SOCIAL STRUCTURE MATTERS IN 3D HUMAN-HUMAN INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Although text-to-motion generation has achieved strong progress in synthesizing realistic single-person motions from language, extending it to text-driven 3D human-human interaction (HHI) remains non-trivial, as HHI requires modeling the underlying **social structure** that governs phase progression, actor roles, and inter-actor coordination. In this paper, we formulate HHI generation as a social structure modeling and grounding problem: the model must first infer how an interaction unfolds and how the two actors coordinate their roles, and then realize this structure as continuous, physically plausible, and partner-aware 3D motion. To study how such structure should be modeled, we first examine the capability boundary of large language models (LLMs) for HHI generation. Our analysis shows that LLMs can *think* by recovering phase decompositions and partner-aware roles, but cannot directly *move*, as they fail to generate dynamic, physically plausible, and interaction-aware motion. This motivates our planner-executor paradigm, **Think with LLM, Move with Motion Skill**. The LLM planner converts implicit interaction semantics into motion-aligned social supervision by decomposing interactions into phases, assigning partner-aware actor roles, and aligning them with motion sequence. The motion executor then grounds the planned social structure into coordinated two-person motion by adapting a pretrained solo motion model with LoRA, previous-phase self-conditioning, and ego-relative partner conditioning. Together, our Solo-to-Social framework bridges social organization and motion realization, producing 3D HHI with improved phase consistency, role alignment, and partner-aware coordination.

1 INTRODUCTION

Text-to-motion generation has made rapid progress in synthesizing realistic single-person 3D motion from natural language Sui et al. (2026); Sahili et al. (2025); Fan et al. (2025). Recent large-scale motion generators further show that pretrained motion backbones can learn strong atomic motion priors from broad single-person motion data Team (2025); Azadi et al. (2023). As generative models move toward embodied AI Liang et al. (2025), multi-agent collaboration Wu et al. (2025a), and social robotics Matheus et al. (2025), the research focus is naturally shifting from individual motion synthesis to text-driven 3D human-human interaction (HHI) generation. Given a global interaction description, HHI generation requires not only realistic individual motions but also coherent coordination between two actors over time.

Existing HHI generation methods have made promising progress by jointly modeling two-person motion, composing individual motion priors, or introducing interaction-aware generation mechanisms Ruiz-Ponce et al. (2024; 2025a); Shafir et al. (2024); Liang et al. (2024); Javed et al. (2025); Wang et al. (2025). However, treating HHI as a direct extension from one actor to two overlooks a key property of HHI: it is not merely a spatial combination of two plausible individual motions, but is organized by an underlying **social structure**, as shown in Fig. 1. We define social structure as the latent interaction organization governing two dimensions: *phase progression*, which describes temporal stages such as approach, contact, and release; and *partner-aware coordination*, which captures the asymmetric but coupled roles of the two actors, such as initiator and receiver or attacker and defender. Without such structure, generated motions may appear plausible for each actor but fail as an interaction: the actors may approach at inconsistent times, miss the contact moment, face the wrong direction, or execute incompatible roles.

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.

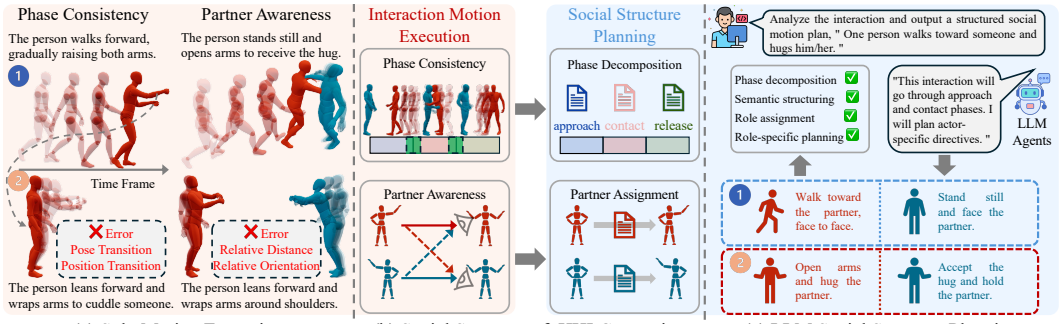

Figure 1: Social structure of text-driven HHI generation. (a) Solo motion execution provides strong intra-personal motion priors but lacks interaction-level coordination. (b) HHI requires social structure along two dimensions: phase-level temporal organization and partner-aware coordination. (c) LLMs offer social planning abilities to make such structure explicit for interaction motion execution.

We therefore formulate HHI generation as a **social structure modeling and grounding** problem: the model must infer phase decomposition and partner-aware role assignment from a global language description and ground them into continuous, physically plausible, and partner-aware 3D motion. This formulation separates two complementary challenges: reasoning about *what* social structure to follow, and executing *how* it should unfold as coordinated motion over time. The former concerns semantic interaction reasoning, whereas the latter requires temporally precise and geometrically grounded motion realization. Large Language Models (LLMs) are natural tools for the former, given their strong language understanding Team (2026), commonsense reasoning Merlo et al. (2025), and structured planning Zou et al. (2025) abilities. Yet whether they can serve as complete HHI generators remains unclear. We therefore conduct a diagnostic study by representing Skinned Multi-Person Linear (SMPL) Loper et al. (2015) parametric human motion in a token-like form and asking an LLM to generate atomic motions from text. The results reveal a clear separation between *thinking* and *moving*: the LLM can infer plausible phase decompositions and partner-aware roles, but fails to reliably produce continuous, dynamic, and physically plausible 3D interaction motion. Thus, LLMs are suitable as social structure planners, but not as direct motion executors.

Based on this insight, we propose a planner-executor paradigm for social-structure-centered HHI generation: **Think with LLM, Move with Motion Skill**. The LLM acts as a social structure planner, while a pretrained motion model provides the executable motion skill. Instead of using the LLM as a direct motion generator, we use it to convert implicit interaction semantics into explicit social structure supervision. Given a global interaction prompt and paired motion sequence, we decompose the interaction into phase-level units, assign partner-aware roles to both actors, and align each phase with its motion segment. This converts coarse HHI text-motion pairs into fine-grained, motion-aligned social annotations, making social structure a trainable bridge between language intent and motion execution. To ground this structure into continuous two-person motion, we introduce a *Solo-to-Social* (S2S) motion execution framework, which adapts a pretrained solo motion backbone into an interaction motion executor rather than training from scratch. S2S preserves the atomic motion prior learned from single-person data while adding the social coordination ability required by HHI. Phase-wise self motion conditioning uses previous-phase motion prefixes to encourage smooth transitions and long-range consistency, while partner-aware motion conditioning injects the partner's latest motion into the actor's ego-centric frame to model relative position, orientation, and interaction geometry. With parameter-efficient Low-Rank Adaptation (LoRA), these mechanisms turn a solo motion model into a socially aware executor for coherent two-person motion. Experiments on standard HHI benchmarks show improved text-motion alignment, phase consistency, and partner-aware coordination over existing baselines, while qualitative results demonstrate clearer phase progression, more role-consistent behaviors, and more plausible inter-person geometry. Overall, our framework treats social structure as the central abstraction of HHI generation, planned by LLM reasoning and grounded through motion skill adaptation.

The main contributions of this paper are summarized as follows:

- We identify **social structure** as a central abstraction for text-driven 3D HHI generation, and formulate HHI generation as a social structure modeling and grounding problem rather than a direct extension of solo motion generation.
- We analyze the capability boundary of LLMs for HHI generation, showing that LLMs are effective at social structure planning but inadequate for direct continuous 3D motion execution.
- We propose an LLM-based social structure planning strategy that reorganizes HHI datasets into fine-grained motion-aligned annotations with explicit phase structure and partner-aware roles.
- We introduce a Solo-to-Social motion execution framework that adapts a pretrained solo motion backbone into a socially aware interaction motion skill through phase-wise self conditioning, ego-relative partner conditioning, and parameter-efficient LoRA adaptation.

2 Related Work

Text-to-Motion Generation. Text-to-motion generation synthesizes single-person 3D motion from natural-language descriptions Zhu et al. (2023); Khani et al. (2025); Chen et al. (2025a). Early methods adopt GANs Amballa et al. (2025); Wang et al. (2020), VAEs Guo et al. (2020); Zhong et al. (2023); Zhang et al. (2023), or autoregressive Transformers Xu et al. (2023); Jiang et al. (2024); Wang et al. (2024); Zhu et al. (2025), while recent diffusion models and large-scale pretrained motion backbones achieve stronger realism, diversity, and text-motion alignment Zhang et al. (2024); Petrovich et al. (2023); Azadi et al. (2023); Team (2025). These models learn strong atomic motion priors but do not explicitly model the temporal and relational structure required for HHI; we build on such priors and adapt them to socially structured two-person motion execution.

