT	$0.288^{\pm.009}$	$8.409^{\pm.089}$	$5.051^{\pm.013}$	✓	✓

Table 2: Comparison of online methods on InterHuman and InterX datasets. Red and Blue indicate the best and the second best results.

tokens $x_{t,p}$ are partially replaced with model predictions $\hat{x}_{t,p}$ with a probability, which increases gradually. This strategy bridges the gap between training and inference, improving the robustness of the model during long-term generation.

3.3 Inference

The inference process begins by obtaining the action tokens $\mathbf{y} \in \{1, \cdots, K\}^{n \times j}$, which serve as the conditioning context. We initialize the reaction with a *<BOS>* token for each body part: $\mathbf{x_1} = [BOS, ..., BOS] \in \{1, \cdots, K\}^{j}$, where j denotes the number of body parts. This design is consistent with our parallel body-part tokenization. Conditioned on the encoded action sequence, the Reaction-Decoder generates reaction tokens autoregressively over time. At each time step t, the model takes the previously generated tokens $\mathbf{x_1}, ..., \mathbf{x_{t-1}}$ along with the action tokens $\mathbf{y}_{\leq t}$ to the current time step, to predict all j body parts simultaneously. This process continues until the maximum sequence length n is reached. The causal masking in both temporal attention and cross attention mechanisms prevents access to future reaction and future action, ensuring consistency with the online setting.

Finally, the initial *<BOS>* tokens are removed, and the generated index map $\mathbf{x} \in \{1, \cdots, K\}^{n \times j}$ is fed into the VQ-VAE decoder to reconstruct the reaction motion $\mathbf{m}_x \in \mathbb{R}^{N \times J \times D}$.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate InterGPT on two datasets for reaction generation: InterHuman [25] and InterX [53]. The InterHuman dataset consists of 7779 motions derived from various categories of human actions. InterX contains 11388 interaction sequences, which has causal interaction order annotations. InterHuman adopts a non-canonical representation $\mathbf{m} = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f]$, where $\mathbf{j}_g^p \in \mathbb{R}^{3J}$ represents global joint positions, $\mathbf{j}_g^v \in \mathbb{R}^{3J}$ represents velocities in the world frame, $\mathbf{j}^r \in \mathbb{R}^{6J}$ denotes 6D representation of local rotations in the root frame, and $\mathbf{c}^f \in \mathbb{R}^4$ is the binary foot-ground contact label. Following SMPL-X [31], the InterX dataset instead adopts a skeleton representation with 55 joints, including body, hand and face joints. The global position $\mathbf{p}_g \in \mathbb{R}^3$ and global velocity $\mathbf{v}_g \in \mathbb{R}^3$ are further included, leading to the final $\mathbf{m} \in \mathbb{R}^{N \times 56 \times 6}$.

Implementation Details. For hierarchical skeleton pooling, we construct three levels of representations. Specifically, InterHuman contains 22 joints, which are progressively pooled into 12 intermediate body parts and 7 final body parts. For InterX, each motion sequence contains 55 joints and an additional global root translation. We model the global translation as a separate token rather than a joint to avoid mixing global trajectory with local joint rotations during spatial modeling. The 55 joints are hierarchically pooled into 14 intermediate and 7 final body parts, combined with the original root translation token. The latent representation in VQ-VAE has a dimension $d = 512$, and the codebook size $|\mathbf{C}| = 1024$. For AST-Transformer, we use $L = 6$ spatio-temporal blocks, each with 6 attention heads. The transformer embedding dimension is $\widetilde{d} = 384$.

Dataset	Method	Reconstruction FID ↓
InterHuman	1D Token Sequence	3.127
	2D Token Map [18]	0.976
	w/o Spatio-Temporal Modeling	0.888
	w/o AGC	0.870
	InterGPT	**0.837**

Table 3: Ablation study results on InterHuman to verify key components of the proposed Structured Motion VQ-VAE. Bold indicates the best result.

