InterEdit: Navigating Text-Guided 3D Dyadic Human Motion Editing

Yebin Yang[1], Di Wen[1], Lei Qi[1], Weitong Kong[1], Junwei Zheng[1], Ruiping Liu[1], Yufan Chen[1], Chengzhi Wu[1], Kailun Yang[2], Yuqian Fu[3], Danda Pani Paudel[4], Luc Van Gool[4], and Kunyu Peng[1,4,*]

[1] Karlsruhe Institute of Technology, Germany
[2] Hunan University, China
[3] KAUST, Saudi Arabia
[4] INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria

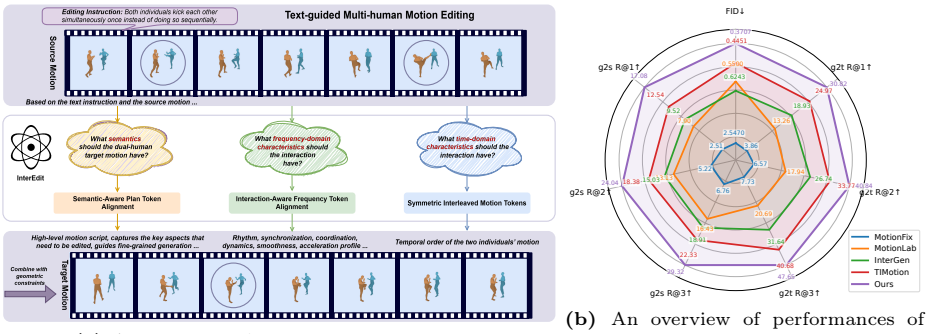
Abstract. Text-guided 3D motion editing has seen success in single-person scenarios, but its extension to multi-person settings is less explored due to limited paired data and the complexity of inter-person interactions. We introduce the task of multi-person 3D motion editing, where a target motion is generated from a source and a text instruction. To support this, we propose **InterEdit3D**, a new dataset with manual two-person motion change annotations, and a Text-guided Multi-human Motion Editing (TMME) benchmark. We present **InterEdit**, a synchronized classifier-free conditional diffusion model for TMME. It introduces Semantic-Aware Plan Token Alignment with learnable tokens to capture high-level interaction cues and an Interaction-Aware Frequency Token Alignment strategy using DCT and energy pooling to model periodic motion dynamics. Experiments show that **InterEdit** improves text-to-motion consistency and edit fidelity, achieving state-of-the-art TMME performance. The dataset and code will be released at `https://github.com/YNG916/InterEdit`.

1 Introduction

Human 3D motion generation [3,11,17,21,22,24,27,53,58,59] has seen significant advancements with large-scale datasets and diffusion-based models, enabling realistic behavior synthesis from textual descriptions. However, pure text generation is often insufficient for practical content creation, where animators, game designers, and AI systems require *controlled modifications* of existing motions rather than full synthesis [23]. Text-guided motion editing [2,12,15,45–47,56,57] addresses this by allowing users to modify a source motion based on a text instruction, producing a target motion that changes only the requested parts while preserving other content. This controllable refinement is crucial for iterative design, personalization, and human-AI collaboration.

While text-driven motion editing [2,12,15,29,57] has shown promising results for single-person scenarios, extending it to multi-human interactions remains

* Correspondence: kunyu.peng@kit.edu



(a) An overview of the task and **InterEdit**.

(b) An overview of performances of baselines and our **InterEdit**.

Fig. 1: An illustration of (a) Text-guided Multi-human 3D Motion Editing (TMME) task and our proposed **InterEdit** model, and (b) the performances of baselines (*i.e.*, MotionFix [2], MotionLab [12], InterGen [24], TIMotion [46]) and our **InterEdit**.

underexplored. Many real-world human behaviors involve interactions—such as collaboration, competition, and physical contact—that require multiple participants [31, 32]. In these cases, motion meaning arises not only from individual movements but from *spatio-temporal coupling*, including synchronization, phase alignment, positioning, role switching, and contact timing. Editing such interactions is crucial for applications like character animation, social robotics, virtual agents, crowd simulation, and training data for embodied intelligence, where fine-grained control over interaction dynamics is essential. Editing multi-human motion is more challenging than single-actor editing due to the interaction semantics in relative timing and configuration. A small modification to one person can disrupt synchronization or spatial consistency. Unlike generation, editing must stay anchored to the source motion and apply only instruction-relevant changes. This "change what is requested, preserve the rest" constraint is harder in interactive settings, where small temporal deviations can alter semantics. Despite its importance, there is no dedicated benchmark for multi-human 3D motion editing.

In this work, we introduce *Text-guided Multi-human Motion Editing* (TMME), a new task that aims to generate a target two-person motion conditioned on both source motion and textual editing instruction (Fig. 1). This task exposes two major challenges. First, paired source–target–instruction triplets for multi-human motion are scarce, as existing datasets are primarily designed for generation. Second, interaction editing requires precise modeling of coordination patterns such as synchrony versus alternation, tempo variation, spatial approach and separation, and role-dependent responses.

To enable systematic study of this problem, we construct **InterEdit3D**, the first large-scale two-person motion editing dataset built on top of InterHuman [24] via a semi-automatic retrieval-and-annotation pipeline. We retrieve motion pairs that share a similar base motion for one individual while differing in interaction structure, and annotate them with editing instructions describing

how to transform the source into the target. This results in 5,161 source–target–text triplets emphasizing spatial, temporal, and coordination-level edits, forming a challenging benchmark for interaction-aware motion editing.

We propose **InterEdit**, a classifier-free conditional diffusion framework specifically designed for dual-person motion editing. To determine what to modify while preserving source consistency, we introduce *Semantic-Aware Plan Token Alignment*: learnable plan tokens aligned with a pretrained motion teacher embedding of the target motion, providing semantic-level guidance and ensuring that the edits align with the high-level intent expressed in the text instruction. To preserve temporal coupling and coordination, we further introduce *Interaction-Aware Frequency Token Alignment. Interaction-Aware Frequency Token Alignment* applies the Discrete Cosine Transformation (DCT) to average and difference interaction signals to obtain band-energy descriptors, which are mapped into frequency control tokens. These tokens are supervised to regress the target band-energy profiles, encouraging edits that respect interaction rhythm, synchrony, and other coupling dynamics. By combining Semantic-Aware Plan Token Alignment and Interaction-Aware Frequency Token Alignment, **InterEdit** effectively captures both high-level editing intent and fine-grained interaction dynamics, which are crucial for generating realistic and semantically accurate multi-human motion edits.

