SOCIALDIT: SOCIAL-AWARE DIFFUSION TRANSFORMER FOR HUMAN-HUMAN INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two people performing the same action move differently depending on who they are and how well they know each other, yet text-driven dyadic motion generators condition on the action description alone. We study text-driven dyadic interaction generation additionally conditioned on per-actor Big Five personality and pair-level familiarity. We introduce **SocialDiT**, a flow-matching diffusion transformer that feeds the caption through text cross-attention and the social attributes through a Social-Aware Cross-person Attention (SACA) module, which modulates the attention between the two participants. We show that on Inter-X, where every subject pair carries a single fixed label, probes that predict these attributes from motion are confounded by pair identity. We therefore test whether the social input is used and controllable with interventions, a controlled familiarity probe, evaluator-free interaction geometry, and a user study. On a subject-disjoint split, SocialDiT reduces FID from 1.90 to 1.03 relative to the strongest socially conditioned baseline. It responds to interventions on its social input and to controlled changes in familiarity, reproduces interpersonal distances more closely than the baselines, and is preferred in a user study, including on matching the target familiarity. A controlled comparison of injection sites further shows that routing social attributes through cross-person attention best reproduces interpersonal geometry, suggesting that social context is most naturally injected where the two participants interact.

1 INTRODUCTION

Generating two-person interactions from natural-language descriptions has become an active problem in human motion synthesis (Liang et al., 2024; Javed et al., 2025; Ponce et al., 2024; Wang et al., 2025). Yet the same caption can be realized in different ways depending on who the participants are and how they relate: a more extraverted person may greet with broader gestures, and close friends and strangers may keep different distances during the same exchange. The caption specifies *what* the two people do but leaves *how* they do it underspecified (Fig. 1). This "how" is shaped by social attributes such as personality and familiarity, which are central to social signal processing (Vinciarelli et al., 2009; Vinciarelli & Mohammadi, 2014). Controlling it matters wherever a scripted action has to be re-rendered for different characters and relationships, as in virtual agents, animation authoring, and social simulation.

Text-driven dyadic generators (Liang et al., 2024; Javed et al., 2025; Ponce et al., 2024; Wang et al., 2025; Tanke et al., 2025; Wu et al., 2025; Wang et al., 2026a; Ruiz-Ponce et al., 2026) model the "what" axis well but condition only on text. Inter-X (Xu et al., 2024a) annotates per-actor Big Five ratings (McCrae & John, 1992) and pair-level familiarity, and defines a stylized generation task in which action-label-conditioned generators also receive these attributes. For text-driven, symmetric dyadic generation, two questions remain open: where such attributes should enter the model, and how to verify that they are expressed in the generated motion.

We present **SocialDiT**, a flow-matching diffusion transformer conditioned on a caption together with per-actor Big Five personality and pair-level familiarity. The caption enters through standard text cross-attention, whereas the social attributes enter the attention *between* the two participants: the behaviors they shape, such as interpersonal distance and contact, are defined by how the two people

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

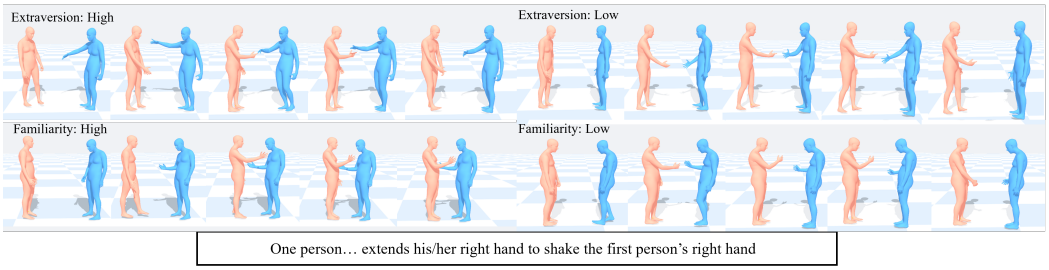


Figure 1: **Same caption, different social input.** Editing a single attribute with SocialDiT: extraversion changes gestural amplitude and posture openness (top); familiarity changes interpersonal distance and contact (bottom).

move relative to each other. Our **Social-Aware Cross-person Attention (SACA)** module modulates this attention with trait-aware FiLM and an input-dependent gate.

Verifying that social attributes are expressed is itself a challenge. The obvious test is a probe that predicts the attributes from generated motion, but on Inter-X each subject pair has one fixed attribute tuple, so a probe trained and tested on the same pairs can succeed by recognizing the pair; on unseen pairs, even real motion drops from $R^2 = 0.544$ to -0.534. We therefore ask whether the social input is *used* (interventions), *controllable* (varying only familiarity), reflected in the *interaction* (3D interpersonal geometry), and *perceived* (a user study). For generated motion, the probe score within the test pairs remains informative, since it collapses when the social input is removed; we report it as condition-consistent separability (Sec. 4).

Contributions.

- We propose SocialDiT, a text-driven dyadic generator conditioned on per-actor Big Five personality and pair-level familiarity, which injects these attributes into cross-person attention through SACA.
- We show that attribute probes on Inter-X are confounded by pair identity, even for real motion, and propose an evaluation protocol based on interventions, controlled generation, evaluator-free interaction geometry, and a user study.
- On a subject-disjoint split, SocialDiT reduces FID from 1.90 to 1.03 over the best socially conditioned baseline, passes all four tests, and reproduces interpersonal distances most accurately, even compared with variants that inject the same attributes elsewhere.

2 Related Work

Text-driven motion and interaction generation. Single-person text-to-motion synthesis has progressed through diffusion models that denoise poses or latents (Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023; Zhang et al., 2023b) and token-based transformers (Zhang et al., 2023a; Jiang et al., 2023; Guo et al., 2024), with extensions for editing and spatial control (Uchida et al., 2025; Karunratanakul et al., 2023; Xie et al., 2023). Dyadic generation adds mutual spatial constraints and inter-person coordination. InterGen uses cooperative shared-weight denoisers (Liang et al., 2024); in2IN adds per-individual descriptions (Ponce et al., 2024); InterMask performs collaborative masked modeling (Javed et al., 2025); TIMotion models temporal and interactive structure (Wang et al., 2025); state-space (Tanke et al., 2025; Wu et al., 2025), mixture-of-experts (Wang et al., 2026a), and autoregressive full-body (Ruiz-Ponce et al., 2026) backbones further improve quality, while unified (Fan et al., 2024; Ma et al., 2025) and action–reaction (Xu et al., 2024b; Ghosh et al., 2024; Cen et al., 2025) formulations broaden the setting. Concurrently, Wang et al. (2026b) use a language-model planner to decompose an interaction into phases and actor roles; their "social structure" describes how an action unfolds, whereas we condition on attributes of the actors and of their relationship. These generators condition on text but not on actor- or dyad-level attributes.