4.2 Quantitative Results

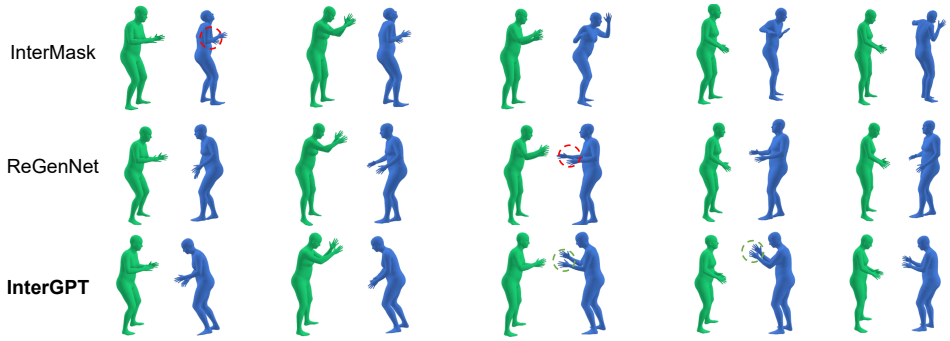
Table 2 shows the comparison between existing methods and our InterGPT. All evaluations are run 20 times, and we report the averaged results along with a 95% confidence interval. ***Online*** refers to the generation setting where only the action up to the current time step can be accessed. ***Causal Online*** indicates that the generation of reactions is determined jointly by the action sequence up to the current time step and previously generated reactions. For InterMask [18] and InterGen [25], we adopt official pretrained weights without text conditioning to suit the online setting. ReGenNet [55] is retrained on InterHuman and InterX following its default configuration. As MDM [45] is originally designed for single-person scenarios, we apply minor architectural adjustments and retrain it on both datasets. On the InterHuman dataset, our method achieves the best performance across all metrics. Although InterMask [18] and InterGen [25] achieve strong performance in general HHI motion generation, their performance degrades in the reaction generation setting. This may be caused by the lack of explicit causal modeling and their reliance on text conditions, which are absent in our task formulation. On the InterX dataset, our method achieves the second-best performance, while ReGenNet [55] achieves the top results. We attribute this gap primarily to differences in motion representation and dataset characteristics. InterX contains fine-grained motions with 55 joints, where diffusion-based ReGenNet preserves well the detailed joint-level information. In contrast, our VQ-VAE compresses the skeleton into a small set of body-part tokens, which may lead to a loss of fine-grained details despite achieving strong overall reconstruction quality. Moreover, InterX sequences are relatively short, reducing the difficulty of long-term temporal modeling and favoring methods with strong local modeling capability. Nevertheless, our method achieves competitive performance while offering significantly higher efficiency for online generation.

4.3 Qualitative Comparison

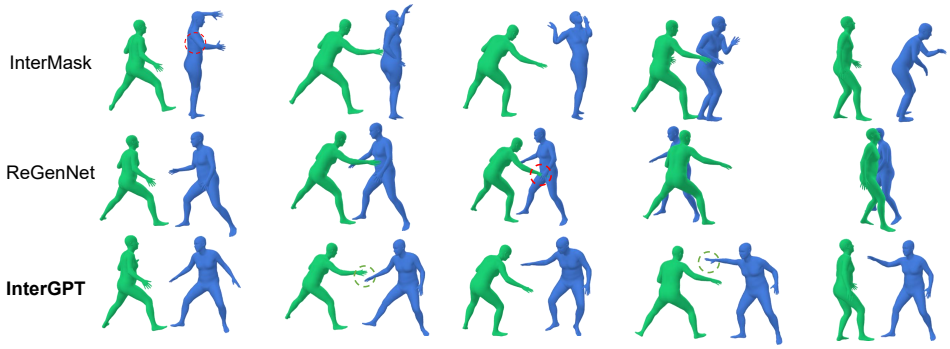
We further compare our method with ReGenNet [55] and InterMask [18] through qualitative visualizations in Figure 4, where both methods are evaluated under the online generation setting on the InterHuman dataset. The results show that our approach produces more coherent interactions and more plausible reaction.

4.4 Ablation Study

Structured Motion VQ-VAE. Table 3 presents an ablation study to evaluate key components of our proposed Structured Motion



(a) The first person gives the basketball to the second.



(b) The first one takes a defensive stance and lunges forward to tag the second one.

Figure 4: Qualitative comparison with InterMask and ReGenNet on InterHuman. Red circles highlight problematic interactions (e.g., wrong direction, incorrect reaction motions), while Green circles emphasize our model's plausibility. Text serves only as semantic reference, not as a generative condition.

Method	FID ↓	Diversity →	MM Dist ↓
Ground Truth	0.273	7.948	3.755
Decoder-Only	3.359	7.787	3.882
w/o Spatio-Temporal Block	7.242	7.558	3.913
w/o Scheduled Sampling	2.463	7.752	3.860
w/o Reaction History	2.176	7.766	**3.838**
InterGPT	**2.039**	**7.781**	3.845

Table 4: Ablation study results on the InterHuman test set to verify key components of the proposed AST-Transformer, and alternative designs. Bold indicates the best result.