Our main contributions are summarized as follows:

- We introduce the Text-guided Multi-human Motion Editing (TMME) benchmark, the first benchmark for multi-person motion editing, with 4 single-person editing and multi-person generation baselines.
- We present **InterEdit3D**, the first large-scale multi-person motion editing dataset containing 5,161 source–target–text triplets, created through a scalable retrieval pipeline.
- We propose **InterEdit**, a multi-human conditional diffusion framework incorporating semantic-aware plan token alignment and interaction-aware frequency token alignment, ensuring precise and temporally synchronized motion edits and delivering state-of-the-art TMME performance.

2 Related Work

3D Human Motion Generation. *Single-Person Human Motion Generation:* Text-driven single-person motion generation [3, 11, 21, 22, 27, 43, 53, 58, 59] has rapidly advanced with diffusion models becoming the dominant paradigm. Representative works include MDM [42], VAE [5,33,34,51], auto-regressive normalization network [13], GAN [1,4,8,50,52], motion diffusion [18,39,54], latent-space diffusion [6, 19, 53] for improved efficiency and fidelity. Beyond direct text conditioning, retrieval-augmented generation further improves text–motion alignment and diversity by incorporating nearest-neighbor motion candidates [55]. Large-scale datasets and stronger text encoders (*e.g.*, CLIP [38]) have also contributed to better language understanding and more expressive motion synthesis. *Multi-Person Motion Generation:* Generating human-human interactions is

more challenging than single-person synthesis due to mutual dependencies and spatio-temporal coupling [17, 24, 46]. Multi-person motion generation is enabled by the proposal of two-person motion datasets [17,24]. InterGen [24] introduces a large-scale interaction dataset and a diffusion model for joint denoising of both participants. In2IN [40] enhances controllability by leveraging individual-level information, while InterControl [48] explores structured control for interaction synthesis. InterMask [17] improves spatio-temporal consistency via masked prediction, and TIMotion [46] emphasizes temporal and interactive dynamics. Recent methods like InterMoE [44] and HINT [25] focus on individual-specific and hierarchical modeling, respectively. However, existing works primarily focus on 3D multi-human generation based on text alone.

Text-Guided Single-Person 3D Motion Editing. Previous work on 3D human motion editing focused on spatial or temporal constraints [9, 10, 20], with recent deep learning methods balancing instruction adherence and source preservation [2, 12, 15, 16, 29, 45, 46, 56, 57]. MotionFix [2] uses diffusion for text-driven single-person editing, but extending this to dual-person interactions is challenging due to the need to preserve realism and spatio-temporal coupling. Similarity-aware methods like SimMotionEdit [23] limit deviation from the source, but multi-person editing is still underexplored. Unlike prior work on human-to-human interaction generation [24,40,46] and single-person editing [2], we explore *multi-person motion editing*, conditioning on a multi-person source motion and instruction. We propose semantic-aware plan token alignment and interaction-aware frequency token alignment to ensure precise motion edits while preserving temporal coordination and minimizing drift in multi-human interactions.

3 Dataset and Benchmark

Dataset: InterEdit3D. We present **InterEdit3D**, a multi-human 3D motion editing dataset composed of (source motion, target motion, edit text) triplets. It is built upon InterHuman [24], a large-scale 3D human–human interaction dataset featuring diverse two-person interactions with natural language descriptions, *e.g.*, daily activities (greeting, handshaking, object passing) and professional interactions (martial arts, dancing). To construct editing pairs, we retrieve motion pairs that share similar motion backbones but differ in interaction semantics via motion-to-motion retrieval. Each pair is further annotated with an instruction that describes how to transform the source motion into the target motion, forming high-quality supervision for multi-person motion editing.

Dataset Comparison. Table 1 compares our dataset with others across *text annotation*, *interaction*, and *editing*. Classical datasets like KIT-ML [36], BABEL [37], and HumanML3D [11] focus on single-person motions but lack multi-person interactions or editing pairs. PoseFix [7] targets pose correction, while MotionFix [2] enables single-person motion editing but lacks interactions. Interaction datasets like InterHuman [24] and Inter-X [49] focus on two-person motion generation but don't offer editing triplets. Our dataset uniquely combines text, interaction, and editing, offering paired source-target motions with edit instruc-

Dataset	Text	Interactive	Editing	Motions	Vocab.
KIT-ML [36]	✓	–	–	3,911	1,623
BABEL [37]	✓	–	–	10,881	1,347
HumanML3D [11]	✓	–	–	14,616	5,371
PoseFix [7]	✓	–	✓	6,157 (pairs)	1,068
MotionFix [2]	✓	–	✓	6,730 (pairs)	1,479
InterHuman [24]	✓	✓	–	6,022	5,656
Inter-X [49]	✓	✓	–	11,388	3,467
Ours	✓	✓	✓	**5,161 (pairs)**	**1754**

Table 1: Comparison with representative existing datasets. "Editing" indicates whether the dataset provides source–target pairs and edit instructions.

tions for two-person interactions. This enables evaluation of models that preserve consistency and follow instructions without disrupting interactions.

Data Collection and Annotation. We convert the motion format provided by InterHuman [24] into AMASS features [28] and encode them using the pretrained TMR [35] motion encoder, which provides semantically meaningful embeddings via contrastive motion–text alignment. Motion-to-motion retrieval is performed in this latent space. We retrieve the top-2 nearest neighbors (cosine similarity), forming source–target candidates that share a similar base motion for one individual but differ in interaction semantics, enabling "edit the change, preserve the rest." Annotators write an *edit instruction* describing how to transform the source into the target. Subtle or unclear pairs are discarded. In total, we obtain 5,161 triplets, split into train/validation/test sets with an 80%/10%/10% ratio. We perform interaction-level disjoint split by manually checking to ensure no overlapping interaction identities among different sets. The dataset is annotated by 8 individuals, and a cross-checking process is performed to ensure consistency in the annotations. Samples with significantly divergent annotations are discarded. Overall, **InterEdit3D** emphasizes relative spatial control, temporal ordering, semantic action changes, and body-part-aware interaction edits, forming a realistic and challenging benchmark for dual-person motion editing.

Baselines. We compare with 4 representative methods from the two closest settings: single-person motion editing models MotionFix [2] and MotionLab [12], and multi-human motion generation models InterGen [24] and TIMotion [46]. Since no prior method is specifically designed for text-guided dyadic motion editing with source–target–instruction triplets, we adapt these methods and retrain them on **InterEdit3D**. All baselines use the same motion representation, data splits, and evaluation metrics as **InterEdit**. Single-Person Editing Baselines. We concatenate both individuals' motion features along the feature dimension, and train the models to predict the edited two-person target motion from the source motion and editing instruction. Multi-Human Motion Generation Baselines. We preserve their original interaction modeling architectures and inject the source motion as an additional condition via AdaLN, enabling motion generation conditioned on both the source motion and instruction.