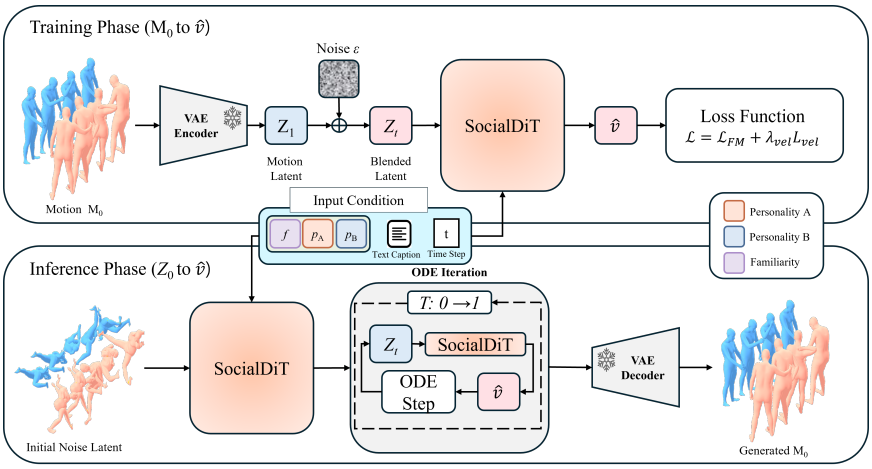


Figure 2: **SocialDiT pipeline.** Training (top): SocialDiT predicts the velocity $\hat{v}$ of the linear flow between noise and VAE latents Z_1, conditioned on text and $s = (f, p_A, p_B)$. Inference (bottom): Euler integration from noise, then VAE decoding.

Conditioning motion on social and affective attributes. For single-person motion, style transfer recomposes an action under a learned style (Aberman et al., 2020; Park et al., 2021; Zhong et al., 2024), and affect-conditioned generators modulate gait, gesture, or full-body motion by emotion (Bhattacharya et al., 2020; Yin et al., 2023; Qi et al., 2024; Ghorbani et al., 2023). Early personality-driven animation modulated gaze or agent behavior with Big Five vectors (Durupinar, 2020; Egges et al., 2004), and perception studies indicate that human raters attribute Big Five personality from motion cues (Liu et al., 2015). Dyadic work has mostly addressed *asymmetric* settings, generating a listener's response to a speaker (Ng et al., 2022; Jung & Kim, 2025; Siniukov et al., 2025) or a reactor's response to a fixed actor (Xu et al., 2024b; Ghosh et al., 2024; Cen et al., 2025; Zhu et al., 2025). Inter-X (Xu et al., 2024a) defines a stylized interaction generation task conditioned on an action label together with relationship and personality, benchmarked by supplying these labels as additional inputs to action-conditioned generators, as well as a personality assessment task. Our setting differs in three respects: the action is given as a free-form caption rather than a label, the social attributes enter through a dedicated cross-person module rather than being appended to an existing condition, and we evaluate whether the specified attributes are expressed rather than generation quality alone. The closest model, DyaDiT (Peng et al., 2026), conditions dyadic gesture generation on social context, including a Big Five vector, through FiLM modulation. SocialDiT differs in being text-driven rather than audio-driven, in generating full-body interaction rather than upper-body gesture, and in generating both participants jointly rather than a single agent. To our knowledge, SocialDiT is the first text-driven, symmetric dyadic generator with a dedicated module that jointly conditions on per-actor Big Five personality and pair-level familiarity.

3 SocialDiT

Problem setup. Given a text caption c, per-person Big Five traits (McCrae & John, 1992) $p_A, p_B \in [0, 1]^5$, and a pair-level familiarity label $f \in \{0, 1, 2, 3\}$ provided by Inter-X (Xu et al., 2024a), we generate two coordinated motion sequences x^A, x^B that follow the described action and reflect the specified social context. The caption fixes *what* is performed, whereas the social attributes act on *how*: expressiveness, interpersonal distance, contact, and responsiveness, properties commonly studied as nonverbal social signals (Vinciarelli et al., 2009). Several of these properties are *relational*—they are defined by how the two participants move with respect to each other rather than by either participant alone. This motivates our central design choice: the caption takes a standard text path, while the social attributes enter the component that models how the two participants influence each other—their cross-person attention. We test whether this placement matters against global and per-stream alternatives in Sec. 5.3.

3.1 LATENT FLOW MATCHING

Fig. 2 gives an overview of training and inference. We encode each participant's motion with a skeleton-aware spatio-temporal VAE, following the design used for Inter-X in InterMoE (Wang et al., 2026a) and learn SocialDiT as a conditional velocity field over its latent space (Rombach et al., 2022; Chen et al., 2023). The VAE is trained separately on the training split and kept frozen (App. C). Each participant's motion is encoded independently as z^i, $i \in \{A, B\}$, and paired into $Z_1 = (z^A, z^B)$ in the normalized latent space. We train with flow matching (Lipman et al., 2023), using $t = 0$ for noise and $t = 1$ for data. With $Z_0 \sim \mathcal{N}(0, I)$, the linear path is

$$Z_t = (1 - t)\, Z_0 + t\, Z_1, \qquad v^\star = Z_1 - Z_0, \tag{1}$$

and SocialDiT predicts $v_\theta(Z_t, t, c, s)$ with $s = (f, p_A, p_B)$. We add a first-order temporal regularizer that encourages the predicted velocity to track the local temporal evolution of the target:

$$\mathcal{L} = \mathbb{E}_{Z_0, Z_1, t}\left[\|v_\theta(Z_t, t, c, s) - v^\star\|_2^2 + \lambda_{\text{vel}}\, \|\Delta_\tau v_\theta - \Delta_\tau v^\star\|_2^2\right], \tag{2}$$

with Δ_τ defined in App. C.

3.2 CONDITION ENCODING

The caption and the flow time describe what is performed and at which generation stage, and form a global condition. We encode c with a frozen CLIP (Radford et al., 2021), producing token features E_c^{tok} for text cross-attention and a pooled embedding e_c^{pool}. The flow time is embedded as $e_t = \phi(t)$ with a sinusoidal encoding and an MLP, and

$$g = \text{MLP}_g\big([e_c^{\text{pool}}, e_t]\big) \tag{3}$$

drives AdaLN modulation throughout the network and produces the SACA gate. Personality and familiarity, in contrast, describe how the participants behave and relate. They are encoded by separate trainable encoders E_{pers} and E_{fam} (App. D) and fused into a per-participant social embedding

$$s_i = \text{MLP}_s([E_{\text{pers}}(p_i),\, E_{\text{fam}}(f)]), \qquad i \in \{A, B\}, \tag{4}$$

which summarizes the actor-level traits and the dyad-level relationship from i's perspective.

3.3 SOCIAL-AWARE CROSS-PERSON ATTENTION

SocialDiT (Fig. 3) maintains one latent stream per participant, so that per-person motion states are preserved and cross-person interaction is modeled explicitly rather than through a single fused representation. The noisy latents (Z_t^A, Z_t^B) are projected to the hidden dimension, combined with positional and identity embeddings, and processed by N blocks of self-attention, text cross-attention, SACA, and a feed-forward network. Self-attention and the feed-forward network follow the DiT design (Peebles & Xie, 2023) with AdaLN modulation from g, and caption semantics enter through text cross-attention,

$$h_i \leftarrow h_i + \text{CrossAttn}(\text{LN}(h_i), E_c^{\text{tok}}), \quad i \in \{A, B\}. \tag{5}$$

SACA augments cross-person attention with three elements with distinct roles: FiLM modulation determines *what* information passes between the two participants, multi-head attention aggregates it under the current motion state, and a residual gate controls *how strongly* the update enters the text-grounded stream. For the $A \leftarrow B$ direction,

$$\tilde{h}_A^q = \text{FiLM}(\text{LN}_q(h_A);\, s_A), \qquad \tilde{h}_B^{kv} = \text{FiLM}(\text{LN}_{kv}(h_B);\, s_B), \tag{6}$$

where FiLM (Perez et al., 2018), $\text{FiLM}(x; s) = (1 + \gamma(s)) \odot x + \beta(s)$, is shared across the two branches but applied with each participant's own embedding, so that the attributes of both participants shape the exchange. The social update

$$u_{A \leftarrow B} = \text{MHA}(\tilde{h}_A^q,\, \tilde{h}_B^{kv}) \tag{7}$$

is added through an input-dependent gate predicted from the global condition,

$$h_A \leftarrow h_A + \alpha_s(g) \odot u_{A \leftarrow B}, \tag{8}$$

with $\alpha_s(g)$ broadcast over latent tokens; $B \leftarrow A$ is computed symmetrically. A final AdaLN layer and an output projection produce $v_\theta(Z_t, t, c, s)$. The social embeddings enter the network *only* through the FiLM terms of SACA, so removing FiLM severs the social path entirely (Sec. 5.4).

