



Fine-Grained Conditioning and Interaction Representation Analysis for Text-Driven Human–Human Interaction Motion Generation

Yonghyun Gwon
Kyung Hee University

Abstract

Text-driven human–human interaction (HHI) motion generation must align detailed language with the motions of two people while preserving contact, distance, timing, and role structure. We study two levels of fine-grained conditioning information for HHI generation. First, we add Word–Motion Cross Attention (WMCA) to an InterMask-based pipeline so that motion tokens can attend directly to word tokens. On InterHuman, WMCA improves R-Precision and MM-Dist relative to InterMask, but its FID is worse. The result supports improved text–motion alignment, not an overall gain in generation quality. Second, we define sequence-level interaction attributes, time-indexed relation features, and event/state constraints, then test whether they provide useful conditioning or candidate-selection signals. Some attributes are predictable from text and help select candidates when combined with a text-quality guard. By themselves, however, sequence-level attributes do not determine motion quality. Time-indexed and event/state representations retain more interaction structure, yet explain little of the denoising residual missed by the evaluated generator. Interaction conditions should explicitly separate temporal intervals, body parts, roles, and event order. The present constraints do not yet provide that control.

1. Introduction

Text-driven motion generation maps a natural-language description to a sequence of three-dimensional human poses. Because motion capture and manual animation are costly, text offers a direct way to specify motion. Human–human interaction (HHI) generation adds a second person and relational constraints. Each motion must remain plausible, while the pair’s distance, contact, timing, and roles must agree. For example, “two people shake hands” describes more than two hands moving closer. The people must face one another, approach, make contact at an appropriate time, and maintain a compatible posture after contact. A sequence may look plausible for each person and still look wrong as a pair.

Recent HHI models and the InterHuman benchmark have advanced paired-motion generation [3, 5, 11, 14, 16]. Even so, the text condition does not always separate cues for a person’s detailed motion from cues for the interaction itself. The first problem is word-level text–motion alignment. Many models rely heavily on one sentence embedding, which does not specify where instructions such as “right hand,” “left foot,” “bend down,” or “step backward” should affect the motion. A model can capture the sentence-level meaning while missing a body part, direction, or event in the generated sequence.

The second problem is interaction-level structure. An action label does not always specify how close the participants should be, when contact should begin, whether one person leads and the other responds, or whether their movements should be synchronized. These factors affect whether an HHI sequence looks coherent, but they can become weak or entangled when compressed into one language condition. A useful condition may therefore need to represent contact, distance change, displacement, synchrony, roles, and event states separately.

We study both problems as forms of fine-grained conditioning, while evaluating them separately because they use different backbones and experimental protocols. For word-level alignment, we insert Word–Motion Cross Attention (WMCA) into an InterMask-based pipeline. Flattened motion tokens query CLIP word tokens, allowing each token to draw on the words most relevant to it. For interaction structure, we define attributes and structured constraints at three resolutions and test their text predictability, conditioning behavior, candidate-selection utility, redundancy, quality relevance, and ability to explain denoising residuals. On InterHuman, WMCA improves R-Precision and MM-Dist relative to InterMask, although FID worsens. The representation analysis finds that some attributes are predictable and useful for selection, but the current representations explain little of the residual structure missed by the generator.

Contributions. The contributions are as follows:

- We evaluate WMCA as a direct word-to-motion conditioning path for HHI generation. It improves the reported alignment metrics, while the worse FID rules out a claim of overall generation-quality improvement.
- We formulate sequence-level attributes, time-indexed relations, and event/state constraints for contact, distance, motion, synchrony, roles, temporal order, and interaction phase.
- We evaluate these representations through predictability, selection utility, redundancy, quality relevance, and residual actionability. The results identify useful but incomplete signals and show that timing, body-part contact, role transition, and event order remain insufficiently separated.

2. Related Work

2.1. Text-to-Motion Generation

Text-to-motion generation produces a three-dimensional human motion sequence from a natural-language description. Early methods learned a shared latent space for language and motion or generated continuous motion directly with a conditional model. More recent approaches often encode motion into a latent representation or a sequence of discrete tokens before learning the generative model.

Vector-quantized variational autoencoders (VQ-VAEs) map continuous latent vectors to discrete entries in a codebook [9]. The resulting motion codes can be handled as tokens, which makes them well suited to Transformer architectures. Transformer self-attention computes dependencies within a sequence in parallel [10]; for motion, these dependencies extend across both time and joints.

T2M-GPT uses a two-stage architecture in which a GPT-style model predicts motion tokens obtained from a VQ-VAE [17]. MoMask instead reconstructs masked motion tokens from bidirectional context and uses residual token quantization to retain finer motion detail [2]. These methods produced strong results for single-person text-to-motion generation. Their use of discrete tokens, masked modeling, and attention-based conditioning also informed later HHI generators.

2.2. Human–Human Interaction Motion Generation

HHI generation imposes more constraints than single-person motion generation. Each person’s pose must remain plausible, but the pair must also agree in distance, contact, relative timing, and response. InterGen introduced the InterHuman dataset for this setting and modeled relations between the two people within a diffusion process [5]. With paired three-dimensional motions and natural-language annotations, InterHuman is a standard benchmark for text-

driven HHI generation.