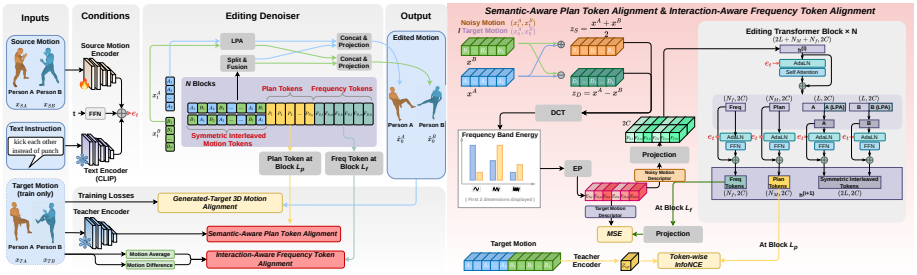


Fig. 2: Overview of the proposed **InterEdit** framework. Given a two-person motion and an editing instruction, **InterEdit** uses a conditional diffusion backbone with symmetric interleaved motion tokens. It introduces (i) *Semantic-Aware Plan Token Alignment* for high-level editing guidance via a motion-teacher embedding, and (ii) *Interaction-Aware Frequency Token Alignment* using DCT-based band-energy descriptors to regulate interaction dynamics.

4 Methodology

As shown in Fig. 2, we formulate two-person motion editing as a conditional diffusion model, where the denoiser is conditioned on the source motion and the text instruction, further guided by semantic-aware plan token alignment and interaction-aware frequency token alignment.

4.1 Multi-Human Motion Representation

We follow the non-canonical motion representation commonly used for two-person 3D motion data following InterHuman [24]. For each person $p \in \{A, B\}$ at frame ℓ, the motion state is as Eq. 1.

$$\mathbf{x}_\ell^p = \left[\mathbf{j}_{g,\ell}^p,\ \mathbf{v}_{g,\ell}^p,\ \mathbf{r}_\ell^p,\ \mathbf{c}_\ell^p\right] \in \mathbb{R}^{d_m}, \tag{1}$$

where $\mathbf{j}_{g,\ell}^p \in \mathbb{R}^{3N_j}$ and $\mathbf{v}_{g,\ell}^p \in \mathbb{R}^{3N_j}$ denote global joint positions and velocities in the world frame, $\mathbf{r}_\ell^p \in \mathbb{R}^{6N_j}$ is the 6D local joint rotation representation in the root frame, N_j denotes the number of joints per person, and $\mathbf{c}_\ell^p \in \mathbb{R}^4$ is the binary foot–ground contact. A two-person motion sequence is denoted as Eq. 2.

$$\mathbf{x}_{1:L} = \left(\mathbf{x}_{1:L}^A, \mathbf{x}_{1:L}^B\right) \in \mathbb{R}^{L \times 2d_m}, \tag{2}$$

where L denotes the motion length (number of frames). Given a two-person source motion sequence $\mathbf{x}_{1:L}^s$ and an edit instruction text $\mathbf{y}$, our goal is to generate an edited two-person target motion $\hat{\mathbf{x}}_{1:L}$ that follows the instruction while preserving source-consistent content.

4.2 Multi-Human Conditional Diffusion for Motion Editing

We formulate multi-human motion editing as conditional generation with a diffusion model $p_\theta(\mathbf{x}_0 \mid \mathbf{x}^s, \mathbf{y})$, where $\mathbf{x}^s$ is the source motion sequence, $\mathbf{y}$ is the

editing instruction, and $\mathbf{x}_0$ denotes the clean target motion sequence. Let $\mathbf{x}_t$ denote the noisy version of $\mathbf{x}_0$ at diffusion step $t \in \{1, \ldots, T\}$. Following a noise schedule $\{\beta_t\}_{t=1}^T$, the forward noising process is as Eq. 3.

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\Big(\sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I}\Big), \qquad q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Equivalently, we can sample according to Eq. 4.

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

Denoising Model and Conditioning. We use a transformer denoiser $\mathcal{D}_\theta$ with START_X parameterization, i.e., it directly predicts the clean motion $\hat{\mathbf{x}}_0$ as shown in Eq. 5 (rather than the added noise $\boldsymbol{\epsilon}$) [14]; this prediction is used to form the reverse-step distribution (and is compatible with DDIM [41] sampling).

$$\hat{\mathbf{x}}_0 = \mathcal{D}_\theta(\mathbf{x}_t, t;\ \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{src}}), \tag{5}$$

where θ denotes trainable parameters, $\mathbf{c}_{\text{text}}$ is the instruction embedding extracted through a frozen CLIP textual encoder [38] via $\mathbf{c}_{\text{text}} = \text{CLIP}(\mathbf{y})$, and $\mathbf{c}_{\text{src}}$ is the source-motion embedding. We encode the source motion $\mathbf{x}^s$ with a lightweight learnable Transformer source-motion encoder to obtain $\mathbf{c}_{\text{src}}$; for this encoding we remove the last 4 foot-contact channels per person and concatenate two persons along the feature dimension, then project it into a latent space, prepend a learnable query (CLS) token, and apply positional encoding followed by a multi-layer Transformer encoder; the source-motion embedding is taken from the output of the query token. Both conditions are injected into the denoiser via AdaLN modulation as Eq. 6.

$$\mathbf{e}_t = \text{EmbedTime}(t) + W_{\text{text}}\mathbf{c}_{\text{text}} + W_{\text{src}}\mathbf{c}_{\text{src}}, \tag{6}$$

where $\text{EmbedTime}(\cdot)$ is a sinusoidal timestep embedding followed by an MLP projection, and W_{text}, W_{src} are linear projections.

Symmetric Interleaved Token Aggregation. To model temporal-order influence and role switching in two-person interactions, we follow TIMotion [46] and construct a symmetric interleaved sequence rather than directly concatenating the two persons into a single token stream. Let $\mathbf{x}_c^A, \mathbf{x}_c^B \in \mathbb{R}^{L\times C}$ be the motion token sequences for two persons, where L is the length of the sequence and C is the dimension of motion embedding. We build a causal interleaving $\mathbf{x}_{\text{cii}} \in \mathbb{R}^{2L\times C}$ and its role-swapped counterpart $\mathbf{x}_{\text{sym}} \in \mathbb{R}^{2L\times C}$ as Eq. 7.

$$\begin{aligned} \mathbf{x}_{\text{cii}}(2\ell{-}1) &= \mathbf{x}_c^A(\ell), \quad & \mathbf{x}_{\text{cii}}(2\ell) &= \mathbf{x}_c^B(\ell), \\ \mathbf{x}_{\text{sym}}(2\ell{-}1) &= \mathbf{x}_c^B(\ell), \quad & \mathbf{x}_{\text{sym}}(2\ell) &= \mathbf{x}_c^A(\ell). \end{aligned} \tag{7}$$