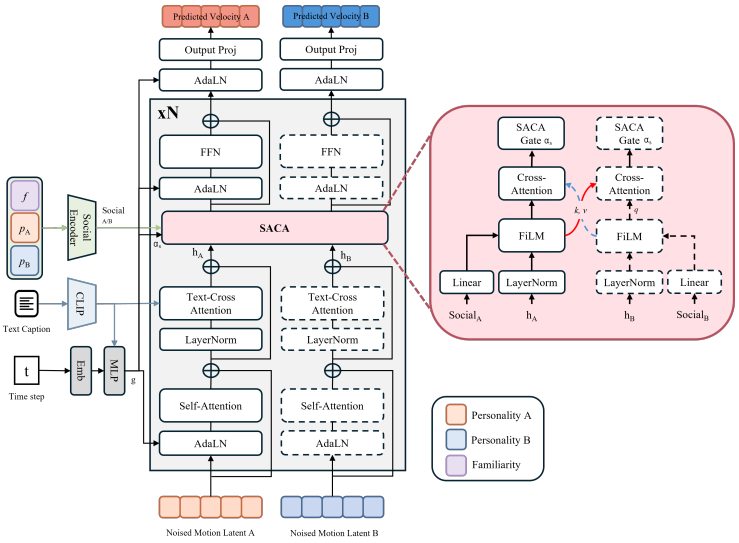


Figure 3: **SocialDiT architecture.** One latent stream per participant. Text cross-attention grounds the action; SACA injects the social embeddings s_A, s_B into cross-person attention through FiLM, and the gate $\alpha_s(g)$ controls the residual contribution of the social update. Dashed modules denote the symmetric branch.

3.4 Training and Inference

Trait-space augmentation. Personality labels in Inter-X are subject-level, so the training split covers only 47 pair-level trait configurations and a sparse region of the five-dimensional trait space. To discourage memorizing these points, we perturb the traits during training with probability p_{aug},

$$\bar{p}_i = \text{clip}(p_i + \epsilon, 0, 1), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{9}$$

and use the specified traits without noise at inference.

Dual-axis classifier-free guidance. We use classifier-free guidance (Ho & Salimans, 2022) in the two-condition form introduced by InstructPix2Pix (Brooks et al., 2023), with the caption as the first condition and the social attributes as the second. We train with three conditioning modes selected at random per sample: an unconditional mode in which text and social inputs are replaced by learned null embeddings, an action-only mode in which only the social input is dropped, and a full mode, with predictions

$$v_{\text{unc}} = v_\theta(Z_t, t, \varnothing, \varnothing), \quad v_{\text{act}} = v_\theta(Z_t, t, c, \varnothing), \quad v_{\text{full}} = v_\theta(Z_t, t, c, s), \tag{10}$$

and combine them at inference as

$$v = v_{\text{unc}} + w_a\big(v_{\text{act}} - v_{\text{unc}}\big) + w_s\big(v_{\text{full}} - v_{\text{act}}\big), \tag{11}$$

where w_a amplifies the action direction and w_s the social residual, so that the strength of social expression can be set independently of caption grounding; the formulation reduces to standard guidance when $w_a = w_s$. We use $(w_a, w_s) = (3.5, 7.0)$ by default; the full guidance sweep is reported in App. H. Sampling integrates $dZ_t/dt = v(Z_t, t, c, s)$ from $t = 0$ to 1 with 50 uniform Euler steps, and the final latent is decoded by the frozen VAE decoder.

4 Evaluating Social Conditioning

Data and split. We use Inter-X (Xu et al., 2024a), whose 59 groups each correspond to a unique subject pair with a fixed $(\text{Big Five}_A, \text{Big Five}_B, f)$ tuple. We partition the groups by subject identity into 47 training, 6 validation, and 6 test groups, so that no test subject—and hence no test trait vector—appears in training; three of the six test groups lie outside the training $\mu \pm \sigma$ band on at least one trait (App. B). This makes social conditioning at test time a generalization problem.

Probe-based evaluation. A natural way to test whether generated motion expresses its conditioning is to train a linear probe (Alain & Bengio, 2016) that predicts the attributes from motion features and to compare its score with that of real motion. When the probe is validated with sample-level random folds, every pair appears in both the training and test folds with the same label, so the probe can succeed by recognizing the pair. Group-wise folds, which hold out entire pairs, drive the personality probe below zero for every source: real motion drops from 0.544 to -0.534, SocialDiT from 0.548 to -0.370, and the baselines to -0.304 (InterGen) and -0.253 (InterMask) (Tab. 7, App. E). With one label per pair, exact trait values of held-out pairs cannot be decoded from this data for any source, including real motion. The sample-level score nevertheless remains informative for generated motion: a probe can separate the test pairs only if the model expresses each specified condition as a consistent, linearly distinguishable style, and the score collapses when the social input is removed while the caption is kept (Tab. 3). We report it as *condition-consistent separability* (CCS), with the real-motion value (0.544) as a reference level.

Evaluation protocol. Our evaluation combines five kinds of evidence, none of which requires decoding labels across pairs. **(E1) Realism and text alignment:** FID, R-precision, MM-Distance, and Diversity, computed with the InterMask evaluator (Javed et al., 2025) following Liang et al. (2024). **(E2) Evaluator-free geometry:** errors of pelvis-to-pelvis and closest-joint interpersonal distances with respect to real motion (MAE and Jensen–Shannon divergence, JSD), and ground-contact errors from the official HUMOS implementation (Tripathi et al., 2024). **(E3) Dependence on the social input:** with captions fixed, we remove the social input, reassign Big Five vectors across pairs, reassign familiarity levels, or both; a model that ignored its social input would be unaffected. **(E4) Controllability:** for every test caption we generate four interactions that differ *only* in the familiarity input (personality fixed at the dataset mean) and train a classifier to predict that input. Every caption appears with all four labels, so pair identity cannot explain success. **(E5) Human judgment:** a pairwise user study on caption alignment, motion quality, and whether the interaction matches the target familiarity level. Protocol details are in App. E, App. F, and App. I.

5 Experiments

Baselines and setup. We compare with InterGen (Liang et al., 2024) and InterMask (Javed et al., 2025), retrained from scratch on our split with their original hyperparameters. Neither includes social conditioning; we extend each by concatenating the social inputs with its text-conditioning input and report both the social-conditioned and the original variants. To isolate where the social attributes should enter, we further compare injection sites within our backbone (Sec. 5.3). Implementation details are in App. A.