Human-Human Interaction and Reaction Generation. Human-human interaction generation extends motion synthesis from individual behavior to coordinated two-person motion Stergiou & Poppe (2018); Fan et al. (2025); Sui et al. (2026). A related line of work studies *reaction generation*, where one actor's observed motion is provided and the model predicts the partner's response Xu et al. (2024b); Siyao et al. (2022); Yao et al. (2023); Siyao et al. (2024); Li et al. (2024); Tan et al. (2025); Cen et al. (2025). For example, Think-Then-React (TTR) Tan et al. (2025) introduces semantic reasoning before reaction generation, but remains conditioned on an observed actor and generates only the reacting partner. This differs from text-driven HHI generation, where neither actor motion is observed and both motions must be jointly inferred from language. Text-driven HHI methods instead generate paired motions directly from interaction descriptions Ghosh et al. (2024); Chopin et al. (2023); Xu et al. (2024a); Ruiz-Ponce et al. (2025b); Park et al. (2025); Goel et al. (2022); Wu et al. (2025b); Geng et al. (2026); Ruiz-Ponce et al. (2025a). Representative approaches improve two-person generation through joint modeling, individual-prior composition, or interaction-aware architectures, including ComMDM Shafir et al. (2024), in2IN Ruiz-Ponce et al. (2024), InterGen Liang et al. (2024), InterMask Javed et al. (2025), and TIMotion Wang et al. (2025). Recent LLM-assisted methods such as in2IN and Text2Interact enrich generation with global or actor-level semantic descriptions, while role-aware methods such as RIG Tanaka & Fujiwara (2023) explicitly condition generation on actor roles. However, these approaches do not jointly expose ordered phase progression and phase-varying coupled roles as a motion-aligned representation: actor descriptions are typically sequence-level, roles are predefined, or interaction structure remains implicit in the motion generator. In contrast, we formulate HHI generation as social structure modeling and grounding, where phase progression and partner-aware coordination serve as explicit representations.

LLMs for Motion Reasoning and Planning. LLMs exhibit strong language understanding, commonsense reasoning, and structured planning abilities Team (2026); Merlo et al. (2025); Zou et al. (2025), making them useful for decomposing high-level human intentions into organized plans. Unlike motion-specialized models such as MotionGPT Wang et al. (2024) and MotionLLM Chen et al. (2025b), we use a general-purpose LLM only for social reasoning and planning. For HHI, such reasoning can reveal how an interaction unfolds and how actors coordinate their roles. However, continuous 3D motion execution requires kinematic precision, temporal smoothness, and physically plausible inter-person geometry, which remain difficult for LLMs to generate directly. We therefore use LLMs not as motion generators, but as social structure planners: the LLM converts implicit interaction semantics into phase-level and role-level supervision, while a pretrained motion model grounds this structure into coordinated two-person motion. This planner-executor view connects the semantic reasoning strength of LLMs with the motion prior of specialized motion generators.

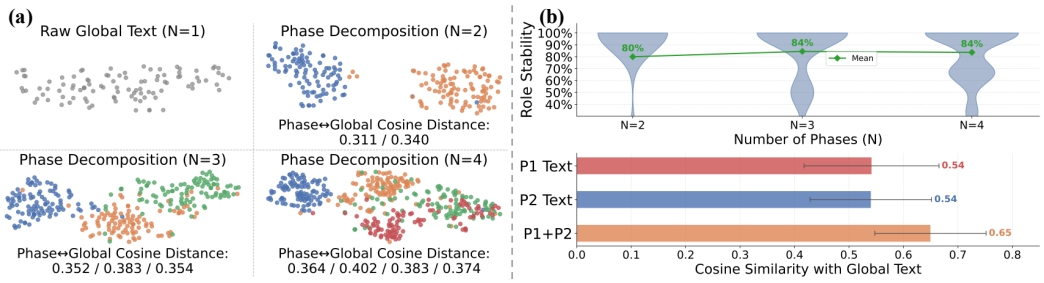

Figure 2: LLM (Qwen3.5) capability analysis in modeling social structures. (a) t-SNE of phase decomposition. The LLM can recover latent phase progression from global text. (b) Role assignment. Decoupled individual semantics are cross-phase consistent and aligned with the global intent. These results indicate that LLMs can model social structure of HHI well at the semantic level.

3 Methodology

3.1 LLM Capability Boundary: Social Structure Planning vs. Motion Execution

Given a natural-language description y, text-driven HHI generation aims to synthesize a two-person motion sequence

$$p\left(\mathbf{X}^{1:T} \mid y\right), \qquad \mathbf{X}^{1:T} = \left\{\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)}\right\}_{t=1}^{T}, \tag{1}$$

where $\mathbf{x}_t^{(i)} \in \mathbb{R}^D$ denotes the SMPL-based motion representation of actor i at frame t.

We define the social structure as a latent intermediate variable S that captures sufficient interaction organization, the text-driven HHI generation problem can be reformulated as

$$p(\mathbf{X}^{1:T} \mid y) = \sum_S p(\mathbf{X}^{1:T}, S \mid y) = \sum_S p(\mathbf{X}^{1:T} \mid S, y)\, p(S \mid y) \approx \sum_S p_\theta(\mathbf{X}^{1:T} \mid S)\, p_\phi(S \mid y), \tag{2}$$

where $p_\phi(S \mid y)$ models the social structure planning while $p_\theta\left(\mathbf{X}^{1:T} \mid S\right)$ models the motion execution. However, since whether LLMs can directly serve for both $p_\phi(S \mid y)$ and $p_\theta\left(\mathbf{X}^{1:T} \mid S\right)$ still remains unclear, we first conduct a diagnostic study along these two dimensions. To make such analysis reproducible and experimentally controlled, we employ Qwen3.5 Team (2026), an open-weight LLM with standard local deployment support, which allows the analysis to be conducted under fixed model parameters and inference settings, without introducing additional variability from service-side API updates or access policies.

Social Structure Planning. We first test whether LLMs can model $p_\phi(S \mid y)$ to capture latent social structure from global interaction text. As shown in Fig. 2(a), Qwen3.5 decomposes a raw global prompt into increasingly separable semantic phases, indicating its ability to impose temporal structure on a coarse interaction description. Fig. 2(b) further shows that Qwen3.5 can infer asymmetric actor roles with high cross-phase consistency, and paired actor descriptions align more closely with the global text than either description alone. These results suggest that LLMs are effective social planners for HHI, especially in resolving phase progression and partner-conditioned actor roles.

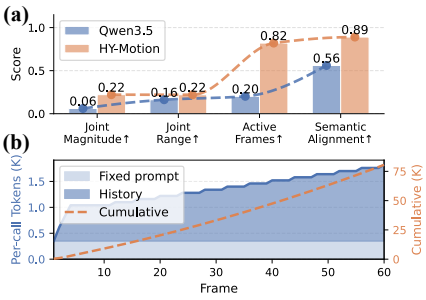

Figure 3: (a) Atomic motion generation capacity. (b) Reasoning efficiency

Direct Motion Execution. We then examine whether LLMs can model $p_\theta\left(\mathbf{X}^{1:T} \mid S\right)$ for motion execution. As shown in Fig. 3(a), Qwen3.5 remains weaker than a dedicated motion model even when motion is represented in the structured SMPL space. It produces limited active motion, with low joint magnitude, narrow motion range, and fewer active frames. Although the generated motion may retain

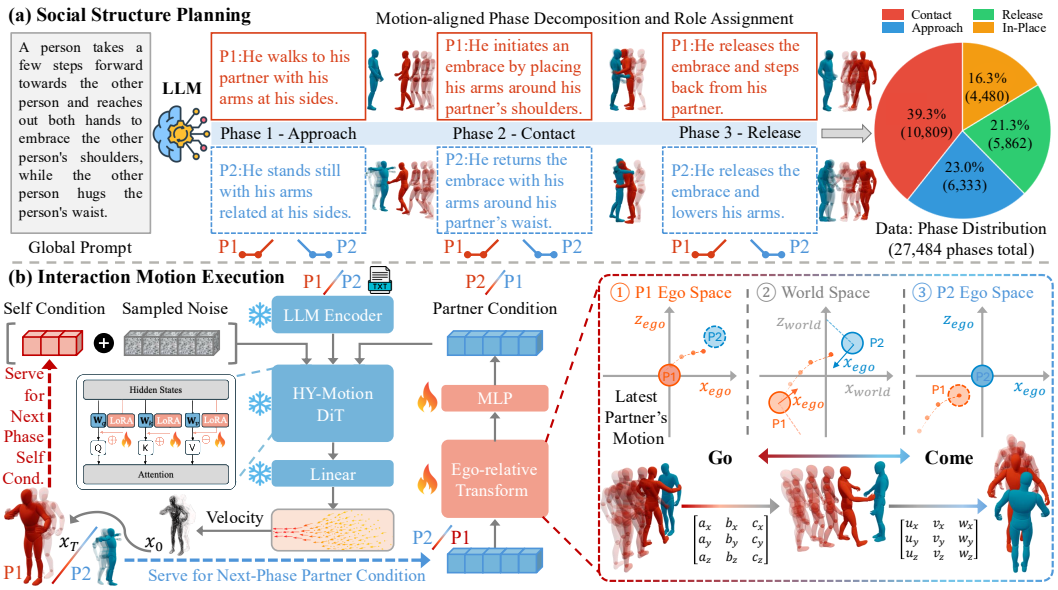

Figure 4: Overview of our proposed planner-executor paradigm for social-structure-centered HHI generation. (a) The LLM serves as a social structure planner which recovers plausible phase decompositions and partner-aware role assignments from global prompt. (b) The motion skill is built on a solo motion backbone equipped with self and partner conditioning for motion execution.