We then feed the concatenation result $\mathbf{x}_{inter} = \text{Concat}(\mathbf{x}_{\text{cii}}, \mathbf{x}_{\text{sym}}) \in \mathbb{R}^{2L\times 2C}$ into a transformer. Given an output $\hat{\mathbf{x}}_{inter} \in \mathbb{R}^{2L\times 2C}$, we split channels $\hat{\mathbf{x}}_{(inter,1)} = \hat{\mathbf{x}}_{inter}[:, :C]$, $\hat{\mathbf{x}}_{(inter,2)} = \hat{\mathbf{x}}_{inter}[:, C:]$ and de-interleave odd/even indices to recover per-person streams, then merge the two role views by element-wise sum

to obtain $\hat{\mathbf{x}}_c^{A,g}, \hat{\mathbf{x}}_c^{B,g} \in \mathbb{R}^{L\times C}$. We further apply a lightweight Localized Pattern Amplification (LPA) [46] branch to refine short-range temporal patterns for each person. Given the per-person motion token sequence, $\mathbf{x}_c^p \in \mathbb{R}^{L\times C}$, we modulate it with the condition embedding $\mathbf{e}_t$ via AdaLN and extract local features with a small 1D convolutional stack as Eq. 8.

$$\hat{\mathbf{x}}_c^p = \mathbf{x}_c^p + \mathrm{Conv}_1(\mathrm{AdaLN}(\mathrm{Conv}_3(\mathrm{AdaLN}(\mathbf{x}_c^p, \mathbf{e}_t)), \mathbf{e}_t)), \tag{8}$$

where Conv_k denotes a 1D convolution with kernel size k. We then fuse global and local features by channel-wise concatenation followed by a linear projection back to C channels as Eq. 9.

$$\mathbf{x}_c^{p,f} = \mathrm{Linear}\big(\mathrm{Concat}\big(\hat{\mathbf{x}}_c^{p,g}, \hat{\mathbf{x}}_c^p\big)\big), \quad p \in \{A, B\}. \tag{9}$$

Then the final outputs are fed into the next transformer block until the final decoder layer to produce the denoiser prediction.

Training objective. With START_X parameterization, we train the denoiser $\mathcal{D}_\theta$ by minimizing the reconstruction error of $\mathbf{x}_0$ as Eq. 10.

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[\big\|\mathbf{x}_0 - \mathcal{D}_\theta(\mathbf{x}_t, t; \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{src}})\big\|_2^2\Big]. \tag{10}$$

4.3 Synchronized Classifier-Free Guidance for Editing

We introduce Synchronized Classifier-Free Guidance (SCFG) for controllable editing. During training, we randomly drop conditioning with probability p_{scfg} by setting both $\mathbf{c}_{\text{text}}$ and $\mathbf{c}_{\text{src}}$ to zeros in a synchronized way. This results in an unconditional branch. At inference, we combine the conditional and unconditional predictions as Eq. 11.

$$\hat{\mathbf{x}}_{0,\text{scfg}} = \gamma\,\mathcal{D}_\theta(\mathbf{x}_t, t; \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{src}}) + (1-\gamma)\,\mathcal{D}_\theta(\mathbf{x}_t, t; \mathbf{0}, \mathbf{0}), \tag{11}$$

which is equivalent to $\hat{\mathbf{x}}_{0,\text{scfg}} = \mathcal{D}_\theta(\cdot; \mathbf{0}, \mathbf{0}) + \gamma\big(\mathcal{D}_\theta(\cdot; \mathbf{c}_{\text{text}}, \mathbf{c}_{\text{src}}) - \mathcal{D}_\theta(\cdot; \mathbf{0}, \mathbf{0})\big)$. Here γ is the guidance scale. Dropping both conditions avoids leakage (source/text bias) and yields a cleaner guidance direction.

4.4 Semantic-Aware Plan Token Alignment

Multi-human editing requires identifying *what to change* while preserving non-edited content. We introduce learnable **plan tokens** to provide semantic-level guidance.

Plan Tokens as Control Tokens. We append N_M learnable plan tokens $\mathbf{P} \in \mathbb{R}^{N_M\times 2C}$ to the denoiser token sequence. Through self-attention, motion tokens can query $\mathbf{P}$ and receive global guidance during denoising. At transformer block L_p, we project plan tokens into a semantic space as Eq. 12.

$$\hat{\mathbf{z}}^{(k)} = g_p\Big(\mathbf{P}_k^{(L_p)}\Big) \in \mathbb{R}^C, \quad k = [1, \ldots, N_M], \tag{12}$$

where g_p is a combination of layer normalization and linear projection.

Teacher Target Embedding. We use a frozen motion teacher encoder $f_T(\cdot)$ (trained contrastively on InterHuman [24] motion–text pairs) to extract a target semantic embedding from the ground-truth target motion as Eq. 13.

$$\mathbf{z}_{\text{tgt}} = f_T(\mathbf{x}_0) \in \mathbb{R}^C. \tag{13}$$

This teacher embedding provides a compact semantic target for aligning plan tokens.

Plan Token Alignment Loss. We apply a token-wise InfoNCE objective with $\mathbf{z}_{\text{tgt}}$ as the positive target and in-batch negatives (n indexes targets in the mini-batch). Let $\tilde{\mathbf{z}}_{\text{tgt}} = \text{norm}(\mathbf{z}_{\text{tgt}})$ and $\tilde{\mathbf{z}}^{(k)} = \text{norm}(\hat{\mathbf{z}}^{(k)})$. The plan loss is shown in Eq. 14.

$$\mathcal{L}_{\text{plan}} = \frac{1}{N_M} \sum_{k=1}^{N_M} \left[-\log \frac{\exp\left((\tilde{\mathbf{z}}^{(k)})^\top \tilde{\mathbf{z}}_{\text{tgt}}/\tau\right)}{\sum_n \exp\left((\tilde{\mathbf{z}}^{(k)})^\top \tilde{\mathbf{z}}_{\text{tgt}}^{(n)}/\tau\right)} \right], \tag{14}$$

where τ is the temperature. We also experimented with cosine and MSE variants, but InfoNCE [30] provides the best overall trade-off.

4.5 Interaction-Aware Frequency Token Alignment

Interaction correctness is highly sensitive to temporal coupling (synchrony vs. alternation, phase alignment, contact timing). We propose Interaction-Aware Frequency Token Alignment, which leverages Discrete Cosine Transformation (DCT) and energy pooling to capture and regularize the frequency dynamics of interaction signals, ensuring precise synchronization and coordination.

Interaction Signals: Average and Difference. We exclude foot-contact channels (last 4 dims per person) and denote the remaining feature dimension as d_f. For a two-person motion sequence $\mathbf{x} = (\mathbf{x}^A, \mathbf{x}^B)$, we construct two interaction-aware sequences as Eq. 15.

$$\mathbf{z}_S = \frac{\mathbf{x}^A + \mathbf{x}^B}{2} \in \mathbb{R}^{L \times d_f}, \qquad \mathbf{z}_D = \mathbf{x}^A - \mathbf{x}^B \in \mathbb{R}^{L \times d_f}, \tag{15}$$

where $\mathbf{z}_S$ captures shared or synchronized components, $\mathbf{z}_D$ captures relative or oppositional components.