5.1 Comparison with Baselines

Realism and text alignment (E1). SocialDiT attains the lowest FID and the highest R-precision at all ranks among generated methods (Tab. 1). Adding social conditioning by concatenation degrades FID for both baselines (InterGen $3.250 \rightarrow 3.570$; InterMask $1.441 \rightarrow 1.902$), whereas SocialDiT with full social conditioning attains a lower FID (1.033) than either baseline without it. As expected, baselines without social input yield negative CCS, whereas every socially conditioned model yields positive CCS.

Table 1: **Comparison on the Inter-X subject-disjoint test split.** "w.o. social": original baselines without social conditioning. CCS: condition-consistent separability (Sec. 4). →: closer to real motion is better. Mean and 95% CI over 20 seeds.

Method	FID ↓	Diversity →	R-1 ↑	R-2 ↑	R-3 ↑	MM Dist. ↓	CCS
Real motion	—	$8.960^{\pm.077}$	$0.287^{\pm.004}$	$0.467^{\pm.004}$	$0.605^{\pm.004}$	$2.967^{\pm.008}$	0.544
InterGen (w. social)	$3.570^{\pm.084}$	$\underline{8.988}^{\pm.071}$	$0.143^{\pm.002}$	$0.262^{\pm.004}$	$0.365^{\pm.005}$	$5.187^{\pm.020}$	$0.367^{\pm.010}$
InterGen (w.o. social)	$3.250^{\pm.082}$	$\mathbf{8.970}^{\pm.071}$	$0.164^{\pm.004}$	$0.290^{\pm.006}$	$0.394^{\pm.005}$	$4.925^{\pm.023}$	$-0.445^{\pm.013}$
InterMask (w. social)	$1.902^{\pm.048}$	$8.561^{\pm.055}$	$0.162^{\pm.003}$	$0.286^{\pm.003}$	$0.384^{\pm.003}$	$4.978^{\pm.021}$	$0.501^{\pm.007}$
InterMask (w.o. social)	$\underline{1.441}^{\pm.036}$	$8.917^{\pm.048}$	$\underline{0.190}^{\pm.005}$	$\underline{0.331}^{\pm.005}$	$\underline{0.444}^{\pm.005}$	$\mathbf{4.320}^{\pm.019}$	$-0.409^{\pm.014}$
Ours	$\mathbf{1.033}^{\pm.021}$	$9.335^{\pm.094}$	$\mathbf{0.191}^{\pm.004}$	$\mathbf{0.338}^{\pm.005}$	$\mathbf{0.458}^{\pm.004}$	$\underline{4.439}^{\pm.019}$	$0.550^{\pm.006}$

Interaction geometry and ground contact (E2). Measured directly from 3D geometry, without a learned evaluator, SocialDiT reproduces interpersonal distance most accurately on all four measures (Tab. 2). The margin is largest on the distributional measures, where InterMask's divergence from real motion is about twice ours (JSD 0.069 vs. 0.035 for pelvis distance and 0.050 vs. 0.029 for closest-joint distance). Ground contact shows trade-offs rather than a single winner, as also observed in HUMOS (Tripathi et al., 2024): InterMask has the lowest penetration and skating but floats the most, InterGen shows the reverse pattern, and SocialDiT ranks second on all three without trading one failure for another. Fig. 4 shows qualitative samples.

Table 2: **Evaluator-free measurements** (test split). Ground contact: HUMOS errors. Interpersonal distance: error with respect to the corresponding real interaction (MAE) and the pooled real distribution (JSD), for pelvis-to-pelvis and closest-joint distances. Baselines are the social-conditioned variants of Tab. 1. Lower is better; mean and 95% CI over 5 seeds.

	Ground contact			Distance MAE		Distance JSD	
Method	Penetrate ↓	Float ↓	Skate ↓	Pelvis ↓	Closest ↓	Pelvis ↓	Closest ↓
Real motion	0.971	1.186	0.080	—	—	—	—
InterGen	$4.089^{\pm.148}$	$\mathbf{2.445}^{\pm.081}$	$0.511^{\pm.007}$	$0.435^{\pm.004}$	$0.326^{\pm.004}$	$0.054^{\pm.002}$	$0.034^{\pm.002}$
InterMask	$\mathbf{2.387}^{\pm.063}$	$6.965^{\pm.226}$	$\mathbf{0.214}^{\pm.006}$	$0.412^{\pm.005}$	$0.328^{\pm.006}$	$0.069^{\pm.002}$	$0.050^{\pm.003}$
Ours	$3.266^{\pm.069}$	$4.157^{\pm.073}$	$0.486^{\pm.005}$	$\mathbf{0.351}^{\pm.005}$	$\mathbf{0.308}^{\pm.005}$	$\mathbf{0.035}^{\pm.002}$	$\mathbf{0.029}^{\pm.002}$

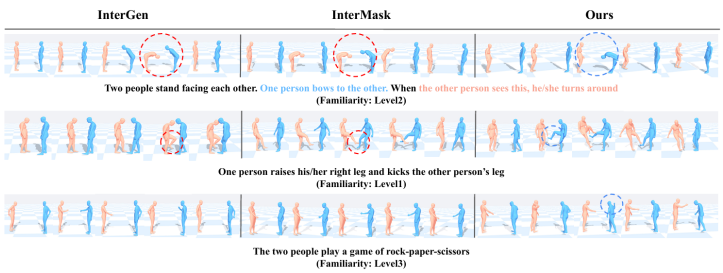


Figure 4: **Qualitative comparison** on test captions. Red circles mark caption–action mismatches in the baselines; blue circles mark the corresponding actions generated by SocialDiT.

5.2 Dependence, Controllability, and Human Judgment

Interventions (E3). With captions fixed, the output depends on the social input (Tab. 3). Removing the social input leaves FID almost unchanged (1.095) while CCS falls below zero, i.e., the model falls back to a caption-driven prior. Reassigning Big Five vectors or familiarity levels across pairs instead raises FID from 1.023 to 2.647 and 3.379, respectively, and CCS collapses to near zero. Fig. 1 shows single-attribute edits with the caption fixed.

Table 3: **Interventions on the social input** at inference, with captions fixed (removed: null embedding; shuffled: reassigned across test pairs). Mean and 95% CI over 5 seeds.

Social input	**FID ↓**	**R-3 ↑**	**MM Dist. ↓**	**CCS**
Unchanged	$1.023^{\pm.037}$	$0.457^{\pm.007}$	$4.426^{\pm.038}$	$0.548^{\pm.003}$
Removed	$1.095^{\pm.030}$	$0.480^{\pm.005}$	$4.085^{\pm.038}$	$-0.342^{\pm.012}$
Shuffled familiarity	$3.379^{\pm.151}$	$0.440^{\pm.006}$	$4.586^{\pm.036}$	$0.057^{\pm.013}$
Shuffled Big Five	$2.647^{\pm.118}$	$0.438^{\pm.008}$	$4.701^{\pm.058}$	$0.040^{\pm.017}$
Shuffled both	$2.657^{\pm.105}$	$0.437^{\pm.011}$	$4.571^{\pm.033}$	$-0.398^{\pm.043}$

Controlled familiarity probe (E4). When only the familiarity input varies, a four-way logistic-regression probe recovers it from SocialDiT's outputs with accuracy 0.834 and macro-F1 0.817,

compared with 0.720/0.649 for InterGen and 0.701/0.648 for InterMask (Fig. 5). SocialDiT reaches a per-level recall of at least 0.76 on every level, whereas InterGen and InterMask drop to 0.51 and 0.54 on their weakest levels. Familiarity is therefore an effective control.