InterMask converts an HHI motion into a two-dimensional map of discrete motion tokens and reconstructs masked tokens for both people jointly [3]. The two-dimensional map retains temporal and spatial organization more directly than a one-dimensional token sequence. Joint prediction also allows the tokens of one person to influence those of the other, which supports the interdependence required by an interaction.

TIMotion arranges the two motion streams as causal interaction sequences and uses Role-Evolving Scanning to account for changes in active and passive roles over time [14]. Its temporal and role-aware modules operate within the generator architecture.

InterMoE introduces a dynamic temporal-selective mixture of experts into a diffusion-based HHI generator [11]. It selects different experts over time using both text semantics and motion context, with the aim of modeling person-specific motion and inter-person interaction together. Our attribute-conditioning experiments use a public generator, checkpoint, and evaluation pipeline from this model family. The representation framework itself, however, is not defined as an InterMoE-specific component.

2.3. Fine-Grained Conditioning and Interaction Structure

Fine-grained conditioning concerns not only the meaning of a sentence as a whole, but also how individual words are grounded in the motion. AttT2M combines sentence-level and word-level text–motion relations to produce more detailed motions [18]. Fg-T2M and Fg-T2M++ use sentence structure, body-part information, and fine-grained semantic relations to improve agreement with the description [12, 13]. These methods divide the text condition into smaller semantic units instead of relying only on one global sentence representation.

Text2Interact’s InterActor lets motion tokens attend to CLIP word tokens and applies an adaptive interaction loss that weights inter-person joint pairs by proximity [16]. These components target fine-grained language alignment and spatial coupling within the generator.

HHI generation adds a second requirement: the model must represent the relation between the two people as well as their individual motions. in2IN uses both an overall interaction description and separate descriptions for each person to increase person-specific motion diversity [7]. InterControl addresses control over joint-pair contact and specified inter-person distances [15]. FineDual uses distance changes and hierarchical inter-person relations to generate dual-human motion [4].

Our experiments examine fine-grained conditioning at two levels: how word-level cues affect individual motion tokens, and whether explicit interaction representations pro-

vide information beyond the generator’s existing text and motion context. We evaluate WMCA on InterMask’s two-dimensional discrete token map for the first question. For the second, we test contact, distance, synchrony, role, and event-state signals in an InterMoE-family generator with a public checkpoint and evaluation environment.

3. Method and Representation Framework

3.1. Overview

We study conditioning at two levels. The first concerns word-level alignment: which words should affect which motion tokens. WMCA addresses this question by adding token-level cross-attention to an InterMask-based generator. The second concerns interaction attributes and constraints: whether contact, distance change, roles, and event progression can be represented as conditions or candidate-selection signals.

WMCA is evaluated on the InterMask pipeline, whereas the interaction representations form a diagnostic framework. Because the two axes use different backbones and answer different questions, they are not direct ablations and their results are interpreted separately. Together, they examine fine-grained text alignment and the distinct problem of representing interaction structure.

3.2. Word–Motion Cross Attention

InterMask encodes input motion as a two-dimensional discrete token map with a VQ-VAE and jointly reconstructs masked tokens for both people. Its text condition is injected mainly through a sentence embedding, which represents the overall meaning but does not directly compute which word is relevant to each motion token.

We retain CLIP word-token embeddings [6] and let flattened motion tokens attend to them. Let $M \in \mathbb{R}^{N_m \times d}$ denote the motion tokens and $C_w \in \mathbb{R}^{N_w \times d}$ the word-token embeddings, where N_m and N_w are the respective token counts and d is the embedding dimension. WMCA computes

$$A = \text{softmax}\left(\frac{(MW_Q)(C_w W_K)^\top}{\sqrt{d_a}}\right)(C_w W_V), \quad (1)$$

followed by the residual update

$$M' = M + AW_O. \quad (2)$$

Here, W_Q , W_K , W_V , and W_O are the query, key, value, and output projections, and d_a is the attention dimension. Each row of the attention map gives a distribution over words for one motion token. The residual form adds word-level conditioning without replacing InterMask’s token reconstruction path. Unlike a single sentence vector, this path lets different motion tokens attend to different lexical cues. It is intended to preserve instructions tied to a particular body part, direction, or event, such as “right leg” or “handshake.”

3.3. Interaction Attributes and Structured Constraints

3.3.1. Sequence-level interaction attributes

Let the paired motion be $X = \{X^{(1)}, X^{(2)}\}$, with $x_{t,j}^{(a)} \in \mathbb{R}^3$ denoting joint j of person $a \in \{1, 2\}$ at frame t . The initial representation computes frame-wise observations and reduces each sequence to a scalar vector. The minimum inter-person joint distance is

$$d_{\min}(t) = \min_{j,k} \|x_{t,j}^{(1)} - x_{t,k}^{(2)}\|_2. \quad (3)$$

The contact ratio is the fraction of frames whose minimum distance is below threshold ϵ_c :

$$p_{\text{contact}}(X) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}[d_{\min}(t) < \epsilon_c]. \quad (4)$$

To characterize the distance during close interaction intervals, we use a low quantile of the distance series:

$$p_{\text{dist}}(X) = \text{Quantile}_q(\{d_{\min}(t)\}_{t=1}^T), \quad q = 0.1. \quad (5)$$

This quantity measures a representative close-range distance, whereas Eq. (4) measures how often the pair is close.