Discrete Cosine Transformation (DCT) and Band-Energy Pooling. We apply discrete cosine transformation along the time axis as Eq. 16.

$$\mathbf{C}_S = \text{DCT}(\mathbf{z}_S), \qquad \mathbf{C}_D = \text{DCT}(\mathbf{z}_D). \tag{16}$$

We compute band-energy descriptors for low/mid/high frequency bins by pooling squared coefficients within each band b as Eq. 17.

$$\mathbf{E}(\mathbf{C}; b) = \sqrt{\frac{1}{|b|} \sum_{k \in b} \mathbf{C}[k]^2 + \epsilon} \in \mathbb{R}^{d_f}, \tag{17}$$

where ϵ is a small constant added for numerical stability. Using normalized cutoffs $(r_{\mathrm{l}}, r_{\mathrm{m}}, r_{\mathrm{h}})$ to define the three bins (low (l), middle (m), and high (h)), we obtain six band-energy descriptors according to Eq. 18.

$$\mathbf{g}(\mathbf{x}) = \Big[\mathbf{E}(\mathbf{C}_S;\mathrm{l}), \mathbf{E}(\mathbf{C}_S;\mathrm{m}), \mathbf{E}(\mathbf{C}_S;\mathrm{h}), \mathbf{E}(\mathbf{C}_D;\mathrm{l}), \mathbf{E}(\mathbf{C}_D;\mathrm{m}), \mathbf{E}(\mathbf{C}_D;\mathrm{h})\Big] \in \mathbb{R}^{6\times d_f}, \tag{18}$$

where $\mathbf{g}(\mathbf{x})$ is computed from $\mathbf{x}$ through $(\mathbf{z}_S, \mathbf{z}_D)$ construction and DCT.

Interaction-Aware Frequency Tokens and Regression Alignment. At each diffusion step, we compute band-energy descriptors from the current noisy motion $\mathbf{x}_t$ and project them into the model token space to form N_f=6 frequency control tokens according to Eq. 19.

$$\mathbf{F}_i = \phi_f^{(i)}(\mathbf{g}_i(\mathbf{x}_t)) \in \mathbb{R}^{2C}, \quad i = [1, \ldots, N_f], \tag{19}$$

where $\phi_f^{(i)}$ is a lightweight network consisting of a layer normalization and a linear projection layer, and $\mathbf{F} = [\mathbf{F}_1; \ldots; \mathbf{F}_{N_f}]$ is appended to the motion token sequence and plan token for joint self-attention. At transformer block L_f, we decode the band-energy descriptors from the corresponding token as Eq. 20.

$$\hat{\mathbf{g}}_i = g_f^{(i)}\left(\mathbf{F}_i^{(L_f)}\right) \in \mathbb{R}^{d_f}, \quad i = [1, \ldots, N_f], \tag{20}$$

where $g_f^{(i)}$ is a combination of layer normalization and linear projection head. We minimize a weighted regression loss against the ground-truth target band-energy descriptors computed from the clean target motion $\mathbf{x}_0$ according to Eq. 21.

$$\mathcal{L}_{\mathrm{freq}} = \frac{1}{N_f}\sum_{i=1}^{N_f} w_i \left\|\hat{\mathbf{g}}_i - \mathbf{g}_i(\mathbf{x}_0)\right\|_2^2, \tag{21}$$

where $w_i \geq 0$ is a per-term weight. We down-weight the two high-frequency terms (the "high" bins for both $\mathbf{z}_S$ and $\mathbf{z}_D$) with a smaller constant weight to reduce sensitivity to high-frequency noise; for all other bins we set $w_i = 1$. We further apply frequency-token dropout during training: with probability p_f, we remove frequency tokens from the denoiser, which regularizes training and helps preserve generation quality.

4.6 Objective Function

Diffusion Reconstruction Loss: The **diffusion reconstruction loss** ($\mathcal{L}_{\mathrm{diff}}$) minimizes the MSE between the predicted motion $\hat{\mathbf{x}}_0$ and the ground-truth motion $\mathbf{x}_0$, ensuring that the denoised motion closely matches the original.

Geometric Losses (Per Person): We apply several geometric losses to preserve realism in each person's motion. The **velocity loss** ($\mathcal{L}_{\mathrm{vel}}$) ensures smooth movement by penalizing the difference between predicted and ground-truth velocities. The **foot-contact loss** ($\mathcal{L}_{\mathrm{foot}}$) ensures realistic foot placement. The **bone-length loss** ($\mathcal{L}_{\mathrm{BL}}$) maintains consistent bone lengths by comparing per-bone distances between predicted and ground-truth joint positions.

Interaction Losses (Between Persons): For realistic interaction, we introduce interaction losses. The **masked distance-map loss** ($\mathcal{L}_{\text{DM}}$) focuses on joint distances in contact regions. The **relative-orientation loss** ($\mathcal{L}_{\text{RO}}$) ensures coherent interaction by penalizing orientation mismatches. The motion objective is shown in Eq. 22.

$$\begin{aligned}\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}} + \lambda_{\text{BL}}\mathcal{L}_{\text{BL}} \\ + \lambda_{\text{DM}}\mathcal{L}_{\text{DM}} + \lambda_{\text{RO}}\mathcal{L}_{\text{RO}}.\end{aligned} \tag{22}$$

Auxiliary Alignment Terms: To further improve the motion generation, we add two auxiliary alignment terms. The **plan token alignment** ($\mathcal{L}_{\text{plan}}$) aligns the learnable plan tokens with the teacher embedding of the ground-truth target, guiding the model to capture and represent the overall motion plan accurately. The **frequency token alignment** ($\mathcal{L}_{\text{freq}}$) aligns the frequency control tokens with the DCT band-energy descriptors derived from interaction signals, preserving the frequency characteristics of the motion and ensuring that the rhythmic aspects of the movement are preserved. The final training objective is shown in Eq. 23.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{motion}} + \lambda_p\mathcal{L}_{\text{plan}} + \lambda_f\mathcal{L}_{\text{freq}}. \tag{23}$$

In this approach, the CLIP text encoder [38] and the motion teacher are kept frozen, and all other parameters are trained end-to-end.