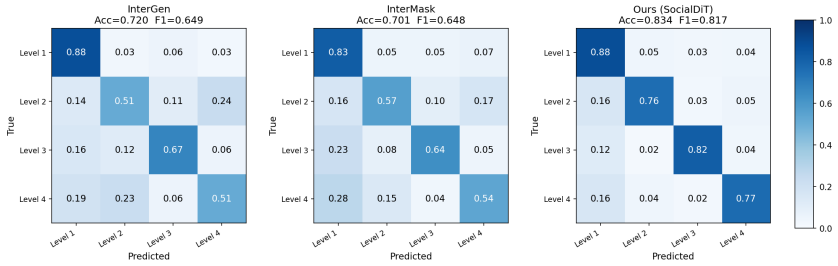


Figure 5: **Familiarity controllability.** Confusion matrices of a four-way probe that predicts the familiarity input from generated motion when only that input varies (levels 1–4 correspond to $f = 0$–3).

Human judgment (E5). Twenty participants compared SocialDiT with each social-conditioned baseline, with the caption and the target familiarity level shown (App. I). SocialDiT was preferred over both baselines on all three axes, including match to the target familiarity (Tab. 4).

Table 4: **User study** (20 participants). Percentage of comparisons in which SocialDiT is preferred (W), tied (T), or the baseline is preferred (L).

Comparison	**Caption** (W/T/L %)	**Quality** (W/T/L %)	**Familiarity** (W/T/L %)
vs. InterGen	**59.2**/6.0/34.8	**57.4**/7.6/35.0	**56.4**/9.8/33.8
vs. InterMask	**56.4**/8.0/35.6	**52.4**/10.6/37.0	**51.4**/16.2/32.4

5.3 Effect of the Social Injection Site

Two variants share everything with SocialDiT except where the social embeddings enter: per-stream injection adds each participant's embedding to its own AdaLN modulation, and shared injection fuses both embeddings into the global condition (App. G). The two variants give better FID and R-precision (Tab. 5), which measure distributional realism and caption alignment but not how the two participants are placed relative to each other. On the interaction itself, cross-person injection is best on all four interpersonal-distance measures, supporting our choice to inject social attributes where the two participants interact.

Table 5: **Effect of the social injection site.** The three variants differ only in where the social embeddings enter (App. G); $(w_a, w_s) = (3.5, 7.0)$. Mean and 95% CI over 5 seeds.

	Realism & alignment		**Distance MAE**		**Distance JSD**	
Injection site	FID ↓	R-3 ↑	Pelvis ↓	Closest ↓	Pelvis ↓	Closest ↓
Per-stream (AdaLN)	**0.852**$^{\pm.048}$	**0.470**$^{\pm.007}$	0.366$^{\pm.006}$	0.328$^{\pm.007}$	0.038$^{\pm.002}$	0.032$^{\pm.001}$
Shared (global)	0.904$^{\pm.040}$	0.469$^{\pm.007}$	0.357$^{\pm.005}$	0.318$^{\pm.005}$	0.038$^{\pm.000}$	0.031$^{\pm.001}$
Cross-person (SACA)	1.023$^{\pm.037}$	0.457$^{\pm.007}$	**0.351**$^{\pm.005}$	**0.308**$^{\pm.005}$	**0.035**$^{\pm.002}$	**0.029**$^{\pm.002}$

5.4 Component Ablation and Guidance

Components. Removing FiLM from SACA collapses CCS to -0.283 while leaving FID unchanged within noise (Tab. 6): FiLM is the only entry point of the social embeddings. Removing the input-dependent gate worsens FID ($1.023 \rightarrow 1.134$) and lowers CCS ($0.548 \rightarrow 0.486$), indicating that the gate helps integrate the social update into the text-grounded stream. Removing trait-space augmentation produces a memorization signature: CCS inflates to 0.767 while FID degrades to 6.011,

Table 6: **Component ablation.** Each row removes one component. Mean and 95% CI over 5 seeds.

Variant	FID ↓	R-3 ↑	MM Dist. ↓	CCS
Full model	**1.023**$^{\pm.037}$	0.457$^{\pm.007}$	<u>4.426</u>$^{\pm.038}$	0.548$^{\pm.003}$
− SACA gate	1.134$^{\pm.018}$	<u>0.469</u>$^{\pm.007}$	4.433$^{\pm.014}$	0.486$^{\pm.013}$
− SACA FiLM	<u>1.034</u>$^{\pm.030}$	**0.471**$^{\pm.006}$	**4.091**$^{\pm.041}$	−0.283$^{\pm.033}$
− trait augmentation	6.011$^{\pm.148}$	0.448$^{\pm.010}$	5.035$^{\pm.048}$	0.767$^{\pm.007}$

consistent with the model reproducing the profiles of the 47 training pairs instead of interpolating between them.

Guidance. Across the social-guidance sweep $w_s \in \{3.5, 5.5, 7.0, 9.0\}$ at $w_a = 3.5$ (Fig. 6; App. H), FID stays between 0.959 and 1.237, below every baseline in Tab. 1, while CCS rises monotonically from 0.387 to 0.586 and R-3 falls from 0.485 to 0.438. The social weight thus acts as a dial on how strongly the social condition is expressed.

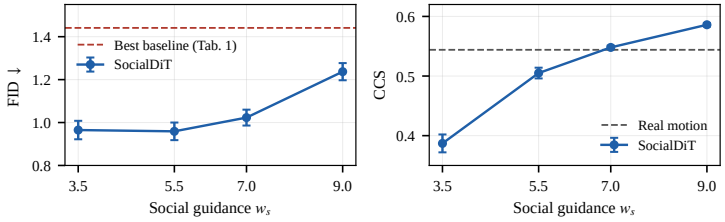


Figure 6: **Social guidance as a dial** ($w_a = 3.5$; 95% CI over 5 seeds). FID stays below the best baseline while CCS rises through the real-motion level.

6 Conclusion

We presented SocialDiT, a flow-matching diffusion transformer for text-driven dyadic interaction generation that injects per-actor Big Five personality and pair-level familiarity into cross-person attention. We showed that probe-based evaluation of social attributes is confounded when every pair carries one fixed label, and evaluated SocialDiT instead through interventions, controlled generation, evaluator-free geometry, and human judgment, under which it substantially improves realism over adapted baselines, responds to its social input, reproduces interpersonal distances most accurately, and is preferred in the user study.

Limitations and future work. Our experiments use a single dataset with six held-out pairs and a fixed attribute vocabulary (subject-level Big Five and four familiarity levels), and, like other data-driven generators, SocialDiT does not enforce physical constraints. Datasets with more annotated pairs or within-pair variation of social attributes, richer social descriptions, and contact-aware objectives are natural next steps (App. J).

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based assistants and an LLM-based coding agent) for the following tasks with required disclosure: providing feedback on the design of evaluation analyses (e.g., the group-wise probe analyses and the evaluator-free geometry metrics), helping interpret experimental results, assisting with translation of research material, and implementing parts of the evaluation code. We have not used generative AI tools to generate data sets or to formulate mathematical claims. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the manuscript, restructuring the paper, checking reference metadata, and searching the literature. We have reviewed all AI-assisted work: AI-assisted code was inspected by the authors and its outputs re-verified (e.g., by re-running the evaluation pipeline and auditing forward kinematics), and all text and reported numbers were checked against our experiment logs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We use the publicly released Inter-X dataset (Xu et al., 2024a) under its terms of use. Its personality and familiarity annotations describe real participants; we use them only as conditioning labels and do not infer the personality of any individual. As with other generative models of human behavior, outputs may reflect associations present in the training data. The user study involved twenty volunteers who were informed of the task (rating generated 3D character motion) and gave informed consent.