partial semantic alignment with the prompt, it lacks the continuity and dynamics required for physically plausible HHI, with visualizations provided in **Appendix C**. Moreover, Fig. 3(b) further shows that autoregressive LLM-based motion execution incurs substantial cumulative context growth, making long-horizon motion generation increasingly inefficient. These results reveal a clear capability boundary: LLMs can expose the social structure of HHI, but cannot reliably execute it as continuous interaction motion. This motivates a planner-executor paradigm for social-structure-centered HHI generation: **Think with LLM, Move with Motion Skill**, as shown in Fig. 4. In this paradigm, the LLM serves as a social structure planner, while a pretrained solo motion model serves as an executable motion skill. In the following sections, we introduce details of these two components.

3.2 Motion-aligned Social Structure Planning

Social structure S provides the temporal and role-level organization needed for coherent HHI generation, but such supervision is not explicitly available in existing HHI datasets. Most datasets pair a global interaction description with a full two-person motion sequence, leaving phase progression and partner-aware actor roles implicit. To obtain the social structure S based on real motion data, we reconstruct existing HHI data into motion-aligned social supervision through our LLM-based social structure planning. To prevent the LLM from producing language-plausible but motion-inconsistent plans, we extract motion facts from real motion data and use it as constraints during planning.

Given a paired HHI motion sequence $\mathbf{X}^{1:T}$, we first detect phases that the interaction goes through

$$\mathcal{C} = f_{\mathrm{p}}(\mathbf{X}^{1:T}), \qquad \mathcal{C} = \{c_k\}_{k=1}^{K}, \quad c_k \in \mathcal{V}_{\mathrm{phase}}, \tag{3}$$

where c_k is the k-th intermediate phase that the interaction goes through, the phase vocabulary is $\mathcal{V}_{\mathrm{phase}} = \{\texttt{approach}, \texttt{contact}, \texttt{release}, \texttt{in-place}\}$, and the mapping $f_{\mathrm{p}}(\cdot)$ uses available InterX segment annotations and otherwise applies a forward state machine to motion-derived signals. More details are provided in **Appendix F**. In this way, the full sequence is decomposed into an explicit phase progression, exposing progress of the interaction, as shown in Fig. 4(a).

For each phase c_k, we extract a structured motion fact representation from the phase-wise motion sequence, represented as

$$\mathcal{M}_k = \{m_k^{(1)}, m_k^{(2)}\}, \quad m_k^{(i)} = \left(\delta_k,\, u_k,\, q_k,\, \mathbf{a}_k^{(i)},\, \mathbf{l}_k^{(i)}\right), \quad i \in \{1, 2\}, \tag{4}$$

where δ_k denotes inter-person distance evolution, u_k denotes the motion initiator, q_k captures contact semantics, $\mathbf{a}_k^{(i)}$ summarizes the motion direction and facing state of actor i, and $\mathbf{l}_k^{(i)}$ describes limb-level cues such as reaching, arm lifting, bending, and foot activity. Motion facts are extracted from 3D motion geometry, with heuristic tie-breaking for ambiguous initiator cues.

Conditioned on the global interaction text y, phase label c_k, and motion facts $\mathcal{M}_k$, the LLM planner reasons actor-specific semantic descriptions,

$$\mathcal{Y}_k \sim p_\phi(\mathcal{Y}_k \mid y,\, c_k,\, \mathcal{M}_k)\,, \qquad \mathcal{Y}_k = (y_k^{(1)}, y_k^{(2)}), \tag{5}$$

where $(y_k^{(1)}, y_k^{(2)})$ represents actor-specific prompts.

Finally, aggregating all phases yields a motion-aligned social structure

$$S = \{(c_k, y_k^{(1)}, y_k^{(2)})\}_{k=1}^{K}. \tag{6}$$

Although the final S contains only phase labels and actor-specific semantic descriptions, it is constructed under phase-wise motion fact constraints and is therefore aligned with the underlying motion sequence. Thus, S is not merely a plausible textual plan, but a motion-aligned intermediate representation that connects phase progression and actor-specific role semantics to the observed interaction motion, making it directly usable for downstream motion execution.

During training, paired motion facts are available and used to constrain the LLM planner. The resulting social structures are therefore guided by real motion evidence rather than free-form language plausibility alone, providing temporally aligned supervision for training the downstream executor. This process also serves as a fine-grained HHI annotation pipeline, converting coarse global text-motion pairs into phase-level and role-aware training annotations. At inference time, ground-truth motion facts are unavailable, so the planner operates purely from the global interaction prompt. This allows the LLM to exploit its semantic generalization ability to infer plausible social structure directly from language, without being tied to a specific observed motion instance, thereby preserving both high-level interaction organization and generation diversity.

3.3 Interaction Motion Execution

Given the structured social plan recovered by the LLM planner, the remaining challenge is to ground it into continuous and coordinated 3D human-human motion. We therefore introduce a Solo-to-Social (S2S) motion execution framework, which adapts a pretrained solo motion backbone into a social interaction executor, as shown in Fig. 4(b). The core idea of S2S is to preserve the atomic motion prior learned from large-scale motion corpus, while injecting the missing mechanisms needed to execute social structure: previous-phase self-conditioning grounds phase progression into continuous motion, ego-relative partner conditioning grounds partner-aware coordination into inter-person geometry, and LoRA adaptation transfers the solo motion prior toward social execution.

Grounding Phase Progression via Self Conditioning. The social plan decomposes an interaction into temporally ordered phases. However, executing each phase independently would produce discontinuities at phase boundaries, breaking the intended phase progression. To ground phase progression into a continuous motion trajectory, we introduce phase-wise self motion conditioning. For phase $k > 1$, the self condition $\mathbf{h}_k^{(i)}$ contains the last e frames of phase $k-1$:

$$\mathbf{x}_k^{(i)}[1:e] = \mathbf{h}_k^{(i)} = \mathbf{x}_{k-1}^{(i)}[T_{k-1} - e + 1 : T_{k-1}], \tag{7}$$

where T_{k-1} denotes the duration of phase $k-1$. The model predicts the remaining frames conditioned on this anchor prefix, encouraging smooth transitions through access to its own recent motion history.

Grounding Partner-aware Coordination via Partner Conditioning. Beyond phase continuity, HHI requires each actor to move with awareness of the partner's behavior. A solo motion backbone operating on each actor independently cannot directly capture relative displacement, facing direction, contact timing, or interaction geometry. To ground partner-aware coordination, we condition the generation of actor i on the partner's motion context:

$$p_\theta\Big(\mathbf{x}_k^{(i)} \mid y_k^{(i)},\, c_k,\, \mathbf{h}_k^{(i)},\, \mathbf{p}_k^{(i)}\Big)\,, \qquad \mathbf{p}_k^{(i)} = R_{j\to i}\Big(\mathbf{s}_{k-1}^{(j)}\Big)\,, \qquad j \neq i, \tag{8}$$

Table 1: Comparison with existing text-driven two-person interaction motion generation methods (mean ± standard deviation over five seeds).