5 Experiments

5.1 Implementation Details

We implement **InterEdit** with $N = 5$ transformer blocks, each with 16 attention heads and motion embedding dimension 512. The text encoder is a frozen CLIP ViT-L/14 [38]. As the motion teacher, we adopt a frozen contrastively trained motion encoder on InterHuman [24] to provide semantic target embeddings for plan-token alignment. We use $N_M = 16$ plan tokens. For auxiliary losses, $\lambda_p = 0.03$ and $\lambda_f = 0.01$, applied at blocks $L_p = 3$ and $L_f = 5$. High-frequency components are down-weighted by 0.25. Three DCT bands are used with $r_{\text{low}} = 0.08$, $r_{\text{mid}} = 0.25$, and $r_{\text{high}} = 0.35$, and frequency tokens are randomly dropped with probability $p_f = 0.04$. Other geometric and interaction losses follow InterGen [24]. Training uses 1000 diffusion steps with a cosine schedule. At inference we adopt DDIM sampling (50 steps, $\eta = 0$) with classifier-free guidance (drop rate $p_{\text{scfg}} = 0.1$, guidance scale $\gamma = 3.5$). Optimization uses AdamW [26] ($\beta_1 = 0.9$, $\beta_2 = 0.999$), weight decay 2×10^{-5}, and peak learning rate 10^{-4} with cosine decay and 10 warm-up epochs. The full model contains 358.8M parameters including frozen encoders, with 85.0M trainable parameters, close to TIMotion's 81.2M and substantially smaller than InterGen's 251M. We train the model for 1500 epochs with batch size 32 on 8 NVIDIA RTX Pro 6000 Blackwell GPUs in about 4h35m, with about 16GB peak GPU memory. At inference, InterEdit achieves 2.842 samples/s, comparable to TIMotion's 2.709 samples/s.

Method	FID ↓	generated-to-source retrieval (%) ↑			**generated-to-target retrieval (%) ↑**		
		R@1	R@2	R@3	**R@1**	**R@2**	**R@3**
MotionFix [2]	2.1089±0.0043	3.10±0.39	7.20±0.34	9.10±0.49	11.60±0.52	17.80±0.47	21.10±0.44
MotionLab [12]	0.4284±0.0121	12.40±0.69	18.90±0.79	23.00±0.38	18.50±0.79	25.60±0.50	30.50±0.44
InterGen [24]	0.6243±0.0062	9.52±0.42	15.03±0.44	18.91±0.47	18.93±0.60	26.74±0.71	31.64±0.72
TIMotion [46]	0.4451±0.0058	12.54±0.34	18.38±0.39	22.33±0.46	24.97±0.59	33.77±0.56	40.68±0.65
Ours	**0.3707±0.0029**	**17.08±0.41**	**24.04±0.46**	**29.32±0.49**	**30.82±0.43**	**40.84±0.70**	**47.65±0.59**

Table 2: Quantitative comparison (mean, 95% CI).

Method	R-Precision@1↑	R-Precision@2↑	R-Precision@3↑
InterGen [24]	0.371±0.010	0.515±0.012	0.624±0.010
TIMotion [46]	0.491±0.005	0.648±0.004	0.724±0.004
Ours	**0.523±0.004**	**0.665±0.005**	**0.753±0.004**

Table 3: Generalization on the InterHuman generation benchmark (mean, 95% CI).

Method	FID ↓	generated-to-source retrieval (%) ↑			**generated-to-target retrieval (%) ↑**		
		R@1	R@2	R@3	**R@1**	**R@2**	**R@3**
without plan/freq token	0.4451±0.0058	12.54±0.34	18.38±0.39	22.33±0.46	24.97±0.59	33.77±0.56	40.68±0.65
only plan token($L_p = 5$)	0.3667±0.0023	14.52±0.36	21.58±0.31	25.87±0.32	28.72±0.49	37.92±0.33	43.50±0.44
only freq token($L_f = 5$)	0.3798±0.0027	14.24±0.34	20.91±0.39	25.48±0.40	28.75±0.47	38.46±0.48	44.05±0.43
with plan&freq token	0.3707±0.0029	17.08±0.41	24.04±0.46	29.32±0.49	30.82±0.43	40.84±0.70	47.65±0.59

Table 4: Ablation of module component (mean, 95% CI).

Method	FID ↓	generated-to-source retrieval (%) ↑			**generated-to-target retrieval (%) ↑**		
		R@1	R@2	R@3	**R@1**	**R@2**	**R@3**
$p_f = 0.05$	0.3477±0.0020	16.11±0.37	23.14±0.50	27.69±0.53	29.10±0.61	39.46±0.65	45.25±0.62
$p_f = 0.04$	0.3655±0.0031	16.33±0.45	22.99±0.38	27.07±0.38	30.51±0.45	40.10±0.32	45.97±0.43
$p_f = 0.03$	0.3693±0.0029	15.20±0.31	21.57±0.39	25.97±0.38	29.70±0.44	40.25±0.49	46.48±0.44

Table 5: Ablation of frequency tokens dropout rate ($L_p = 5$, $L_f = 5$, mean, 95% CI).

5.2 Analysis of the Benchmark

Metrics. We use retrieval-based metrics as primary measures following MotionFix [2]. For a 3D multi-human motion editing result, we evaluate: (i) generated-to-target retrieval (g2t) and (ii) generated-to-source retrieval (g2s) in a learned motion embedding space, reporting Recall@K ($K \in 1, 2, 3$). We use the InterGen text-to-motion retrieval model [24] as the feature extractor, where motions are L2-normalized and ranked by cosine similarity against the full test set. g2t measures instruction adherence, while g2s reflects source preservation. A strong editor should achieve high g2t with reasonably high g2s, balancing semantic modification and content preservation. We also report FID in the same embedding space to assess motion realism, measuring the distribution distance between generated and real target motions (lower is better). All methods are evaluated for 20 independent runs, and we report the mean with 95% confidence intervals.

Benchmark Results and Baseline Comparison. Table 2 presents quantita-

tive comparisons on our dataset with four adapted baselines: single-person editing (MotionFix [2], MotionLab [12]) and two-person generation (InterGen [24], TIMotion [46]). Our method outperforms all baselines in g2t and g2s, and achieves the lowest FID, demonstrating superior instruction adherence, edit fidelity, and motion realism. Compared to the strongest baseline TIMotion [46], InterEdit improves g2t R@1/2/3 by +5.85, +7.07, and +6.97, improves g2s R@1/2/3 by +4.54, +5.66, and +6.99, and reduces FID by 16.7%, indicating better edits without disrupting interaction coupling. The adapted single-person editors lag clearly behind in g2t, highlighting the importance of interaction-aware modeling for dual-person motions. Two-person generation baselines achieve better g2t than adapted single-person editors, but g2s remains limited, since they are not designed to balance instruction following with source preservation, which can lead to global drift or disrupted interaction consistency. In contrast, **InterEdit** combines semantic-aware plan token alignment and interaction-aware frequency token alignment to better capture editing intent and temporal coordination, producing more faithful and realistic dyadic motion edits.

Human Preference Evaluation. To complement automatic metrics, we conduct a human preference study against TIMotion. InterEdit is preferred by 75.5%, 78.5%, 71.0%, and 81.0% on overall preference, instruction adherence, source preservation, and interaction realism, respectively, further confirming its advantages in faithful and realistic interaction editing.