Let $r_t^{(a)}$ be the root or pelvis position of person a . Displacement is the larger mean frame-to-frame root motion of the two people:

$$p_{\text{disp}}(X) = \max_{a \in \{1,2\}} \frac{1}{T-1} \sum_{t=1}^{T-1} \|r_{t+1}^{(a)} - r_t^{(a)}\|_2. \quad (6)$$

If $v^{(a)}(t)$ is the frame-wise magnitude of the full-joint velocity for person a , synchrony is the Pearson correlation between the two velocity series:

$$p_{\text{sync}}(X) = \text{corr}\left(\{v^{(1)}(t)\}_{t=1}^{T-1}, \{v^{(2)}(t)\}_{t=1}^{T-1}\right). \quad (7)$$

Motion similarity averages the absolute cosine similarity between their pose-change vectors:

$$p_{\text{sim}}(X) = \frac{1}{T-1} \sum_{t=1}^{T-1} \left| \cos(\Delta x_t^{(1)}, \Delta x_t^{(2)}) \right|. \quad (8)$$

The resulting sequence-level representation is

$$p_{\text{seq}}(X) = [p_{\text{contact}}, p_{\text{dist}}, p_{\text{disp}}, p_{\text{sync}}, p_{\text{sim}}]. \quad (9)$$

These attributes are simple to compute and useful for measuring selected interaction properties. Their reduction over the whole sequence discards the onset of contact, the body parts involved, and the approach or separation pattern before and after contact.

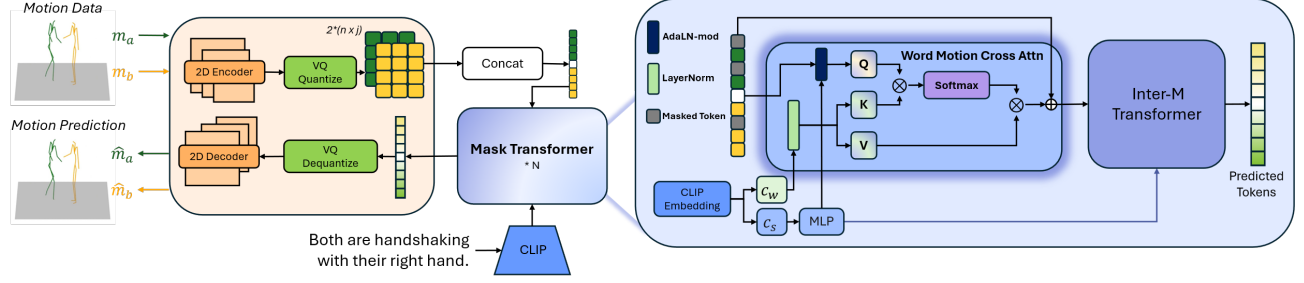


Figure 1. Word-Motion Cross Attention inserted into the InterMask-based pipeline. Motion tokens query CLIP word tokens before masked token prediction.

3.3.2. Time-indexed interaction attributes

We next retain a feature vector at each frame or latent-time bin:

$$P_{\text{time}}(X) = [p_1(X), p_2(X), \dots, p_{T'}(X)], \quad p_\tau(X) \in \mathbb{R}^P. \quad (10)$$

The length T' may equal the frame count or a reduced temporal resolution aligned with the generator. Unlike a sequence average, this representation records where contact strengthens or distance changes. It is an intermediate representation between sequence-level summaries and structured constraints.

For body-part contact, let $\mathcal{B}$ contain meaningful part pairs, such as hand-hand, hand-torso, and foot-lower-body. If $d_b(t)$ is the minimum distance for pair b , a soft contact score is

$$c_b(t) = \sigma\left(\frac{\epsilon_b - d_b(t)}{\tau_b}\right), \quad (11)$$

where ϵ_b is a part-specific contact distance and τ_b controls the transition around it. Scores for different part pairs can be combined as

$$c_{\text{body}}(t) = \sum_{b \in \mathcal{B}} \rho_b c_b(t), \quad \sum_{b \in \mathcal{B}} \rho_b = 1. \quad (12)$$

Uniform weights use $\rho_b = 1/|\mathcal{B}|$; larger weights can emphasize a part pair such as hand contact. This is a general time-indexed construction, not a parameter required by the sequence-level contact ratio.