Method	R-Prec@1 (%) ↑	FID ↓	Interpenetration ↓	Contact ↑	Foot-slide ↓
RIG Tanaka & Fujiwara (2023)	12.87 ± 0.61	0.998 ± 0.005	0.150 ± 0.004	0.328 ± 0.006	0.338 ± 0.004
ComMDM Shafir et al. (2024)	15.21 ± 0.60	0.901 ± 0.004	0.189 ± 0.005	0.346 ± 0.010	0.334 ± 0.005
in2IN Ruiz-Ponce et al. (2024)	17.34 ± 0.52	0.779 ± 0.002	0.138 ± 0.005	0.353 ± 0.009	0.307 ± 0.005
InterGen Liang et al. (2024)	17.34 ± 0.68	0.757 ± 0.004	0.160 ± 0.007	0.378 ± 0.008	0.293 ± 0.003
InterMask Javed et al. (2025)	17.23 ± 0.59	0.761 ± 0.002	0.144 ± 0.005	0.364 ± 0.008	0.289 ± 0.004
TIMotion Wang et al. (2025)	17.07 ± 0.34	0.704 ± 0.003	0.137 ± 0.004	0.360 ± 0.009	0.256 ± 0.005
Text2Interact Wu et al. (2026)	17.72 ± 0.35	0.705 ± 0.003	0.122 ± 0.006	0.384 ± 0.009	0.241 ± 0.003
Social Planning + HYM1	20.25 ± 0.36	0.703 ± 0.004	0.110 ± 0.005	0.389 ± 0.010	0.232 ± 0.005
Ours (Social Planning + S2S)	$\mathbf{24.67 \pm 0.42}$	$\mathbf{0.661 \pm 0.003}$	$\mathbf{0.103 \pm 0.004}$	$\mathbf{0.397 \pm 0.009}$	$\mathbf{0.218 \pm 0.004}$

where $y_k^{(i)}$ is the actor-specific description, c_k is the phase label, and $\mathbf{s}_{k-1}^{(j)}$ is the partner's latest generated state after phase $k-1$. The transform $R_{j \to i}$ expresses this state in actor i's ego-centric frame. This ego-relative representation makes the partner's position, orientation, and motion dynamics directly observable to the current actor. During inference, S2S is instantiated as two coupled actor-wise executors, one for each actor. The two motions are rolled out synchronously in a phase-by-phase manner according to the planned phase sequence. After each phase, both actors update their states to condition the next phase; neither requires the partner's not-yet-generated motion in that phase.

Transferring Solo Motion Priors to Social Interaction. We employ HY-Motion 1.0 Team (2025), a state-of-the-art pretrained solo motion generation model, as the core backbone of our motion executor. Instead of training the HHI generator from scratch, we adapt this solo backbone for social interaction execution using parameter-efficient LoRA modules. The overall fine-tuning objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + w_{\text{smooth}}\mathcal{L}_{\text{smooth}} + w_{\text{dist}}\mathcal{L}_{\text{rel_dist}} + w_{\text{ori}}\mathcal{L}_{\text{rel_ori}}, \tag{9}$$

where $\mathcal{L}_{\text{FM}}$ is the original flow-matching objective of HY-Motion 1.0 and preserves intra-person motion quality. $\mathcal{L}_{\text{smooth}}$ penalizes abrupt frame-to-frame changes in the generated region, encouraging temporally smooth phase execution. $\mathcal{L}_{\text{rel_dist}}$ and $\mathcal{L}_{\text{rel_ori}}$ constrain the generated actor to match the target partner-relative distance and orientation, respectively, thereby grounding partner-aware coordination in inter-person geometry. More details are provided in **Appendix B**.

4 Experiments

4.1 Implementation Details

Baselines and Metrics. We compare S2S with seven existing two-person motion generation methods: RIG Tanaka & Fujiwara (2023), ComMDM Shafir et al. (2024), in2IN Ruiz-Ponce et al. (2024), InterGen Liang et al. (2024), InterMask Javed et al. (2025), TIMotion Wang et al. (2025), and Text2Interact Wu et al. (2026). We additionally evaluate *Planning + HYM1*, which combines our LLM-based social structure planner with the pretrained HY-Motion 1.0 model Team (2025) without fine-tuning. We evaluate text-motion alignment using top-1 retrieval precision (R-Prec@1), motion distribution fidelity using FID, and physical interaction quality using Interpenetration, Contact, and Foot-slide, which assess inter-person body penetration, inter-person contact, and foot sliding, respectively. Furthermore, a user study is provided in **Appendix D**.

Datasets. We use InterHuman Liang et al. (2024) and InterX Xu et al. (2024a), two widely adopted SMPL-based HHI datasets. We reorganize them with our motion-aligned social structure planning. The resulting fine-grained HHI training set contains 27,484 phase segments, each paired with a phase label, partner-aware role descriptions, and the corresponding motion sequence. Evaluation is conducted on 914 sequences that consist of 454 InterHuman cases and 460 InterX cases.

Setup. We initialize the motion executor with the open-source HY-Motion-1.0-Lite weights and finetune it with LoRA ($r = 16$ and $\alpha = 32$), which introduces only approximately 8M trainable parameters (1.7% of the backbone parameter size). We train S2S for 100 epochs with a batch size of 64, with AdamW (learning rate 1×10^{-4} and weight decay 1×10^{-3}) on an H100 GPU cluster. We use 50 flow-matching steps and a maximum of 300 frames per phase (10 seconds at 30 fps). The loss contribution ratio is set to 7:1:1:1 for $\mathcal{L}_{\text{FM}}$, $\mathcal{L}_{\text{smooth}}$, $\mathcal{L}_{\text{rel_dist}}$, and $\mathcal{L}_{\text{rel_ori}}$, respectively.



Figure 5: Qualitative evaluation results. We compare InterGen, ComMDM, a baseline using only our structure planning with the raw HY-Motion 1.0 model (Planning + HYM1), and Ours (Our structure planning + Our S2S framework). Social structure planning results used to condition HYM1 and our S2S are provided in **Appendix E**.

4.2 Main Results

Quantitative Evaluation. Table 1 shows that S2S achieves the best mean performance across all five metrics. It obtains an R-Prec@1 of 24.67%, surpassing Text2Interact by 6.95 percentage points, and the lowest FID of 0.661, compared with 0.704 for TIMotion. Lower Interpenetration and Foot-slide, together with higher Contact, further indicate improved physical interaction quality. Compared with Planning + HYM1, S2S improves R-Prec@1 by 4.42 percentage points and reduces FID from 0.703 to 0.661, with gains across all three physical metrics. These results support the value of adapting the solo backbone to ground social structure beyond planning alone.

Qualitative Evaluation. Fig. 5 illustrates improvements in both *phase progression* and *partner-aware coordination*. S2S produces clearer contact-to-separation progression in greeting (f) and preparation-to-lifting progression in assisted lifting (g), while better aligning the actors' complementary gestures in pushing and blocking (a) and paper passing (c). In contrast, Planning + HYM1 exhibits mismatched hand positions in (f) and fails to realize the assisted lift in (g). These examples suggest the adapted executor translates phase-level plans into coordinated two-person motion, grounding complementary actor roles in the geometry of physical interaction.

Table 2: Ablation of social planning and motion conditioning (mean ± standard deviation over five seeds). The full model leads all five metrics.

Ablation	R-Prec@1 (%) ↑	FID ↓	Interpenetration ↓	Contact ↑	Foot-slide ↓
w/o social planning	15.25 ± 0.87	0.783 ± 0.003	0.208 ± 0.017	0.376 ± 0.038	0.289 ± 0.016
w/o motion facts	23.61 ± 0.31	0.742 ± 0.011	0.120 ± 0.004	0.365 ± 0.016	0.294 ± 0.002
w/o self condition	20.65 ± 0.92	0.704 ± 0.006	0.121 ± 0.007	0.349 ± 0.009	0.242 ± 0.006
w/o partner condition	21.03 ± 0.45	0.691 ± 0.003	0.141 ± 0.001	0.352 ± 0.014	0.235 ± 0.002
Ours (full)	$\mathbf{24.67 \pm 0.42}$	$\mathbf{0.661 \pm 0.003}$	$\mathbf{0.103 \pm 0.004}$	$\mathbf{0.397 \pm 0.009}$	$\mathbf{0.218 \pm 0.004}$

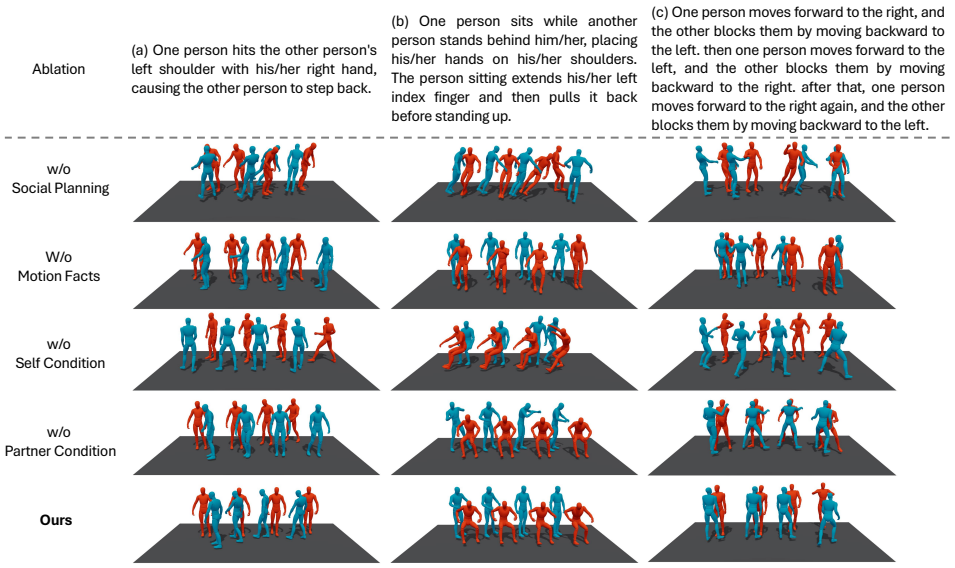

Figure 6: Qualitative ablation results.