VQ-VAE. The results show that the 1D token representation yields the worst reconstruction performance in terms of FID. Using a 2D token map, where joints are randomly grouped into five body parts as in InterMask [18], leads to noticeable improvement. Building upon this, incorporating the proposed AGC and spatio-temporal modeling module further reduces FID to **0.837**, achieving the best performance. These results demonstrate that both the structured spatial modeling and the joint spatio-temporal modeling are critical for high-quality motion reconstruction.

AST-Transformer. Table 4 presents an ablation study of AST-Transformer, including key components, alternative designs such as using a decoder-only architecture with flattened token sequence, as well as the impact of reaction history. The results show that the proposed spatio-temporal block and encoder–decoder architecture consistently improve the performance. While decoder-only models are common in autoregressive generation, the performance is inferior to our method. Moreover, removing reaction history causes performance degradation, confirming that modeling past reactions contributes to generation quality.

5 Discussion and Conclusion

In this paper, we present InterGPT for human reaction generation via autoregressive modeling in a discrete space, adhering to real-world scenarios. First, a structured VQ-VAE was designed to encode motions into compact spatio-temporal 2D tokens. Next, AST-Transformer was introduced to generate the reaction token sequence conditioned on both action tokens and previously generated reaction tokens. Extensive experimental results demonstrate that our method achieves impressive results. However, discrete tokenization may lose fine-grained motion details (e.g., in InterX), and the model lacks explicit long-term intention modeling for complex interactions. Future efforts will be devoted to more expressive motion representation learning, intention-aware modeling, and multi-person interaction generation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62476131 and Grant 62377004, in part by the Major Science and Technology Projects in Jiangsu Province under Grant BG2024042.

References

[1] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6774–6783.

[2] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. 2025. 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.

[3] Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo Bolkart, et al. 2024. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1942–1953.

[4] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. 2023. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia* 25 (2023), 8842–8854.

[5] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9760–9770.

[6] P Kingma Diederik and Welling Max. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392.

[7] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*. Springer, 93–109.

[8] F.A. Gers, J. Schmidhuber, and F. Cummins. 1999. Learning to forget: continual prediction with LSTM. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, Vol. 2. 850–855 vol.2. doi:10.1049/cp:19991218

[9] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 206–216.

[10] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. 2024. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European conference on computer vision*. Springer, 418–437.

[11] 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.

[12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5152–5161.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM international conference on multimedia*. 2021–2029.

[15] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[16] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. 2025. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7158–7168.

[17] Seyed Rohollah Hosseyni, Ali Ahmad Rahmani, Seyed Jamal Seyedmohammadi, Sanaz Seyedin, and Arash Mohammadi. 2025. Bad: Bidirectional auto-regressive diffusion for text-to-motion generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[18] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. 2025. InterMask: 3D Human Interaction Generation via Collaborative Masked Modeling. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=ZAyuwJYN8N

[19] Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. 2025. Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10173–10183.

[20] Wentao Jiang, Jingya Wang, Kaiyang Ji, Baoxiong Jia, Siyuan Huang, and Ye Shi. 2025. Arflow: Human action-reaction flow matching with physical guidance. *arXiv preprint arXiv:2503.16973* (2025).

[21] Zicheng Jiao, Yunlian Sun, Hongwen Zhang, Jinhui Tang, and Massimo Tistarelli. 2026. HiTMM: Generative Temporal Masked Modeling of Human Interactive Motions. *IEEE Transactions on Visualization and Computer Graphics* (2026).

[22] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 8255–8263.

[23] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 10 (2017), 1901–1913.

[24] Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence Yang. 2025. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. In *International Conference on Learning Representations*, Vol. 2025. 84238–84250.

[25] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision* 132, 9 (2024), 3463–3483.

[26] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

[27] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024. 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*. 1144–1154.

[28] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.

[29] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. 2025. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 10929–10939.

[30] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1388–1398.

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. 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*. 10975–10985.

[32] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. 2024. Bamm: bidirectional autoregressive motion model. In *European Conference on Computer Vision*. Springer, 172–190.

[33] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. 2024. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1546–1555.

[34] Qiaosong Qi, Le Zhuo, Aixi Zhang, Yue Liao, Fei Fang, Si Liu, and Shuicheng Yan. 2023. Diffdance: Cascaded human motion diffusion model for dance generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1374–1382.

[35] Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. 2024. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10424–10434.

[36] Xingqun Qi, Hengyuan Zhang, Yatian Wang, Jiahao Pan, Chen Liu, Muyi Sun, Wei Xue, Shanghang Zhang, Sirui Han, Qifeng Liu, et al. 2025. CoCoGesture: Towards coherent co-speech 3D gesture generation in the wild. *Information Fusion* (2025), 103613.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[38] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. 2024. in2in: Leveraging individual information to generate human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1941–1951.