3.3.3. Structured constraint representation

Time-indexed attributes retain temporal position but remain low-level scalar measurements. Rather than prescribing an exact value at every frame, we group them into states and event relations that language may specify:

$$C^*(X) = \{C_{\text{trend}}, C_{\text{contact}}, C_{\text{role}}, C_{\text{order}}, C_{\text{phase}}\}. \quad (13)$$

Let I_t be an interval over which an interaction state is maintained, and assign a state token z_t to it. Intervals are formed from temporally varying observations such as root

distance, soft contact, and motion energy. Their number is not a fixed parameter chosen to tune performance: each interval covers a period during which the inferred interaction state is maintained. A short local window is used so that the state does not depend on a single frame. The spatial trend of interval I_t is

$$\beta_{I_t} = \text{slope}\left(\left\{\left\|r_t^{(1)} - r_t^{(2)}\right\|_2\right\}_{t \in I_t}\right), \quad (14)$$

$$C_{\text{trend}}(I_t) = \text{sign}(\beta_{I_t}).$$

Body-part contact averages part-specific soft-contact scores within the interval. Let $e^{(1)}(I_t)$ and $e^{(2)}(I_t)$ denote the motion energy of each person. Role or agency is represented by their normalized difference:

$$C_{\text{role}}(I_t) = \frac{|e^{(1)}(I_t) - e^{(2)}(I_t)|}{e^{(1)}(I_t) + e^{(2)}(I_t) + \epsilon}. \quad (15)$$

The identity-free magnitude increases when one person is more active and approaches zero when their activity is similar.

Temporal order compares state start and end times and assigns relations such as before, overlap, adjacent, and after. These edges encode event order that may be expressed in text rather than memorizing frame-level values. One simple event chain is

$$C_{\text{order}} = \{t_{\text{app}} < t_{\text{con}}, t_{\text{con}} < t_{\text{rel}}\}, \quad (16)$$

where approach precedes contact and release follows contact. Event phase combines primitives into states such as approach, contact, hold, and release. This representation contains interval states and order edges, not only a table of scalar values. The intervals provide a middle ground between frame-level measurements and a single sequence summary. We evaluate the representation through text predictability, redundancy, quality relevance, and residual actionability rather than assuming that additional structure makes it a useful generator condition.

3.4. Attribute Prediction and Conditioning

3.4.1. Text-to-attribute prediction

An attribute can serve as a text-derived condition only if it can be predicted from text. For description y , we predict the target attribute from the text encoder output:

$$\hat{p}(y) = f_\phi(E_{\text{text}}(y)). \quad (17)$$

Attributes computed from ground-truth motion provide supervision. Binary targets use binary cross-entropy, while continuous targets use L_1 regression. The two cases can be written in the following general form:

$$\begin{aligned} \mathcal{L}_{\text{attr}} = & \lambda_{\text{bin}} \text{BCE}(\hat{p}_{\text{bin}}(y), p_{\text{bin}}(X)) \\ & + \lambda_{\text{reg}} \|\hat{p}_{\text{reg}}(y) - p_{\text{reg}}(X)\|_1. \end{aligned} \quad (18)$$

Descriptions such as “hug” or “handshake” indicate contact, and “walk away” indicates a distance trend. Attributes whose precise values cannot be inferred from the description are poor candidates for text-derived conditions.

3.4.2. Injection into the motion generator

The predicted attributes can be injected into denoising representation H_t . Additive injection gives

$$H'_t = H_t + \alpha W_p \hat{p}(y). \quad (19)$$

This form is simple, but it merges the text and attribute signals immediately in the same representation. A gate lets the generator regulate the attribute contribution:

$$H'_t = H_t + \sigma(W_g[H_t; \hat{p}(y)]) \odot W_p \hat{p}(y). \quad (20)$$

Alternatively, attribute tokens P remain on a separate condition path and are accessed through cross-attention:

$$\text{Attn}(H_t, P) = \text{softmax}\left(\frac{(H_t W_Q)(P W_K)^\top}{\sqrt{d}}\right) (P W_V). \quad (21)$$

The separate path makes it easier to examine which attributes affect each temporal part of the denoising representation.

3.4.3. Fixed-generator candidate selection

Attributes can also rank samples from a fixed generator. Given text y , generate K candidates:

$$X_1, \dots, X_K \sim G_\theta(y). \quad (22)$$

Because attributes have different units, each attribute j is normalized with reference-set statistics. Without normalization, variables with larger numerical ranges, such as distance or displacement, can dominate the selection score:

$$\tilde{p}_j(X) = \frac{p_j(X) - \mu_j}{\sigma_j + \epsilon}. \quad (23)$$

The attribute score is the negative distance to a normalized target $\tilde{p}^*(y)$, predicted from text or computed from the ground-truth motion:

$$s_{\text{attr}}(X_i, y) = -\|\tilde{p}(X_i) - \tilde{p}^*(y)\|_1. \quad (24)$$

Let $q(X_i, y)$ be a text-quality score in the evaluation text-motion embedding space. The combined score is

$$s_{\text{sel}}(X_i, y) = s_{\text{attr}}(X_i, y) + \lambda q(X_i, y). \quad (25)$$

The text term guards against candidates that match the target attributes but depart from the sentence meaning. This test measures attribute utility without retraining the motion generator.