4.3 Ablation Study

Table 2 evaluates each component's contribution. Removing social structure planning causes the largest R-Prec@1 drop (24.67% to 15.25%) and approximately doubles Interpenetration (0.103 to 0.208), highlighting its importance for semantic alignment and interaction geometry. Removing motion facts increases FID from 0.661 to 0.742 and yields the highest Foot-slide (0.294), suggesting that motion-grounded planning benefits motion quality beyond semantic alignment. Without self conditioning, R-Prec@1 falls to 20.65% and Contact drops to 0.349, indicating that motion history supports both semantic execution and contact realization. Removing partner conditioning raises Interpenetration to 0.141 and lowers Contact to 0.352, supporting its role in coordinating inter-person geometry. The full model leads all five metrics on average, supporting the complementary roles of motion-aligned planning, self conditioning, and partner conditioning.

5 Conclusion

We formulate text-driven 3D human-human interaction generation as social structure modeling and grounding, with phase progression and partner-aware coordination as its organizing principles. Our diagnostic study shows that LLMs can recover high-level social structure but struggle to execute continuous, physically plausible interaction motion. We propose **Think with LLM, Move with Motion Skill**: an LLM planner reconstructs motion-aligned social supervision from existing HHI data, while a Solo-to-Social executor adapts a pretrained solo motion model to ground the planned structure into coordinated two-person motion. Experiments on standard HHI benchmarks show improved text-motion alignment and physical interaction quality, with qualitative results illustrating clearer phase progression and partner-aware coordination. These findings support both explicit social structure planning and motion-level grounding, encouraging future research on generative systems that model social organization for human interaction and embodied agents.

AI USE STATEMENT

Large language models are part of the proposed method: they generate phase-wise, actor-specific social structure descriptions for motion-aligned annotation and text-only inference. We also evaluate Qwen3.5 in the planning and motion-execution diagnostic study described in Section 3.1. OpenAI ChatGPT is used for writing polishing and grammar checking.

ETHICS STATEMENT

This work studies two-person motion generation using the existing InterHuman and InterX datasets. The user study asks participants to assess generated animations and phase descriptions; its protocol is described in Appendix D. Potential benefits include animation and interactive virtual agents, while potential risks include misleading synthetic depictions of human interactions and the reproduction of biases in the source motion data. The phase and role representation simplifies human social behavior and should not be interpreted as a comprehensive account of human intent. Further discussion of applications and limitations appears in Appendix I.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the datasets, evaluation split, backbone, optimization settings, and generation settings. Appendix B gives the model formulation and training objectives; Appendix F details the motion-based annotation heuristics; and Appendix E provides the inference prompt and example plans. Metric definitions appear in Appendix A. The main comparison and ablation tables report means and standard deviations over five seeds. Planner diagnostics and the human evaluation protocol are provided in Appendices H and D, respectively.

REFERENCES

Avinash Amballa, Gayathri Akkinapalli, and Vinitra Muralikrishnan. Ls-gan: Human motion synthesis with latent-space gans. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 326–335, 2025.

Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15039–15048, 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. In *ICLR*, 2025.

Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6200–6211, 2025a.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

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

Siyuan Fan, Wenke Huang, Xiantao Cai, and Bo Du. 3d human interaction generation: A survey. *arXiv preprint arXiv:2503.13120*, 2025.

Zichen Geng, Zeeshan Hayder, Bo Miao, Jian Liu, Wei Liu, and Ajmal Mian. Disentangled hierarchical vae for 3d human-human interaction generation. *arXiv preprint arXiv:2603.00144*, 2026.

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 (ECCV)*, 2024.

Aman Goel, Qianhui Men, and Edmond S. L. Ho. Interaction Mix and Match: Synthesizing Close Interaction using Conditional Hierarchical GAN with Multi-Hot Class Embedding. *Computer Graphics Forum*, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14647.

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

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZAyuwJYN8N.

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, 2024.

Aliasghar Khani, Arianna Rampini, Bruno Roy, Larasika Nadela, Noa Kaplan, Evan Atherton, Derek Cheung, and Jacky Bibliowicz. Motion generation: A survey of generative approaches and benchmarks. *arXiv preprint arXiv:2507.05419*, 2025.

Ronghui Li, Youliang Zhang, Yachao Zhang, Yuxiang Zhang, Mingyang Su, Jie Guo, Ziwei Liu, Yebin Liu, and Xiu Li. Interdance: Reactive 3d dance generation with realistic duet interactions. *arXiv preprint arXiv:2412.16982*, 2024.

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*, pp. 1–21, 2024.

Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. Large model empowered embodied ai: A survey on decision-making and embodied learning. *arXiv preprint arXiv:2508.10399*, 2025.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. Long-term interactions with social robots: Trends, insights, and recommendations. *ACM Transactions on Human-Robot Interaction*, 14(3):1–42, 2025.

Elena Merlo, Marta Lagomarsino, and Arash Ajoudani. A human-in-the-loop approach to robot action replanning through llm common-sense reasoning. *IEEE Robotics and Automation Letters*, 2025.

Jeongeun Park, Sungjoon Choi, and Sangdoo Yun. A unified framework for motion reasoning and generation in human interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10698–10707, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9488–9497, 2023.

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

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. Mixermdm: Learnable composition of human motion diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12380–12390, 2025a.

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. *arXiv preprint arXiv:2512.19692*, 2025b.

Ali Rida Sahili, Najett Neji, and Hedi Tabia. Text-driven motion generation: Overview, challenges and directions. *arXiv preprint arXiv:2505.09379*, 2025.

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059, 2022.

Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In *The Twelfth International Conference on Learning Representations*, 2024.

Alexandros Stergiou and Ronald Poppe. Understanding human-human interactions: a survey. *arXiv preprint arXiv:1808.00022*, 2, 2018.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *International Journal of Computer Vision*, 134(3):113, 2026.

Wenhui Tan, Boyuan Li, Chuhao Jin, Wenbing Huang, Xiting Wang, and Ruihua Song. Think then react: Towards unconstrained action-to-reaction motion generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UxzKcIZedp.

Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15953–15963. IEEE, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Tencent Hunyuan 3D Digital Human Team. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

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.

Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024.

Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12281–12288, 2020.

Di Wu, Xian Wei, Guang Chen, Hao Shen, Xiangfeng Wang, Wenhao Li, and Bo Jin. Generative multi-agent collaboration in embodied ai: A systematic review. *arXiv preprint arXiv:2502.11518*, 2025a.

Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2interact: High-fidelity and diverse two-person interaction generation from text. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

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*, 2025b.

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

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

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.

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*, pp. 509–519, 2023.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2023.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. Llm-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753*, 2025.

A Metric Details

We describe the five metrics reported in Tables 1 and 2. R-Prec@1 and FID assess semantic alignment and motion distribution fidelity, respectively; Interpenetration, Contact, and Foot-slide assess complementary aspects of physical interaction quality.

- **R-Prec@1 (↑).** We evaluate top-1 retrieval precision at the phase level in a shared text-motion embedding space. For each generated motion phase, its paired phase description is ranked by embedding similarity within a pool of 16 candidate descriptions: the paired description and 15 mismatched descriptions. R-Prec@1 is the percentage of phase queries for which the paired description ranks first. Higher values indicate stronger phase-level semantic alignment.
- **FID (↓).** Fréchet Inception Distance measures the distributional discrepancy between generated and reference motions in a learned motion-feature space. Gaussian distributions are fitted to the two sets of motion embeddings, and their Fréchet distance is computed from the feature means and covariances. Lower values indicate that generated motions more closely match the reference motion distribution.

- **Interpenetration (↓).** Each of the 22 body joints carries a sphere with a body-part-dependent radius, ranging from 0.13 m at the pelvis and 0.10 m at the hips and spine to 0.05 m at the hands and feet. A frame is classified as interpenetrating if any of the 22×22 cross-person joint-pair distances falls below the sum of the corresponding radii. The metric is the fraction of such frames over all frames, providing a conservative, mesh-free proxy for body collision. Lower values indicate less interpenetration.
- **Contact (↑).** A frame is classified as a contact frame if at least one cross-person joint-pair distance is below 0.3 m and the frame contains no joint-sphere interpenetration as defined above. Contact is the fraction of such frames over all frames. Higher values indicate more inter-person contact without penetration under this geometric approximation.
- **Foot-slide (↓).** After shifting each sequence vertically so that its floor is at zero height, a foot joint is considered grounded when its height is below 0.05 m. We use both ankles and toes, compute their horizontal (XZ) speed from consecutive-frame displacements at 30 fps, and average the speed over grounded observations for each actor and then over both actors. The metric is reported in m/s. Lower values indicate less foot skating, since a planted foot should have zero horizontal velocity.