REPRODUCIBILITY STATEMENT

Architecture, optimization, conditioning dropout, sampling, and evaluation settings are given in App. A; the motion VAE in App. C; the condition encoders in App. D; the subject-disjoint split, including the identities of the test groups, in App. B; the probe, intervention, and geometry protocols in App. E and App. F; the injection-site variants in App. G; and the complete guidance sweep in App. H. All reported numbers are averaged over fixed seed sets (App. A). We will release code, trained VAE and SocialDiT weights, latent normalization statistics, split definitions, and evaluation scripts.

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions On Graphics (TOG)*, 39(4):64–1, 2020.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1342–1350, 2020.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-react: Online reaction policy for two-character interaction generation. *arXiv preprint arXiv:2502.20370*, 2025.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18000–18010, 2023.

Funda Durupinar. Personality-driven gaze animation with conditional generative adversarial networks. *arXiv preprint arXiv:2012.02224*, 2020.

Arjan Egges, Sumedha Kshirsagar, and Nadia Magnenat-Thalmann. Generic personality and emotion simulation for conversational agents. *Computer animation and virtual worlds*, 15(1):1–13, 2004.

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

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, volume 42, pp. 206–216. Wiley Online Library, 2023.

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

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. In *International Conference on Learning Representations*, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Siyeol Jung and Taehwan Kim. Difflistener: Discrete diffusion model for listener generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2151–2162, 2023.

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*, 132(9):3463–3483, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Kris Liu, Jackson Tolins, Jean E Fox Tree, Michael Neff, and Marilyn A Walker. Two techniques for assessing virtual agent personality. *IEEE Transactions on Affective Computing*, 7(1):94–105, 2015.

Yiyi Ma, Yuanzhi Liang, Xiu Li, Chi Zhang, and Xuelong Li. Intersyn: Interleaved learning for dynamic motion synthesis in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12832–12841, 2025.

Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20395–20405, 2022.

Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on computer graphics and interactive techniques*, 4(3):1–17, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Yichen Peng, Jyun-Ting Song, Siyeol Jung, Ruofan Liu, Haiyang Liu, Xuangeng Chu, Ruicong Liu, Erwin Wu, Hideki Koike, and Kris Kitani. Dyadit: A multi-modal diffusion transformer for socially favorable dyadic gesture generation. *arXiv preprint arXiv:2602.23165*, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2in: Leveraging individual information to generate human interactions. *arXiv preprint arXiv:2404.09988*, 2024.

Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*, 26: 10420–10430, 2024.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Pablo Ruiz-Ponce, Sergio Escalera, José García-Rodríguez, Jiankang Deng, and Rolandos Alexandros Potamias. Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23559–23569, 2026.

Maksim Siniukov, Di Chang, Minh Tran, Hongkun Gong, Ashutosh Chaubey, and Mohammad Soleymani. Ditailistener: Controllable high fidelity listener video generation with diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11991–12001, 2025.

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

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael J. Black, Daniel Holden, and Carsten Stoll. HUMOS: Human motion model conditioned on body shape. In *European Conference on Computer Vision*, 2024.

Kengo Uchida, Takashi Shibuya, Yuhta Takida, Naoki Murata, Julian Tanke, Shusuke Takahashi, and Yuki Mitsufuji. Mola: Motion generation and editing with latent diffusion enhanced by adversarial training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2901–2910. IEEE, 2025.

Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

Lipeng Wang, Hongxing Fan, Haohua Chen, Zehuan Huang, and Lu Sheng. Intermoe: Individual-specific 3d human interaction generation via dynamic temporal-selective moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2137–2145, 2026a.

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 Computer Vision and Pattern Recognition Conference*, pp. 7169–7178, 2025.

Zhongju Wang, Beier Wang, Yatao Bian, Pichao Wang, Zhi Wang, Daoyi Dong, Hongdong Li, Huadong Mo, and Zhenhong Sun. Social structure matters in 3d human–human interaction generation. *arXiv preprint arXiv:2606.24255*, 2026b.

Zizhao Wu, Yingying Sun, Yiming Chen, Xiaoling Gu, Ruyu Liu, and Jiazhou Chen. Intermamba: Efficient human-human interaction generation with adaptive spatio-temporal mamba. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023.

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

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

Lianying Yin, Yijun Wang, Tianyu He, Jinming Liu, Wei Zhao, Bohan Li, Xin Jin, and Jianxin Lin. Emog: Synthesizing emotive co-speech 3d gesture with diffusion model. *arXiv preprint arXiv:2306.11496*, 2023.

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

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023b.

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

Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European conference on computer vision*, pp. 405–421. Springer, 2024.

Chen Zhu, Buzhen Huang, Zijing Wu, Binghui Zuo, and Yangang Wang. E-react: Towards emotionally controlled synthesis of human reactions. *arXiv preprint arXiv:2508.06093*, 2025.

A Implementation Details

Dataset and splits. Inter-X (Xu et al., 2024a) contains 11,388 two-person interaction sequences over 59 groups. To prevent identity leakage we split at the *group* level so that no subject appears in more than one split: 47 groups for training (9,005 motions), 6 for validation (1,155), and 6 for testing (1,228), stratified by familiarity level. Sequences are truncated or padded to 152 frames at 30 fps.

Architecture. SocialDiT stacks $N = 8$ blocks with hidden dimension $D = 512$, $H = 8$ attention heads, MLP ratio 4, and dropout 0.1. Captions are encoded by a frozen CLIP ViT-L/14 (336) text encoder (Radford et al., 2021) followed by a trainable linear projection to D. AdaLN-Zero modulates self-attention, the SACA residual gate, and the feed-forward layers; text cross-attention uses a standard LayerNorm and residual connection. The social attributes enter only through FiLM inside SACA. The motion VAE is described in App. C.

Training. We train for 500 epochs with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, no weight decay), a linear warm-up over the first 1,000 steps followed by a constant learning rate of 10^{-4}, gradient clipping at 1.0, and an EMA of the weights with decay 0.9999, in FP32 on a single NVIDIA RTX 3090 with batch size 32. The flow time is sampled from $\mathcal{U}(0, 1)$. The objective is $\mathcal{L}_{\mathrm{FM}} + 0.5\,\mathcal{L}_{\mathrm{vel}}$ (Eq. 2). Checkpoints are selected on the validation split.

Conditioning dropout. With probability 0.15 both caption and social attributes are dropped (unconditional mode), with probability 0.45 only the social attributes are dropped (action-only mode), and the remaining 40% of samples use full conditioning. Dropped inputs are replaced by learnable null embeddings.

Inference. We integrate the learned flow with 50 Euler steps from Gaussian noise and apply dual-axis guidance (Eq. 11) with $(w_a, w_s) = (3.5, 7.0)$ unless stated otherwise; App. H explains how this setting was chosen. Latents are denormalized and decoded by the frozen VAE decoder.

Evaluation. FID, R-precision, MM-Distance, and Diversity use the pretrained motion–text evaluator of InterMask (Javed et al., 2025). Main results average 20 generation runs (seeds 42–61) and all other experiments 5 runs (seeds 42–46). All $\pm$ values are 95% confidence intervals over these runs. Diversity is the average distance between 1,000 randomly selected pairs of generated samples.