4. Experiments and Analysis

4.1. Experimental Setup and Evaluation Protocol

Dataset and implementation setting. The experiments use the InterHuman benchmark, which contains paired three-dimensional interaction motions, natural-language annotations, and train, validation, and test splits. The WMCA experiment follows the main InterMask evaluation setting and compares WMCA with InterMask and prior HHI models on the InterHuman test split.

The attribute conditioning and candidate-selection experiments require generated motions from which interaction attributes can be recomputed. A public InterMoE-family generator and checkpoint provide the experimental environment for the interaction-representation analysis.

Metrics and comparisons. R-Precision measures text-motion retrieval alignment and is better when higher. MM-Dist is the distance between text and motion embeddings and is better when lower. FID measures the distance between generated and real motion distributions and is better when lower. Diversity and MultiModality measure variation in generated samples.

The attribute experiments distinguish four comparisons. The released checkpoint provides a reference point for the warm-started runs in the current evaluation setup. Matched controls compare an added attribute path with no sample-specific attribute against the same path supplied with sample-specific attributes. Candidate-selection experiments compare rules over one fixed candidate bank. Representation validation asks whether a representation can be predicted from text, contains information beyond simpler attributes, relates to evaluation quality, and explains residual structure missed by the generator.

4.2. WMCA Evaluation

Table 1 compares WMCA with existing HHI methods, including ComMDM [8] and MoMat-MoGen [1], and with

InterMask on InterHuman. Relative to InterMask, WMCA raises R@1 from 0.449 to 0.472, R@2 from 0.599 to 0.633, and R@3 from 0.683 to 0.715. MM-Dist decreases from 3.790 to 3.786. These changes support improved semantic alignment between text and motion.

FID moves in the opposite direction, increasing from 5.154 to 7.018. The experiment therefore does not establish an overall improvement in generation quality: retrieval-based alignment improves, while distribution quality regresses.

The qualitative examples in Fig. 2 show the same alignment tendency. For a description containing approach, handshake, and bending down, InterMask weakens or blurs some events, whereas WMCA shows them more clearly. WMCA also better preserves the specified right leg.

The qualitative examples assess instruction fidelity; FID still worsens.

4.3. Attribute Conditioning and Candidate Selection

4.3.1. Attribute-conditioned generation

Table 2 evaluates the prediction and injection paths from Sec. 3.4. The released InterMoE checkpoint provides a reference. Within each added structure, $\emptyset$ is a matched control that does not supply a different attribute vector for each sample. SSA denotes sample-specific attributes: the condition path receives the sequence-level contact, displacement, synchrony, motion-similarity, and related values computed for each sample instead of a shared zero or null attribute.

SSA tests whether the condition path can use inter-sample interaction differences. It is not a text-only inference result because the reported sample-specific values are computed from motion. The comparison asks how the additional signal changes alignment and distribution metrics.

For the gated adapter, SSA weakly improves attribute adherence but is worse than its control in R-Precision and FID. For attribute cross-attention, SSA improves R@1, R@3, and MM-Dist relative to its matched control, while FID worsens. Thus, the attribute paths affect the results: cross-attention changes retrieval alignment, and the gated adapter changes adherence. Neither path, however, improves text alignment and distribution quality together. We therefore next test a more limited use of attributes: selecting among samples from an unchanged generator.

4.3.2. Fixed-generator candidate selection

The fixed-generator experiment samples $K = 128$ motions for each text with the released InterMoE checkpoint and applies different selection rules to the same candidate bank. Single sample uses one draw from the generator. Random selection picks one of the 128 candidates. Property-balanced selection favors candidates whose recomputed attributes approach the targets. Text-quality guarded attribute

selection removes candidates with weak text–motion alignment before minimizing attribute error. Oracle selection uses attributes computed from the ground-truth motion as the target.

Property-balanced selection reduces attribute error from 1.648131 for the three-seed random mean to 1.121577, without a comparable gain in R-Precision or MM-Dist. Oracle attribute selection obtains the lowest attribute error but is not the best rule on text–motion alignment. Text-quality guarded selection improves MM-Dist, R@1, R@3, and attribute error relative to the three-seed random mean. Its FID is slightly higher, so the result does not show improved distribution quality.

These comparisons separate two objectives. Matching a sequence-level attribute target is not equivalent to selecting a good HHI motion. In this candidate bank, checking text–motion alignment first yields better alignment metrics than the attribute-only and oracle rules while still reducing attribute error relative to random selection.

4.4. Interaction Representation Validity Analysis

4.4.1. Validation framework

Let E denote one candidate representation from Sec. 3.3. Let B collect information already available to the generator or evaluator, including the text embedding, diffusion timestep, motion length, and latent-time index. We test four questions, denoted Q1–Q4.

Q1 is text predictability. It asks whether representation E can be inferred from description y :

$$\hat{E} = f_{\psi}(E_{\text{text}}(y)). \quad (26)$$

Binary states are evaluated with F1, and continuous values with an error measure or R^2 . Q2 tests non-triviality and redundancy. A phase label adds little new structure if contact and distance already predict it almost completely.