B Modeling Derivation of Social-Structure-Centered HHI Generation

B.1 Planner-Executor Factorization

Given a global interaction prompt y, text-driven HHI generation aims to model the conditional distribution of a two-person motion sequence

$$p(\mathbf{X}^{1:T} \mid y), \qquad \mathbf{X}^{1:T} = \{\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)}\}_{t=1}^{T}. \tag{10}$$

We use the SMPL-based motion representation Pavlakos et al. (2019), where the motion state of actor i at frame t is

$$\mathbf{x}_t^{(i)} = \big(\mathbf{r}_t^{(i)}, \mathbf{o}_t^{(i)}, \boldsymbol{\theta}_t^{(i)}, \boldsymbol{\eta}_t^{(i)}\big), \tag{11}$$

with $\mathbf{r}_t^{(i)} \in \mathbb{R}^3$ the root translation, $\mathbf{o}_t^{(i)} \in \mathbb{R}^6$ the root orientation, $\boldsymbol{\theta}_t^{(i)} \in \mathbb{R}^{21\times 6}$ the body joint rotations, and $\boldsymbol{\eta}_t^{(i)} \in \mathbb{R}^{22\times 3}$ the root-relative joint positions obtained by forward kinematics.

We introduce a latent social structure variable S as an intermediate representation that organizes the interaction before motion execution. Marginalizing over S gives

$$p(\mathbf{X}^{1:T} \mid y) = \sum_S p(\mathbf{X}^{1:T}, S \mid y) = \sum_S p(\mathbf{X}^{1:T} \mid S, y)\, p(S \mid y). \tag{12}$$

The term $p(S \mid y)$ corresponds to social structure planning, while $p(\mathbf{X}^{1:T} \mid S, y)$ corresponds to motion execution conditioned on the planned structure. We adopt the modeling approximation that, once S captures the phase progression and actor-specific role semantics, the global prompt provides largely redundant information for execution:

$$p(\mathbf{X}^{1:T} \mid S, y) \approx p_\theta(\mathbf{X}^{1:T} \mid S), \qquad p(S \mid y) \approx p_\phi(S \mid y). \tag{13}$$

This yields the planner-executor formulation

$$p(\mathbf{X}^{1:T} \mid y) \approx \sum_S p_\theta(\mathbf{X}^{1:T} \mid S)\, p_\phi(S \mid y), \tag{14}$$

where p_ϕ is implemented by the LLM planner and p_θ is implemented by the motion skill executor.

B.2 Social Structure Derivation

Since existing HHI datasets provide paired global text and two-person motion, we can derive S from paired motion data by constraining the LLM planner with phase-wise motion facts.

Given a paired HHI motion sequence $\mathbf{X}^{1:T}$, we first detect the phase sequence

$$\mathcal{C} = f_{\mathrm{p}}(\mathbf{X}^{1:T}), \qquad \mathcal{C} = \{c_k\}_{k=1}^{K}, \qquad c_k \in \mathcal{V}_{\text{phase}}, \tag{15}$$

where

$$\mathcal{V}_{\text{phase}} = \{\texttt{approach}, \texttt{contact}, \texttt{release}, \texttt{in-place}\}. \tag{16}$$

The mapping $f_{\text{p}}(\cdot)$ uses available InterX segment annotations and otherwise applies a forward state machine to motion-derived signals, as detailed in Appendix F.

For each phase c_k, we extract motion facts from the corresponding motion segment:

$$\mathcal{M}_k = \{m_k^{(1)}, m_k^{(2)}\}, \qquad m_k^{(i)} = \big(\delta_k,\, u_k,\, q_k,\, \mathbf{a}_k^{(i)},\, \mathbf{l}_k^{(i)}\big), \qquad i \in \{1, 2\}. \tag{17}$$

Here, δ_k denotes inter-person distance evolution, u_k denotes the motion initiator, q_k captures contact semantics, $\mathbf{a}_k^{(i)}$ summarizes the motion direction and facing state of actor i, and $\mathbf{l}_k^{(i)}$ describes limb-level cues such as reaching, arm lifting, bending, and foot activity.

Conditioned on the global interaction text y, phase label c_k, and motion facts $\mathcal{M}_k$, the LLM planner produces actor-specific semantic descriptions:

$$\mathcal{Y}_k \sim p_\phi(\mathcal{Y}_k \mid y,\, c_k,\, \mathcal{M}_k)\,, \qquad \mathcal{Y}_k = (y_k^{(1)}, y_k^{(2)}). \tag{18}$$

Finally, aggregating all phases yields the social structure used by the executor:

$$S = \{(c_k, y_k^{(1)}, y_k^{(2)})\}_{k=1}^{K}. \tag{19}$$

Although the final S contains only phase labels and actor-specific semantic descriptions, it is derived under motion fact constraints and is therefore aligned with the observed motion sequence.

B.3 Phase-wise Motion Execution Modeling

Given the planned social structure S, the executor generates motion phase by phase:

$$p_\theta(\mathbf{X}^{1:T} \mid S) \approx \prod_{k=1}^{K} \prod_{i=1}^{2} p_\theta\left(\mathbf{x}_k^{(i)} \mid c_k,\, y_k^{(i)},\, \mathbf{h}_k^{(i)},\, \mathbf{p}_k^{(i)}\right), \tag{20}$$

where $\mathbf{x}_k^{(i)}$ is actor i's phase motion, $\mathbf{h}_k^{(i)}$ is its self-history prefix, and $\mathbf{p}_k^{(i)}$ is the partner condition available before generating phase k. The two actors are coupled through states propagated between phases. Phase progression is grounded by self conditioning, where the anchor region of phase k is taken from the last e frames of the previous phase:

$$\mathbf{x}_k^{(i)}[1:e] = \mathbf{h}_k^{(i)} = \mathbf{x}_{k-1}^{(i)}[T_{k-1} - e + 1 : T_{k-1}], \tag{21}$$

where T_{k-1} denotes the duration of phase $k-1$. For $k > 1$, partner conditioning uses the latest generated state $\mathbf{s}_{k-1}^{(j)}$ after the preceding phase, transformed into the current actor's ego-centric frame:

$$\mathbf{p}_k^{(i)} = R_{j \to i}\left(\mathbf{s}_{k-1}^{(j)}\right), \qquad j \neq i. \tag{22}$$

B.4 Solo-to-Social Optimization

We employ HY-Motion 1.0 Team (2025) as the backbone of the motion executor and adapt it with LoRA. For selected linear layers with pretrained weight $\mathbf{W}_0$, the adapted weight is

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \tag{23}$$

where $\mathbf{A}$ and $\mathbf{B}$ are trainable low-rank matrices, r is the rank, and α is the scaling factor.

The overall fine-tuning objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + w_{\text{smooth}}\mathcal{L}_{\text{smooth}} + w_{\text{dist}}\mathcal{L}_{\text{rel_dist}} + w_{\text{ori}}\mathcal{L}_{\text{rel_ori}}. \tag{24}$$

Let $\mathbf{x}_{\text{gt}}$ denote the clean target motion, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ the Gaussian noise, and $s \sim \mathcal{U}(0, 1)$ the flow interpolation time. The noised motion is

$$\mathbf{x}_s = (1 - s)\boldsymbol{\epsilon} + s\,\mathbf{x}_{\text{gt}}, \tag{25}$$

and the flow-matching loss is

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \| v_\theta(\mathbf{x}_s, s)_t - (\mathbf{x}_{\mathrm{gt}} - \boldsymbol{\epsilon})_t \|_2^2 , \tag{26}$$

where $\mathcal{V}$ denotes the valid generated frames excluding anchor-prefix frames.

The smoothness loss is

$$\mathcal{L}_{\mathrm{smooth}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \left\| \hat{\mathbf{x}}^t - \hat{\mathbf{x}}^{t-1} \right\|_2^2 . \tag{27}$$

During training, the partner is treated as a fixed geometric reference using its ground-truth motion. The relative distance loss is

$$\mathcal{L}_{\mathrm{rel_dist}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \left(\left\| \hat{\mathbf{r}}^t - \mathbf{r}_p^t \right\|_2 - \left\| \mathbf{r}^t - \mathbf{r}_p^t \right\|_2 \right)^2 . \tag{28}$$

The relative orientation loss is

$$\mathcal{L}_{\mathrm{rel_ori}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \left(\hat{\mathbf{f}}^t \cdot \mathbf{d}(\hat{\mathbf{r}}^t, \mathbf{r}_p^t) - \mathbf{f}^t \cdot \mathbf{d}(\mathbf{r}^t, \mathbf{r}_p^t) \right)^2 , \tag{29}$$

where

$$\mathbf{d}(a, b) = \frac{b - a}{\|b - a\|_2 + \varepsilon} \tag{30}$$

is the normalized direction vector from a to b. Together, these objectives adapt the solo motion backbone into a socially aware executor by preserving atomic motion quality, improving temporal smoothness, and enforcing partner-relative coordination.