B Subject-Disjoint Split and Test-Set Trait Distribution

To our knowledge, Inter-X (Xu et al., 2024a) is the only public dyadic motion dataset that combines per-subject Big Five annotations with pair-level familiarity. The training split contains the trait profiles of only 47 subject pairs, leaving substantial regions of the five-dimensional Big Five space uncovered. The six test groups are G016, G020, G024, G029, G041, and G048. Three of them (G020, G024, G048) include trait values outside the training mean $\pm$ standard deviation band on at least one dimension, while the other three (G016, G029, G041) lie mostly within it. G020 and G024 deviate in opposite directions on Conscientiousness and Extraversion: G020 lies above the band on both, G024 below. The test set therefore requires both interpolation and extrapolation in trait space (Fig. 7).

C Motion VAE

Architecture. We follow the skeleton-aware spatio-temporal VAE design applied to Inter-X in InterMoE (Wang et al., 2026a). We use a per-joint linear input embedding, a hierarchical skeleto-temporal convolutional encoder and decoder with fixed skeleton-aware spatial pooling over hand-crafted body-part groupings, and per-spatial-token diagonal Gaussian latent heads. Specifically, we (i) define the skeletal adjacency for 56 joints (55 body, hand, and face joints plus an auxiliary composite joint encoding global translation and per-frame velocity); (ii) use a three-level pooling hierarchy ($56 \to 14 \to 7 \to 7$) that yields seven spatial tokens corresponding to body parts (root, left/right lower body, spine, left/right upper body, head), which we call joint groups; and (iii) use a per-token feature dimension $d_z = 128$. The encoder downsamples time by a factor of 8, yielding latents of shape $(T_{\mathrm{lat}}, 7, 128)$ per participant.

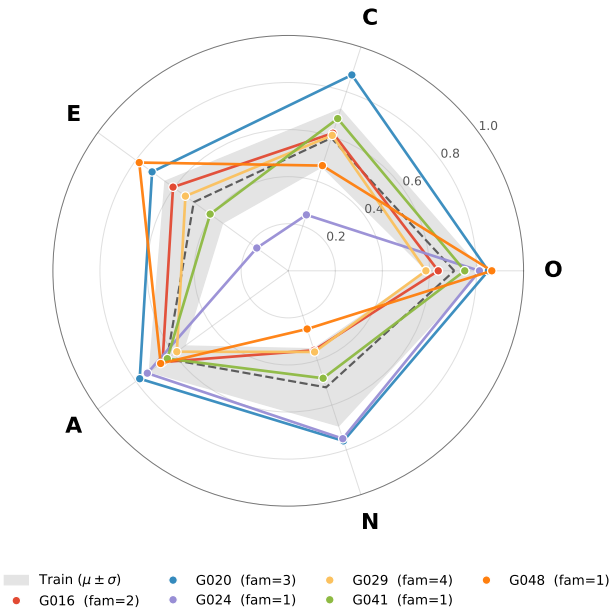


Figure 7: **Big Five profiles of the six test groups** relative to the training distribution (gray: training mean $\pm$ standard deviation). Each polygon is one test group, labeled with its familiarity level.

Normalization. Each per-participant latent is normalized per (joint group, channel) with statistics computed on the training split; SocialDiT operates in this normalized space, and outputs are denormalized before decoding.

Training. The VAE is trained from scratch on the training split with AdamW, using a full reconstruction term, position and velocity terms, and a Gaussian KL term, plus a geodesic rotation loss on the 6D joint rotations; reconstruction terms use L1. We do not condition on SMPL-X body shape. Checkpoints are selected by validation reconstruction FID.

Finite-difference operator. Δ_τ in Eq. 2 is the temporal first difference applied per joint group j after reshaping the token sequence to (T_{lat}, J, d_z): $\Delta_\tau f_{\tau,j} = f_{\tau+1,j} - f_{\tau,j}$.

D Condition Encoders

Text encoder. We use the frozen CLIP ViT-L/14 (336) text encoder (Radford et al., 2021), followed by a trainable linear projection to the hidden dimension. Both the token-level features E_c^{tok} (keys and values of text cross-attention) and the EOS-pooled embedding e_c^{pool} (global condition) are used.

Personality encoder (E_{pers}). Each Big Five dimension is encoded by its own two-layer MLP, producing five trait tokens. A learnable trait-identity embedding is added to each token, and the tokens are aggregated by single-query multi-head attention pooling followed by LayerNorm, which preserves trait-specific contributions before pooling.

Familiarity encoder (E_{fam}). The familiarity label $f \in \{0, 1, 2, 3\}$ is both categorical and ordinal (stranger to close). E_{fam} sums a learnable categorical embedding of f and a linear projection of $f/3$, followed by LayerNorm.

E PROBE AND INTERVENTION PROTOCOLS

Motion features. Unless stated otherwise, probes operate on the 512-dimensional motion embeddings of the pretrained InterMask evaluator (Javed et al., 2025), the encoder used for FID and R-precision. Generated and real motions pass through the same frozen encoder.

Condition-consistent separability (CCS). For each Big Five dimension, a Ridge regressor ($\alpha = 1.0$) on standardized features predicts the trait averaged over the two participants. We use 5-fold cross-validation with sample-level random folds, report the mean R^2 over folds, and average over the five dimensions. Because every test pair appears in both training and test folds with a constant label, this score measures whether a source expresses each condition as a consistent, linearly distinguishable style—not whether traits can be decoded for unseen pairs.

Group-wise evaluation. The group-wise variant uses the same features and regressor but forms folds by test group, so that the pairs in each test fold are unseen by the probe; the real-motion reference is recomputed with the identical protocol. All sources fall below zero (Tab. 7): real motion -0.534, SocialDiT $-0.370^{\pm.023}$, InterGen $-0.304^{\pm.023}$, and InterMask $-0.253^{\pm.016}$ (5 seeds).

Table 7: **Personality-probe R^2 under sample-level vs. group-wise cross-validation** (test split). Group-wise folds hold out entire pairs. Mean and 95% CI over 5 seeds; real motion is a single realization. † CCS from Tab. 1

Source	Sample-level CV	Group-wise CV
Real motion	0.544	-0.534
InterGen (w. social)	$0.367^{\dagger}$	$-0.304^{\pm.023}$
InterMask (w. social)	$0.501^{\dagger}$	$-0.253^{\pm.016}$
SocialDiT	$0.548^{\pm.003}$	$-0.370^{\pm.023}$

Controlled familiarity probe. For each test sample, we generate four motions by fixing the personality input at the dataset mean and setting the familiarity input to each of $\{0, 1, 2, 3\}$. A logistic-regression classifier ($C = 1.0$) predicts the familiarity input from motion features under 5-fold stratified cross-validation; we report mean accuracy and macro-F1 over folds. Because each caption appears with all four labels, the classifier cannot succeed by recognizing captions or pairs. The resulting confusion matrices are shown in Fig. 5.

Interventions. With captions fixed, we (i) replace the social input by its learned null embedding ("removed"), (ii) reassign Big Five vectors across test pairs using a fixed cross-group mapping computed once over the test set, (iii) reassign each sample's familiarity to a different, randomly chosen level, or (iv) apply both (ii) and (iii).

F EVALUATOR-FREE GEOMETRY AND GROUND CONTACT

Interpersonal distance. For every frame we compute two distances from 3D joint positions: the pelvis-to-pelvis (root) distance and the closest distance between any joint of one participant and any joint of the other. Each is compared with real motion in two ways. (1) *MAE*: the absolute difference between the sequence-mean distance of a generated interaction and that of the corresponding real interaction, averaged over the test set; it measures how accurately the spacing of each interaction is reproduced. (2) *JSD*: the Jensen–Shannon divergence between the generated and real distance distributions pooled over all frames; it measures how realistic the overall range of spacings is.