Q3 tests quality relevance. Let m denote a target derived from the official text–motion embedding or from a candidate-quality measure. The incremental explanatory power of E is

$$\Delta R_{Q3}^2 = R^2(m; B, E) - R^2(m; B). \quad (27)$$

A positive value means that E explains part of the target not explained by B . It does not show that the motion generator can use the representation.

Q4 tests residual actionability. If the baseline diffusion generator predicts clean motion $\hat{x}_0^{\text{base}}$, define the missed denoising component as

$$r(x, t) = x_0 - \hat{x}_0^{\text{base}}(x_t, t, y). \quad (28)$$

The extra residual variance explained by E is

$$\Delta R_{Q4}^2 = R^2(r; B, E) - R^2(r; B). \quad (29)$$

Table 1. Quantitative WMCA evaluation on InterHuman. Arrows show the preferred direction; the horizontal arrow marks a diversity metric without a single monotonic optimum.

Method	R@1 $\uparrow$	R@2 $\uparrow$	R@3 $\uparrow$	FID $\downarrow$	MM-Dist $\downarrow$	Div. $\rightarrow$	MMod. $\uparrow$
ComMDM	0.223 $\pm$.009	0.334 $\pm$.008	0.466 $\pm$.010	7.069 $\pm$.054	6.212 $\pm$.021	7.244 $\pm$.038	1.822 $\pm$.052
InterGen	0.371 $\pm$.010	0.515 $\pm$.012	0.624 $\pm$.010	5.918 $\pm$.079	5.108 $\pm$.014	7.387 $\pm$.029	2.141$\pm$.063
MoMat-MoGen	0.449 $\pm$.004	0.591 $\pm$.003	0.666 $\pm$.004	5.674 $\pm$.085	3.790 $\pm$.001	8.021 $\pm$.035	1.295 $\pm$.023
in2IN	0.425 $\pm$.008	0.576 $\pm$.008	0.662 $\pm$.009	5.535 $\pm$.120	3.803 $\pm$.002	7.953 $\pm$.047	1.215 $\pm$.023
InterMask	0.449 $\pm$.004	0.599 $\pm$.005	0.683 $\pm$.004	5.154$\pm$.061	3.790 $\pm$.002	7.944 $\pm$.033	1.737 $\pm$.020
WMCA (Ours)	0.472$\pm$.004	0.633$\pm$.004	0.715$\pm$.004	7.018 $\pm$.101	3.786$\pm$.001	8.229 $\pm$.032	1.986 $\pm$.077

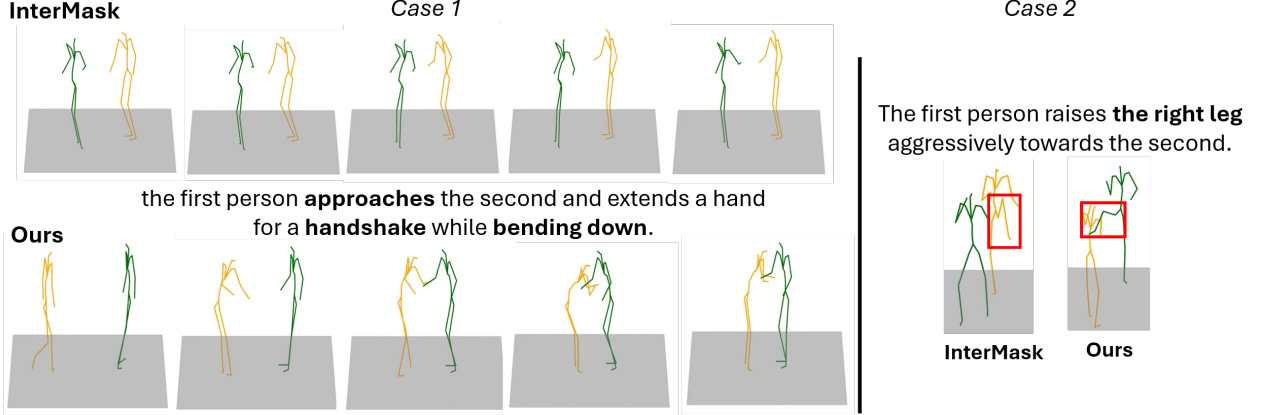


Figure 2. Qualitative WMCA comparisons. The left example contains approach, handshake, and bending-down events; the right example specifies the right leg.

Q4 is the most direct test for generator conditioning. A representation may be predictable from text and related to quality yet still add little if it does not explain what the generator misses.

4.4.2. Sequence-level attribute validation

Before the analysis, we verified the correspondence between InterHuman motion files and extracted rows, checked for missing rows, and confirmed agreement between recomputed formulas and stored values. The verified sequence attributes were then used in the subsequent analysis.

Several attributes are predictable from text: contact reaches F1 70.1%, displacement 63.6%, synchrony 66.0%, and motion similarity 63.5%. The descriptions therefore contain some interaction-attribute information. For example, handshake and hug descriptions imply contact, while dancing together can imply synchrony or similar movement.