C LLM-BASED HUMAN ATOMIC MOTION EXECUTION

To examine whether LLMs can directly serve as motion executors, we ask the LLM to control SMPL-based human motions from a set of atomic action prompts, such as stepping forward, raising an arm, reaching out, bowing, and waving. These prompts describe simple single-person motions and therefore provide a basic test of motion execution ability before considering more complex two-person interactions.

As shown in Fig. 7, the LLM can sometimes recover coarse pose-level semantics, such as lifting the specified arm or orienting the body toward the intended direction. However, the generated results are often close to static poses, with limited motion amplitude, weak temporal evolution, and insufficient coordination among body joints. For locomotion-related prompts, the model frequently fails to produce dynamic displacement or realistic stepping motion. These observations indicate that LLMs may understand the semantic meaning of atomic motion descriptions, but they do not reliably generate continuous, dynamic, and physically plausible 3D motion. This further supports our planner-executor design: LLMs are suitable for semantic-level planning, while motion execution should be handled by specialized motion models.

D USER STUDY

We conducted a user study with a Gradio-based interface, as illustrated in Fig. 10. Each participant evaluated 10 randomly selected cases, resulting in 120 case evaluations from 12 participants in total. All participants had research backgrounds in computer vision or closely related areas. For each case, participants viewed animations generated by seven models and answered the questions listed in Table 3. Participants first rated the phase decomposition accuracy of the phase-level prompts on a 5-point Likert scale, where 1 denotes "not accurate at all" and 5 denotes "very accurate." They then evaluated the generated animations along three motion-related aspects: global text-motion alignment, partner coordination, and phase-level alignment. For these three aspects, participants ranked the seven methods from 1 (best) to 7 (worst). To reduce ordering bias, the presentation order of the seven animations was randomly shuffled for each case.

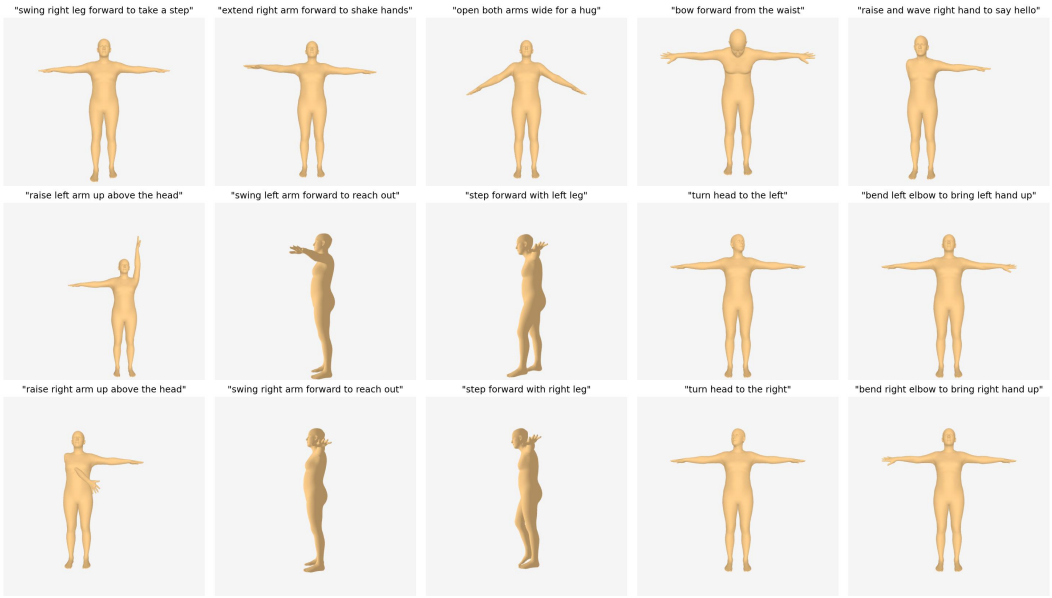

Figure 7: Visualizations of LLM-based human atomic motion execution. Given atomic action prompts, the LLM can sometimes recover coarse pose-level semantics, but the generated motions are often static, low-amplitude, and lack continuous dynamics. This supports our observation that LLMs can reason about social structure planning but are insufficient for direct motion execution.

Table 3: User study questionnaire. We evaluate generated HHI motions along four dimensions: phase decomposition accuracy, global text-motion alignment, partner coordination, and phase-level alignment. Participants judge whether the phase-level prompts faithfully reflect the global interaction description and whether the generated motions correctly realize the intended interaction, temporal phase progression, and partner-aware coordination.

Dimension	**Question**
(a) Phase Decomposition Accuracy	How accurately do the phase-level text prompts decompose the global text description?
(b) Global Text-Motion Alignment	How well does the generated full motion sequence align with the global text description? Consider whether the interaction, movement direction, action intention, and final outcome correspond to the global prompt.
(c) Partner Coordination	How well do the two partners coordinate during the interaction? Consider whether they respond appropriately, maintain reasonable spatial relationships, and perform coordinated movements.
(d) Phase-Level Alignment	How well does each generated motion phase align with its corresponding phase-level text description? Consider the phase type, i.e., approach, contact, release, or in-place, and whether the intended phase-specific actions occur at the appropriate time.

The results are summarized in Fig. 8. Fig. 8(a) shows that most responses for phase decomposition accuracy fall into the “Accurate” and “Very accurate” categories, with a mean score of 4.03, indicating that the LLM-generated phase prompts are generally faithful to the original global descriptions. Fig. 8(b) reports the ranking distribution for global text-motion alignment, where our method achieves the most favorable overall rankings, suggesting stronger faithfulness to the intended interaction semantics. Fig. 8(c) shows that our method obtains the best average ranking for partner coordination, indicating stronger mutual responsiveness and more coherent inter-person spatial relationships. Fig. 8(d) further shows that our method better follows the decomposed phase-

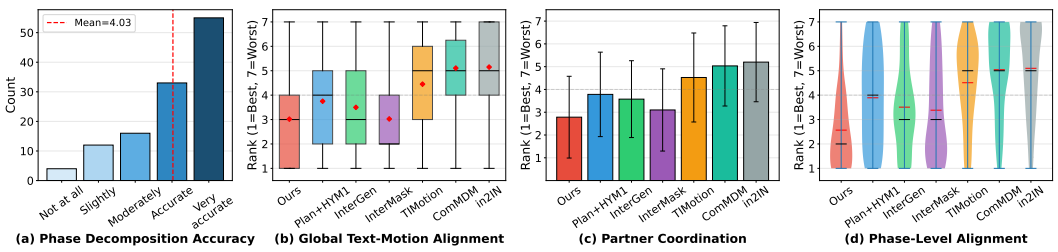

Figure 8: User study results across four evaluation aspects. (a) Phase decomposition accuracy of the LLM-generated phase-level prompts. (b) Global text-motion alignment between the generated full motion sequence and the global prompt. (c) Partner coordination between the two actors. (d) Phase-level alignment between generated motions and decomposed phase prompts.

level prompts, preserving clearer temporal structure across interaction phases. Overall, the user study supports our quantitative findings and demonstrates improvements in global text-motion faithfulness, partner-aware coordination, and phase-level temporal consistency.

E Social Structure Planning Results

We present the inference-time prompt for LLM-based social structure planning in Fig. 9. The prompt asks the LLM to convert a global two-person interaction description into a phase-wise plan, where each phase is assigned one of four interaction types, `approach`, `contact`, `release`, or `in-place`, together with partner-aware action descriptions for P1 and P2. These four types define the phase vocabulary, not a limit on plan length: phase types may recur as the interaction unfolds, allowing longer phase sequences such as those examined in Appendix G. This format makes the implicit temporal progression and role coordination in the global prompt explicit, providing structured conditions for interaction motion execution.

Table 5 provides the social structure planning results for the eight qualitative cases shown in Fig. 5. For each global interaction prompt, the LLM planner decomposes the interaction into one or more temporally ordered phases and assigns partner-aware action descriptions to the two actors within each phase. These phase-level plans make the implicit interaction organization explicit, specifying both how the interaction progresses over time and how the two actors should coordinate their roles. They are used as structured conditions for interaction motion execution.