Ground contact. We use the official implementation of the penetration, floating, and skating metrics of HUMOS (Tripathi et al., 2024), computed from body geometry. As HUMOS notes, the three metrics trade off against one another (e.g., reducing penetration can increase floating), so we report all three.

G Injection-Site Variants

All variants share the SocialDiT backbone, the VAE, the social encoders (Eq. 4), the training schedule, trait augmentation, the conditioning dropout, and the guidance setting; they differ only in where the social embeddings s_A, s_B enter the network. **Per-stream (AdaLN).** Each participant's social embedding is routed into the AdaLN modulation of that participant's stream; cross-person attention is standard, without FiLM. **Shared (global).** Both social embeddings are fused into the caption-level global condition g shared by the two streams. **Cross-person (SACA).** The social embeddings modulate cross-person attention through FiLM (Sec. 3.3); this is SocialDiT.

H Guidance Sensitivity

We sweep the action and social guidance weights to characterize how SocialDiT trades motion quality, caption alignment, and the strength of social expression (Tab. 8; 5 seeds, test split).

How the default was chosen. The sweep was run on the test split, and the default $(w_a, w_s) = (3.5, 7.0)$ is a balance point on it rather than the test optimum of any single metric (FID is lowest at $w_s = 5.5$, R-3 at $w_s = 3.5$, and CCS at $w_s = 9.0$). For this reason we report the complete sweep and base our claims on those that hold across it; in particular, SocialDiT's FID is below that of every baseline in Tab. 1 at every swept setting.

Behavior across the sweep. Increasing w_s from 3.5 to 9.0 (with $w_a = 3.5$) raises CCS monotonically from 0.387 to 0.586 and lowers R-3 from 0.485 to 0.438, while FID is U-shaped with its minimum at $w_s = 5.5$. The social residual direction thus provides a controllable trade-off between motion quality and the strength of social expression. Lowering w_a improves FID but reduces R-precision, indicating weaker caption grounding, and raising it gives limited alignment gains at a cost in FID.

Table 8: **Dual-axis guidance sensitivity** (test split, 5 seeds). Left: w_s varied with $w_a = 3.5$. Right: w_a varied with $w_s = 7.0$.

w_a/w_s	FID ↓	R-3 ↑	MM Dist. ↓	CCS
3.5 / 3.5	$0.965^{\pm.043}$	$\mathbf{0.485}^{\pm.006}$	$\mathbf{4.219}^{\pm.033}$	$0.387^{\pm.015}$
3.5 / 5.5	$\mathbf{0.959}^{\pm.041}$	$0.472^{\pm.009}$	$4.340^{\pm.040}$	$0.505^{\pm.009}$
3.5 / 7.0 (default)	$1.023^{\pm.037}$	$0.457^{\pm.007}$	$4.426^{\pm.038}$	$0.548^{\pm.003}$
3.5 / 9.0	$1.237^{\pm.040}$	$0.438^{\pm.005}$	$4.548^{\pm.032}$	$0.586^{\pm.003}$

w_a/w_s	FID ↓	R-3 ↑	MM Dist. ↓	CCS
1.5 / 7.0	$\mathbf{0.670}^{\pm.039}$	$0.404^{\pm.007}$	$4.887^{\pm.033}$	$0.691^{\pm.008}$
2.5 / 7.0	$0.745^{\pm.034}$	$0.444^{\pm.007}$	$4.531^{\pm.035}$	$0.621^{\pm.007}$
3.5 / 7.0 (default)	$1.023^{\pm.037}$	$0.457^{\pm.007}$	$4.426^{\pm.038}$	$0.548^{\pm.003}$
4.5 / 7.0	$1.300^{\pm.053}$	$\mathbf{0.468}^{\pm.004}$	$\mathbf{4.406}^{\pm.040}$	$0.480^{\pm.012}$

I User Study

Sampling. We randomly sampled 25 test captions, distributed across the six test groups to cover their personality and familiarity configurations. For each caption we generated clips with SocialDiT and the two social-conditioned baselines (InterGen (Liang et al., 2024) and InterMask (Javed et al., 2025)) under the same text and social input.

Design. Participants compared SocialDiT with each baseline on three axes: caption alignment (whether the motion matches the described action), motion quality (naturalness and realism), and perceived familiarity (whether the relational closeness between the two participants matches the conditioned familiarity level). Each comparison showed two side-by-side rendered clips in random order, and participants chose the preferred clip on each axis, with a tie option. The interface is shown in Fig. 8.

Participants. Twenty participants completed all comparisons. They were informed of the nature of the task (rating generated 3D character motion) and provided informed consent.

Results. Win/tie/loss percentages are reported in Tab. 4. The samples of rendered clips used in the study are included in the supplementary material.

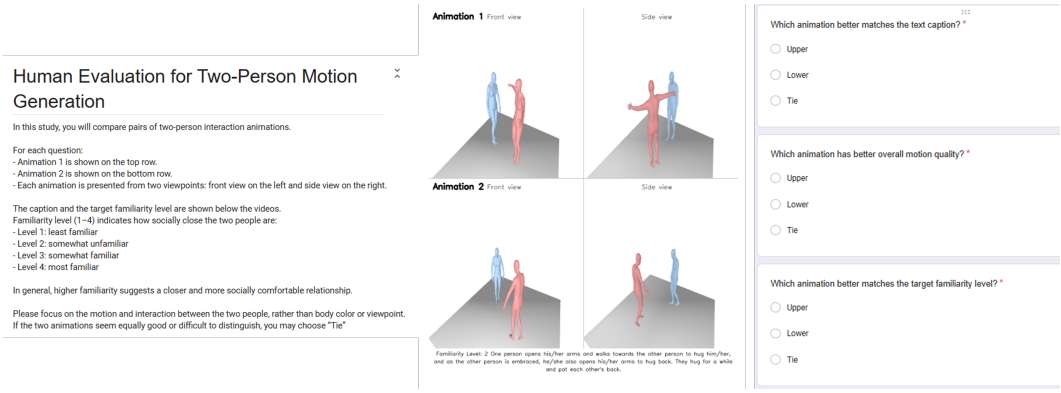


Figure 8: **User study interface.** Two side-by-side clips per comparison; participants chose the preferred clip on each axis, with a tie option.

J Limitations, Applications, and Future Work

Scope. Our experiments use Inter-X under a subject-disjoint split with six held-out pairs. Datasets with more annotated pairs, or with social attributes that vary within a pair, would allow finer-grained tests of how specific attribute changes should alter behavior.

Attribute vocabulary. Inter-X provides subject-level Big Five scores and four familiarity levels, which cannot distinguish relationships at the same level (e.g., colleague vs. sibling), asymmetric relations, or context-dependent variation within a pair. Richer, open-ended descriptions of interaction style (e.g., generated by language or vision–language models) could provide more flexible control.

Physical plausibility. Like other data-driven interaction generators, SocialDiT does not enforce physical constraints; contact-aware objectives or physics-based post-processing are natural extensions.

Applications. Social control is useful for (1) virtual agents and social simulation, where interaction styles should adapt to personalities and relationships; (2) animation authoring, where the same caption can be re-rendered under different social settings (Fig. 1); (3) human–robot interaction, where proxemics and contact behavior should depend on the relationship with the partner; and (4) data augmentation for social-perception models, particularly for rare trait and familiarity combinations.