Generalization. Beyond our TMME benchmark in Table 2, we further evaluate InterEdit on the standard InterHuman two-person generation benchmark, where motions are generated from text only. As shown in Table 3, InterEdit achieves the best R-Precision across all ranks, suggesting that our token alignment design also benefits two-person motion modeling beyond editing.

5.3 Analysis of the Ablation Study

Effect of the two main modules. Table 4 evaluates our core modules, including semantic-aware latent representation alignment and interaction-oriented frequency alignment, individually and jointly. Removing both modules ("without plan/freq token") results in the weakest performance across g2s/g2t retrieval and FID, indicating that the base diffusion model alone struggles to preserve source content and capture target semantics. Introducing only plan tokens improves both retrieval metrics and motion quality, showing that semantic alignment provides a global editing signal. Using only frequency tokens enhances retrieval and maintains competitive FID, emphasizing the importance of regularizing interaction dynamics. Combining both plan and frequency tokens yields the best results, highlighting their complementary roles: plan tokens guide what to edit semantically, while frequency tokens stabilize interaction dynamics.

Randomly dropping frequency tokens during training. Table 5 studies random frequency-token dropout (p_f) where we stochastically disable the frequency-token pathway during training. We observe that a moderate drop rate provides the best balance: it improves retrieval while keeping FID favorable. Intuitively, this stochastic dropping regularizes training by preventing the model

from over-relying on the frequency tokens, encouraging robust generation quality even when frequency guidance is partially absent. Too small a drop rate may lead to over-dependence, while too large a drop rate weakens the intended frequency alignment signal; both cases yield inferior overall trade-offs compared to the moderate setting. More ablation studies are shown in the supplementary.

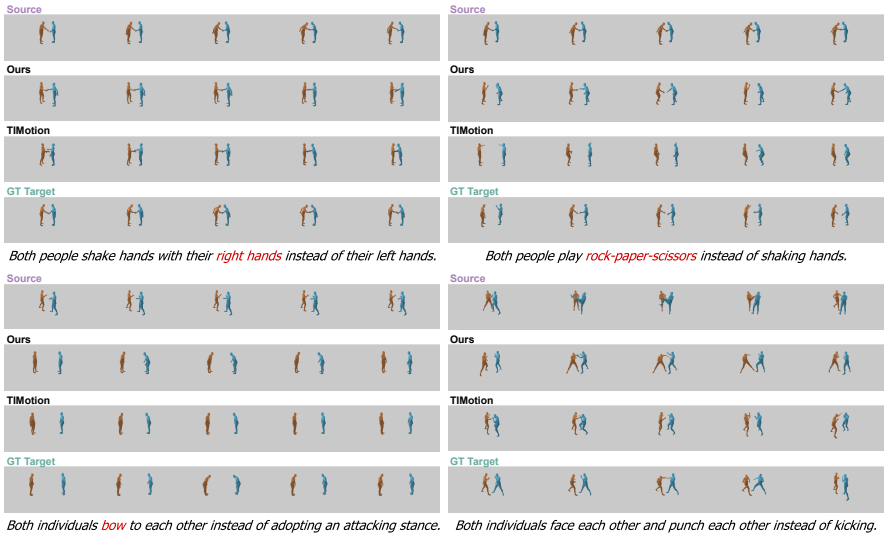


Fig. 3: Qualitative results comparison of our **InterEdit** and TIMotion [46].

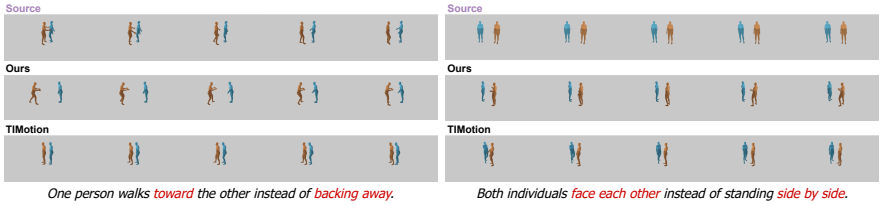


Fig. 4: Qualitative results comparison under custom prompts.

5.4 Analysis of the Qualitative Results

We present qualitative results on test-set prompts (Fig. 3) and custom prompts (Fig. 4), comparing our results with the TIMotion [46] baseline. The examples cover action-category changes, temporal coordination, and spatial relations.

"Rock-paper-scissors instead of shaking hands." requires both semantic switch and correct temporal coupling. Our result exhibits synchronized throws, while TIMotion produces offset gestures. This highlights the advantage of regularizing interaction dynamics to enforce synchrony. For *"Both individuals face each other and punch each other instead of kicking."*, our result shows an asymmetric attack-defense coupling, matching the target's intent. TIMotion produces a symmetric response, illustrating the importance of capturing high-level semantics in dual-person editing. For *"One person walks toward the other instead of backing away."*, our model accurately reflects the intended approach behavior, while TIMotion fails to capture the directional change, demonstrating the value of semantic-level guidance.

6 Conclusion

We introduce the task of **text-guided multi-human motion editing**, where a model applies instruction-driven changes while preserving source consistency and spatio-temporal coupling. To support this task, we construct a dataset of source-target-edit triplets and establish a benchmark with retrieval-based metrics. We introduce **InterEdit3D** dataset and propose **InterEdit**, a conditional diffusion framework that uses plan tokens for semantic alignment and frequency tokens for interaction-aware DCT descriptors, ensuring precise edits while preserving interaction dynamics. Our method outperforms single-person editors and interaction generators in instruction adherence, source preservation, and motion realism, providing a foundation for future research in interaction editing and long-horizon motion modifications.

Acknowledgment

The project served to prepare the SFB 1574 Circular Factory for the Perpetual Product (project ID: 471687386), approved by the German Research Foundation (DFG, German Research Foundation) with a start date of April 1, 2024. This work was also partially supported by the SmartAge project sponsored by the Carl Zeiss Stiftung (P2019-01-003; 2021-2026). This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. This project was also supported partially by the National Natural Science Foundation of China under Grant No. 62473139, partially by the Hunan Provincial Research and Development Project under Grant No. 2025QK3019, and partially by the Open Research Project of the State Key Laboratory of Industrial Control Technology, China, under Grant No. ICT2025B20.