Predictability does not make an attribute sufficient for generation quality. In Tab. 3, oracle attribute selection minimizes attribute error but does not give the best R-Precision or MM-Dist. The text-quality guarded rule has slightly higher attribute error and better alignment. The conditioning results show the same mismatch: standard metrics and attribute adherence do not move consistently together. Sequence-level attributes summarize whether contact, dis-

placement, or synchrony is high, but omit contact timing, body part, role transition, and event order.

4.4.3. Time-indexed relation/property validation

The time-indexed analysis aligns relation and property values with the generator’s compressed temporal positions instead of averaging over the sequence. Each sequence is aligned to at most 75 latent-time positions. These positions do not result from truncating the raw motion to 75 frames; they are the compressed temporal locations used inside the generator. The representation is coarser than frame-level coordinates but preserves more local structure than a single sequence average. The basic representation records 13 relations or properties, $R_{TP13} \in \mathbb{R}^{N \times 75 \times 13}$, where N is the number of sequences. An expanded representation records the raw value, activation, strength, temporal change, neighborhood mean or persistence, and confidence for each item: $R_{TPF13 \times 6} \in \mathbb{R}^{N \times 75 \times 13 \times 6}$.

An initial validation flattened each sequence into one summary before explaining residual $r_{i,\tau}$ at latent position τ . Because $r_{i,\tau}$ is local to one compressed time position, a sequence-wide summary can mix the relation at that position with unrelated observations elsewhere in the motion. The resulting ΔR_{Q4}^2 was -0.000311 for R_{TP13} and -0.000277 for $R_{TPF13 \times 6}$, close to the shuffled control at

Table 2. Attribute-conditioned generation. The central comparison is between $\emptyset$ and SSA within the same added structure.

Setting	Condition	MM-Dist ↓	R@1 ↑	R@3 ↑	FID ↓	Interpretation
Released InterMoE checkpoint	$\emptyset$	3.7630±.0012	0.5115±.0047	0.7480±.0044	4.6966±.0932	baseline
Gated attribute adapter	$\emptyset$	3.7620	0.5208	0.7535	4.6030	matched control
Gated attribute adapter	SSA	3.7631	0.5019	0.7481	4.7433	adherence tradeoff
Attribute cross-attention	$\emptyset$	3.7650±.0007	0.5106±.0044	0.7443±.0032	4.6226±.0655	matched control
Attribute cross-attention	SSA	3.7630±.0012	0.5188±.0047	0.7554±.0051	4.8605±.0686	alignment tradeoff

Table 3. Fixed-generator candidate selection from a bank of 128 candidates per text. The final row reports the mean of random selection over three seeds.

Selection rule	MM-Dist ↓	R@1 ↑	R@3 ↑	FID ↓	Attr. Err. ↓
Single sample from generator	3.7657	0.3929	0.6590	4.7083	1.655318
Random selection from 128 candidates	3.7681	0.4048	0.6437	4.5679	1.628226
Property-balanced selection	3.7649	0.4090	0.6624	4.6264	1.121577
Text-quality guarded attribute selection	3.7460	0.4515	0.7117	4.5465	1.133559
Oracle attribute selection	3.7678	0.4056	0.6480	4.7151	1.069742
Random selection mean, 3 seeds	3.7670	0.3901	0.6486	4.5173	1.648131

−0.000248.

The corrected validation pairs $r_{i,\tau}$ directly with relation feature $R_{i,\tau}$ at the same latent-time position. The largest signal occurs for spatial/contact topology. For $R_{\text{TPF13} \times 6}$, test $\Delta R_{Q4}^2 = +0.0002312159$ and $\Delta \text{MSE} = +0.0003594362$. In the same group, R_{TP13} gives $+0.000206$, and adding adjacent temporal neighbors gives $+0.000215$. Temporal/similarity and role/asymmetry remain near zero or negative. Positive values for orientation and event phase are at most $+0.000030$.

The positive values after temporal alignment show that spatial/contact features carry some information about the residual at the same latent-time position. The magnitude is very small, however, and the other relation groups provide no consistent positive signal. Matching temporal positions corrects an overly lossy initial test, but adding a temporal axis alone does not solve the representation problem. Time indexing preserves more structure than a sequence average but does not, in the current formulation, adequately explain the interaction structure missed by the generator. This result motivates the event/state analysis below.

4.4.4. Event/state constraint validation

The structured representation groups observations into spatial trend, body-part contact, role/agency, temporal order, and event phase. Text predictability differs across axes. Contact reaches F1 0.7202, role/agency 0.6391, and spatial trend 0.7822. Temporal order is not sufficiently stable under the text-prediction criterion. Contact and approach/separation trends are often stated directly, whereas action order and role transitions may be omitted or implied.

The redundancy test exposes a limitation of event phase. Spatial/contact information alone explains it up to $R^2 =$

0.9761, suggesting that the current phase definition mostly recombines contact and distance states. In the quality-relevance test, the largest ΔR_{Q3}^2 is $+0.00169$ for the spatial component and $+0.00094$ for the full constraint representation. The relation to candidate quality is measurable but weak.

Residual actionability is weaker still. The largest ΔR_{Q4}^2 values are $+0.000064$ for event phase and $+0.000024$ for spatial/contact. A residual-PCA positive control responds much more strongly. Because that control directly compresses the residual, its response shows that the validation can detect a strong residual signal. The small constraint results are therefore evidence of weak residual explanatory power, not evidence that the validator is insensitive.