[39] 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*.

[40] Xu Shi, Wei Yao, Chuanchen Luo, Junran Peng, Hongwen Zhang, and Yunlian Sun. 2024. FG-MDM: Towards Zero-Shot Human Motion Generation via ChatGPT-Refined Descriptions. In *International Conference on Pattern Recognition*. Springer, 446–461.

[41] Mingze Sun, Chao Xu, Xinyu Jiang, Yang Liu, Baigui Sun, and Ruqi Huang. 2025. Beyond talking–generating holistic 3d human dyadic motion for communication. *International Journal of Computer Vision* 133, 5 (2025), 2910–2926.

[42] Wenhui Tan, Boyuan Li, Chuhao Jin, Wenbing Huang, Xiting Wang, and Ruihua Song. 2025. Think-then-react: Towards unconstrained human action-to-reaction generation. *arXiv preprint arXiv:2503.16451* (2025).

[43] Mikihiro Tanaka and Kent Fujiwara. 2023. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF international conference on computer vision*. 15999–16009.

[44] Julian Tanke, Takashi Shibuya, Kengo Uchida, Koichi Saito, and Yuki Mitsufuji. 2025. Dyadic Mamba: Long-term dyadic human motion synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2868–2877.

[45] 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*. https://openreview.net/forum?id=SJ1kSyO2jwu

[46] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence* 45, 12 (2023), 15406–15425.

[47] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[49] Lipeng Wang, Hongxing Fan, Haohua Chen, Zehuan Huang, and Lu Sheng. 2026. InterMoE: Individual-Specific 3D Human Interaction Generation via Dynamic Temporal-Selective MoE. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 2137–2145.

[50] Yabiao Wang, Shuo Wang, Jiangning Zhang, Jiafu Wu, Qingdong He, and Yong Liu. 2026. Marrs: Masked autoregressive unit-based reaction synthesis. *IEEE Transactions on Visualization and Computer Graphics* (2026).

[51] Zizhao Wu, Yingying Sun, Yiming Chen, Xiaoling Gu, Ruyu Liu, and Jiazhou Chen. 2025. InterMamba: Efficient Human-Human Interaction Generation With Adaptive Spatio-Temporal Mamba. *IEEE Transactions on Visualization and Computer Graphics* (2025).

[52] Yong Xie, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. 2026. ReCoM: Realistic Co-Speech Motion Generation with Recurrent Embedded Transformer. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 13162–13166.

[53] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. 2024. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22260–22271.

[54] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. 2023. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2228–2238.

[55] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. 2024. Regennet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1759–1769.

[56] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. 2024. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems* 37 (2024), 20055–20080.

[57] Haiwei Xue, Sicheng Yang, Zhensong Zhang, Zhiyong Wu, Minglei Li, Zonghong Dai, and Helen Meng. 2024. Conversational co-speech gesture generation via modeling dialog intention, emotion, and context with diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8296–8300.

[58] Zhao Yang, Bing Su, and Ji-Rong Wen. 2023. Synthesizing long-term human motions with diffusion models via coherent sampling. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3954–3964.

[59] Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. 2023. Dance with you: The diversity controllable dancer generation via diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8504–8514.

[60] Wei Yao, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. 2026. Hosig: Full-body human-object-scene interaction generation with hierarchical scene perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 11901–11909.

[61] Wei Yao, Hongwen Zhang, Yunlian Sun, and Jinhui Tang. 2024. Staf: 3d human mesh recovery from video with spatio-temporal alignment fusion. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 11 (2024), 10564–10577.

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 469–480.

[63] Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. 2024. Mogents: Motion generation based on spatial-temporal joint modeling. *Advances in Neural Information Processing Systems* 37 (2024), 130739–130763.

[64] Jinming Zhang, Yunlian Sun, Hongwen Zhang, and Jinhui Tang. 2025. EDMG: Towards Efficient Long Dance Motion Generation with Fundamental Movements from Dance Genres. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10447–10456.

[65] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14730–14740.

[66] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence* 46, 6 (2024), 4115–4128.

[67] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–17.

[68] Zongye Zhang, Bohan Kong, Qingjie Liu, and Yunhong Wang. 2025. Towards robust and controllable text-to-motion via masked autoregressive diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 9326–9335.

[69] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. 2023. Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 179–185.

[70] Chen Zhu, Buzhen Huang, Zijing Wu, Binghui Zuo, and Yangang Wang. 2025. E-React: Towards Emotionally Controlled Synthesis of Human Reactions. *arXiv preprint arXiv:2508.06093* (2025).