References

1. Amballa, A., Akkinapalli, G., Muralikrishnan, V.: LS-GAN: Human motion synthesis with latent-space GANs. In: WACVW (2025)
2. Athanasiou, N., Ceske, A., Diomataris, M., Black, M.J., Varol, G.: MotionFix: Text-driven 3D human motion editing. In: SIGGRAPH Asia (2024)
3. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-An-Animation: Large-scale text-conditional 3D human motion generation. In: ICCV (2023)
4. Barsoum, E., Kender, J., Liu, Z.: HP-GAN: Probabilistic 3D human motion prediction via GAN. In: CVPRW (2018)
5. Bie, X., Guo, W., Leglaive, S., Girin, L., Moreno-Noguer, F., Alameda-Pineda, X.: HIT-DVAE: Human motion generation via hierarchical transformer dynamical VAE. arXiv preprint arXiv:2204.01565 (2022)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
7. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: PoseFix: Correcting 3D human poses with natural language. In: ICCV (2023)
8. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: ICCV (2021)
9. Gleicher, M.: Motion editing with spacetime constraints. In: I3D (1997)
10. Gleicher, M.: Motion path editing. In: I3D (2001)
11. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3D human motions from text. In: CVPR (2022)
12. Guo, Z., Hu, Z., Soh, D.W., Zhao, N.: MotionLab: Unified human motion generation and editing via the motion-condition-motion paradigm. In: ICCV (2025)
13. Henter, G.E., Alexanderson, S., Beskow, J.: MoGlow: Probabilistic and controllable motion synthesis using normalising flows. TOG (2020)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
15. Hong, S., Kim, C., Yoon, S., Nam, J., Cha, S., Noh, J.: SALAD: Skeleton-aware latent diffusion for text-driven motion generation and editing. In: CVPR (2025)
16. Hu, V.T., Yin, W., Ma, P., Chen, Y., Fernando, B., Asano, Y.M., Gavves, E., Mettes, P., Ommer, B., Snoek, C.G.: Motion flow matching for human motion synthesis and editing. arXiv preprint arXiv:2312.0889 (2023)
17. Javed, M.G., Guo, C., Cheng, L., Li, X.: InterMask: 3D human interaction generation via collaborative masked modelling. In: ICLR (2025)
18. Kapon, R., Tevet, G., Cohen-Or, D., Bermano, A.H.: MAS: Multi-view ancestral sampling for 3D motion generation using 2D diffusion. In: CVPR (2024)
19. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: ICCV (2023)
20. Lee, J., Shin, S.Y.: A hierarchical approach to interactive motion editing for human-like figures. In: SIGGRAPH (1999)
21. Lee, T., Baradel, F., Lucas, T., Lee, K.M., Rogez, G.: T2LM: Long-term 3D human motion generation from multiple sentences. In: CVPRW (2024)
22. Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., Pons-Moll, G.: UniMotion: Unifying 3D human motion synthesis and understanding. In: 3DV (2025)
23. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: SimMotionEdit: Text-based human motion editing with motion similarity prediction. In: CVPR (2025)

24. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: InterGen: Diffusion-based multi-human motion generation under complex interactions. IJCV (2024)
25. 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)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
27. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: PoseGPT: Quantization-based 3D human motion generation and forecasting. In: ECCV (2022)
28. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: ICCV (2019)
29. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: CVPR (2025)
30. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2019)
31. Peng, K., Fu, J., Yang, K., Wen, D., Chen, Y., Liu, R., Zheng, J., Zhang, J., Sarfraz, M.S., Stiefelhagen, R., Roitberg, A.: Referring atomic video action recognition. In: ECCV (2024)
32. Peng, K., Huang, J., Huang, X., Wen, D., Zheng, J., Chen, Y., Yang, K., Wu, J., Hao, C., Stiefelhagen, R.: HopaDIFF: Holistic-partial aware fourier conditioned diffusion for referring human action segmentation in multi-person scenarios. In: NeurIPS (2025)
33. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: ICCV (2021)
34. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: ECCV (2022)
35. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: ICCV (2023)
36. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big Data (2016)
37. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: CVPR (2021)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
39. Ren, Z., Pan, Z., Zhou, X., Kang, L.: Diffusion Motion: Generate text-guided 3D human motion by diffusion model. In: ICASSP (2023)
40. Ruiz-Ponce, P., Barquero, G., Palmero, C., Escalera, S., García-Rodríguez, J.: in2IN: Leveraging individual information to generate human interactions. In: CVPRW (2024)
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
42. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
43. Uchida, K., Shibuya, T., Takida, Y., Murata, N., Tanke, J., Takahashi, S., Mitsufuji, Y.: MoLA: Motion generation and editing with latent diffusion enhanced by adversarial training. In: CVPRW (2025)
44. Wang, L., Fan, H., Chen, H., Huang, Z., Sheng, L.: InterMoE: Individual-specific 3D human interaction generation via dynamic temporal-selective MoE. In: AAAI (2026)
45. Wang, R., He, Y., Sun, T., Li, X., Shi, T.: UniTMGE: Uniform text-motion generation and editing model via diffusion. In: WACV (2025)

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: CVPR (2025)
47. Wang, Y., Li, M., Liu, J., Leng, Z., Li, F.W., Zhang, Z., Liang, X.: Fg-T2M++: LLMs-augmented fine-grained text driven human motion generation. IJCV (2025)
48. Wang, Z., Wang, J., Li, Y., Lin, D., Dai, B.: InterControl: Zero-shot human interaction generation by controlling every joint. In: NeurIPS (2024)
49. Xu, L., Lv, X., Yan, Y., Jin, X., Wu, S., Xu, C., Liu, Y., Zhou, Y., Rao, F., Sheng, X., Liu, Y., Zeng, W., Yang, X.: Inter-X: Towards versatile human-human interaction analysis. In: CVPR (2024)
50. Xu, L., Song, Z., Wang, D., Su, J., Fang, Z., Ding, C., Gan, W., Yan, Y., Jin, X., Yang, X., Zeng, W., Wu, W.: ActFormer: A GAN-based transformer towards general action-conditioned 3D human motion generation. In: ICCV (2023)
51. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: MT-VAE: Learning motion transformations to generate multimodal human dynamics. In: ECCV (2018)
52. Yan, Y., Xu, J., Ni, B., Zhang, W., Yang, X.: Skeleton-aided articulated motion generation. In: ACMMM (2017)
53. Zhang, J., Fan, H., Yang, Y.: EnergyMoGen: Compositional human motion generation with energy-based diffusion model in latent space. In: CVPR (2025)
54. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-driven human motion generation with diffusion model. TPAMI (2024)
55. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: ReMoDiffuse: Retrieval-augmented motion diffusion model. In: ICCV (2023)
56. Zhang, P., Liu, P., Garrido, P., Kim, H., Chaudhuri, B.: KinMo: Kinematic-aware human motion understanding and generation. In: ICCV (2025)
57. Zheng, C., Liu, X., Peng, Q., Wu, T., Wang, P., Chen, C.: DiffMesh: A motion-aware diffusion framework for human mesh recovery from videos. In: WACV (2025)
58. Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., Wang, Y.: Human Motion Generation: A survey. TPAMI (2024)
59. Zhuo, W., Ma, F., Fan, H.: InfiniDreamer: Arbitrarily long human motion generation via segment score distillation. In: ICCV (2025)