The results are mixed across axes. Some axes, such as contact and spatial trend, are predictable from text but only weakly related to the residual. Other axes may contain residual information, but their current definitions remain close to contact and distance. Event phase overlaps with spatial/contact, temporal order is unstable from text, and the residual signal is small. A revised representation must distinguish role transitions, event order, and body-part contact timing from contact/distance-derived values.

4.5. Discussion

Both experimental axes test finer-grained conditioning, at the word–motion and interaction-structure levels, respectively. WMCA directly connects word and motion tokens and improves retrieval-based alignment. The attribute analysis finds predictable signals and some candidate-selection utility, but scalar attributes and attribute error do not reliably identify a good HHI sample. The evidence does not show

that adding attributes automatically improves generation.

The representation validation distinguishes the limitations of the three representation levels. Sequence-level attributes discard contact time, body part, role transitions, and event order. Time-indexed properties retain temporal position but explain little of the matched denoising residual. Event/state constraints better reflect concepts expressible in language, but their present phase and contact dimensions are redundant and their residual signal remains weak.

Recent HHI generators have reported gains from temporal role modeling, word-level conditioning, and adaptive joint-relation supervision [14, 16]. In our experiments, the explicit representations provide little additional information beyond the tested generator. This result is specific to those representations and does not imply that role, contact, or phase information is generally uninformative.

Future interaction conditions should retain word-level alignment while separating temporal interval, body part, role, and event order. The current experiments motivate that redesign; they do not establish that the present structured constraints solve interaction-level control.

5. Conclusion

This work examined two ways to use conditioning information more precisely in text-driven HHI motion generation. The first addresses word-level text–motion alignment through WMCA. The second analyzes whether interaction attributes and structured constraints can serve as generation conditions or candidate-selection signals.

WMCA gives motion tokens direct access to word tokens in an InterMask-based pipeline. On InterHuman, it improves R-Precision and MM-Dist relative to InterMask, and the qualitative examples show clearer responses to named events and body parts. FID worsens, however. The experiment therefore supports improved fine-grained text alignment, not an overall improvement in generation quality.

The interaction analysis defines sequence-level attributes, time-indexed relations, and structured event/state constraints, then examines them through text prediction, generator conditioning, fixed-generator candidate selection, and representation validation. Sequence-level attributes are partly predictable from text and can assist selection when a text-quality criterion is enforced, but they do not determine the quality of an HHI motion by themselves. Time-indexed relations preserve temporal position, yet their explanatory power remains very weak even after they are aligned with residuals at the same latent-time locations; the surviving signal is concentrated in spatial/contact topology. Event/state constraints provide a more suitable structure for representing interaction events, but several current axes overlap with contact and distance and still explain little of the denoising residual.

Sequence-level summaries cannot retain when and

where interaction events occur. Future representations should preserve temporally localized body-part relations, role transitions, event order, and phase transitions. They should be predictable from language, add information beyond contact and distance, and be usable through a condition path separate from the word-level text path. The current experiments motivate, but do not validate, this design.

References

- [1] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, Liang Pan, Xiangyu Fan, Han Du, Peng Gao, Zhitao Yang, Yang Gao, Jiaqi Li, Tianxiang Ren, Yukun Wei, Xiaogang Wang, Chen Change Loy, Lei Yang, and Ziwei Liu. Digital life project: Autonomous 3D characters with social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 582–592, 2024.
- [2] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng. MoMask: Generative masked modeling of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024.
- [3] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3D human interaction generation via collaborative masked modelling. In *International Conference on Learning Representations*, 2025.
- [4] M. Li, Y. Wang, Z. Leng, J. Liu, F. W. B. Li, and X. Liang. Fine-grained text-driven dual-human motion generation via dynamic hierarchical interaction. *arXiv preprint arXiv:2510.08260*, 2025.
- [5] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 2024.
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.
- [7] Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2IN: Leveraging individual information to generate human interactions. In *CVPR Workshop on Human Motion Generation*, 2024.
- [8] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.
- [9] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

- [11] Lipeng Wang, Hongxing Fan, Haohua Chen, Zehuan Huang, and Lu Sheng. InterMoE: Individual-specific 3D human interaction generation via dynamic temporal-selective MoE. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):2137–2145, 2026.
- [12] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-T2M: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [13] Y. Wang, M. Li, J. Liu, Z. Leng, F. W. B. Li, Z. Zhang, and X. Liang. Fg-T2M++: LLMs-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, 133(7):4277–4293, 2025.
- [14] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. TIMotion: Temporal and interactive framework for efficient human–human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7169–7178, 2025.
- [15] Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. InterControl: Zero-shot human interaction generation by controlling every joint. In *Advances in Neural Information Processing Systems*, 2024.
- [16] Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2Interact: High-fidelity and diverse text-to-two-person interaction generation. In *The Fourteenth International Conference on Learning Representations*, 2026.
- [17] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [18] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. AttT2M: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.