F Phase-Detection Heuristics for Motion-Aligned Annotation

This section details the motion-based annotation procedure introduced in Section 3.2. Given a paired motion sequence, the procedure identifies phase segments and extracts motion facts that guide the LLM's actor-specific descriptions. It uses the vocabulary of four phase types $\mathcal{V}_{\text{phase}} = \{\texttt{approach}, \texttt{contact}, \texttt{release}, \texttt{in-place}\}$. These are reusable labels, not a four-phase limit on the social plan. The procedure requires observed motion and is used to construct annotations; at inference time, the planner instead predicts a social plan from the global text prompt.

Official InterX segment annotations take priority when available. A forward state machine over motion-derived signals provides segments for InterHuman and for InterX sequences without such annotations, followed by human verification. Wherever the detector is used, its thresholds are fixed across both datasets.

- **Contact cues for annotation.** A contact candidate is detected when an inter-person joint-pair distance falls below 0.15 m. The detector requires at least 5 consecutive frames and uses an 8-frame wind-up guard to suppress spurious triggers. Contact gaps shorter than 30 frames are bridged to avoid fragmenting sustained interactions such as a hug or a dance hold. This temporally filtered cue locates annotation boundaries; it differs from the evaluation metric *Contact* in Appendix A, which uses a 0.3 m threshold and excludes interpenetrating frames.

- **Phase-type assignment.** The detector assigns `approach` to convergence before contact, refining its late boundary using a wrist-to-partner-upper-body distance below 0.8 m. Detected contact spans receive `contact`; subsequent withdrawal receives `release`, with its boundary refined using a wrist distance exceeding 0.6 m. For non-contact segments, mean wrist speed above 0.3 m/s provides a cue for `in-place` interaction, such as waving or gesturing; otherwise, the fallback label is `approach`. These rules operationalize the phase vocabulary for annotation rather than define all possible realizations of each semantic phase type.
- **Motion-consistency correction.** A post-processing pass checks phase labels against the motion cues. Segments containing detected contact are assigned `contact`, while a `contact` label without supporting contact cues is corrected to `approach`. Post-contact withdrawal is identified by increasing inter-person distance and assigned `release`. These checks reduce mismatches between the phase labels and the observed interaction geometry.
- **Segment consolidation.** Segments shorter than 15 frames are merged into a neighboring segment, and durations are mapped to the nearest duration bucket supported by the motion executor. Phase lengths may vary, subject to a maximum of 300 frames per phase (10 seconds at 30 fps), consistent with Section 4.1. This step prepares the detected segments for motion-aligned supervision.
- **Initiator-cue extraction.** For contact phases, the actor moving more toward the partner during the 10 frames preceding contact is marked as the initiator. For non-contact phases, the cue is based on the larger horizontal (XZ) displacement. Near ties are resolved by alternating the assignment across phases; this is a tie-breaking convention rather than evidence of a distinct semantic role. The resulting initiator cue is one component of the motion facts in Section 3.2. The LLM combines these facts with the global text and phase label to infer complementary actor-specific action descriptions.

G Error Analysis

To evaluate error accumulation beyond the 1–4-phase benchmark setting, we construct synthetic 14-phase chains by concatenating randomly sampled plans from 150 inference-time planner outputs. Phase descriptions remain fixed to assess error propagation through the executor's self-conditioning chain without replanning. Between the first and final phases of the 14-phase rollout, R-Prec@1 decreases by 6.9 percentage points and FID increases by 0.10 relative to initial values of 24.7% and 0.64, respectively, indicating long-horizon drift in semantic alignment and motion quality.

As a reference-assisted diagnostic, we re-anchor the propagated state to a ground-truth motion frame every eight phases. At the eighth phase, calibration yields a 4.8-percentage-point improvement in R-Prec@1 and a 0.07 reduction in FID relative to the uncalibrated rollout. At the final phase, the corresponding advantages are 1.8 percentage points and 0.03, respectively, with calibrated R-Prec@1 of 19.6% and FID of 0.71. These results characterize long-horizon drift and show that reference-assisted calibration mitigates, but does not eliminate, its accumulation.

H Inference-Time Planning: Reference Agreement and Downstream Impact

During annotation, observed motion provides phase boundaries and motion facts that constrain the social plan. At inference time, only the global text prompt is available to the planner (Section 3.2). We examine this information gap on a subset of 150 validation sequences through two questions: how closely do text-only plans agree with motion-aligned reference annotations, and how does the plan source affect motion generated by the same executor? This subset analysis complements the 914-sequence main evaluation in Section 4.1.

Reference-Count-Controlled Diagnostic. To control for phase-count mismatch, we separately provide the reference number of phases to the planner. This diagnostic uses additional reference information and is distinct from unconstrained text-only inference. With the reference phase count provided, phase-type accuracy is 59% and role accuracy is 63%. Equal phase counts facilitate comparison but do not guarantee matching phase boundaries or semantics. In this setting, predicted per-actor descriptions have higher TMR cosine similarity to matched reference annotations than to shuffled annotations (0.149 vs. 0.114; paired $p = .002$; $n = 466$ phase pairs). This comparison

Table 4: Motion quality when the Executor is driven by oracle plans (built from ground-truth motion facts) versus inference-time plans (text-only Planner, no motion facts) on the same 150 validation sequences.

Plan source	R-Prec@1 (%) ↑	FID ↓	Interpenetration ↓	Contact ↑	Foot-slide ↓
Oracle (GT plans)	25.47	0.611	0.101	0.193	0.204
Planner (end-to-end)	24.61	0.667	0.119	0.204	0.215

supports semantic correspondence between the descriptions beyond agreement on discrete phase labels.

Downstream Motion Quality. We compare text-only plans with motion-informed reference plans constructed using ground-truth motion facts, passing both through the same executor on the same 150 sequences. We call the latter *oracle plans* to indicate their access to reference motion, rather than a guarantee of optimal generation. Table 4 reports the resulting motion quality. Relative to oracle plans, text-only plans yield R-Prec@1 that is 0.86 percentage points lower and FID that is 0.056 higher. The corresponding increases in Interpenetration and Foot-slide are 0.018 and 0.011, respectively. These differences characterize the downstream gap associated with the planning information available at inference time.

Contact is 0.011 higher with text-only plans, indicating more penetration-free contact frames. This increase can coexist with more interpenetrating frames elsewhere in the sequences and does not establish more accurate contact timing. Taken together, the results show that planning conditions affect the quality of motion-level grounding, while leaving room to improve text-only planning. The aggregate metrics do not isolate the effects of phase-count, phase-type, and role-description errors.

I Discussion

Relation to Action-to-Reaction Generation. Our executor shares a conditional-generation mechanism with reaction models: each actor uses the partner's motion context to guide its own motion. The distinction lies in the overall task formulation. In the action-to-reaction setting considered here, one actor's motion is supplied as a condition for generating the other actor's response. Our framework instead starts from a global interaction description and generates both actors under a shared social structure plan. Phase progression and complementary actor roles organize the interaction, while partner conditioning grounds this organization in motion. Thus, reaction-like conditioning serves as a local execution mechanism within our planner–executor framework.

Future Work. This work formulates text-driven 3D human-human interaction generation as social structure modeling and grounding, but several directions remain open. First, our current framework focuses on two-person interactions. Extending social structure modeling to multi-person scenarios is an important future direction, where interaction organization may involve group roles, changing subgroups, and more complex coordination patterns. Second, our phase vocabulary captures common interaction stages such as approach, contact, release, and in-place coordination, but real human interactions can involve more subtle temporal and social dynamics. Future work may explore richer or adaptive social structure representations, uncertainty-aware planning, and more diverse interaction data that better cover different body types, cultural contexts, and social behaviors.

Societal Impact. This work may have positive impact by enabling more controllable and semantically grounded generation of two-person human interactions. It can benefit animation production, virtual avatars, embodied AI, social robotics, human-computer interaction, and simulation environments for training or education by reducing manual motion authoring effort and supporting richer virtual social scenarios. Potential negative impacts should also be considered. Generated human interactions may be misused to create misleading synthetic content, and models trained on existing motion datasets may inherit biases in body types, interaction styles, cultural norms, or social behaviors. Moreover, representing interaction through planned phases and roles may simplify the complexity of real human social behavior, especially in ambiguous, emotional, or culturally specific settings. The method does not involve private personal data or high-risk deployment, but responsible

LLM Prompt for Social Structure Planning

```
You are a professional social psychology expert specializing in analyzing the
     social behaviors of human-human interactions.

Your task is to convert a global two-person interaction description into a phase-
     wise social structure plan.
The plan should describe how the interaction unfolds over time and what each
     person does in each phase.

Definitions:
- Person 1 / P1 refers to the first person mentioned in the global prompt.
- Person 2 / P2 refers to the other person.
- Use only the following phase types:
  1. approach: one or both people move closer or prepare to interact before
     contact.
  2. contact: physical contact, object transfer, blocking, supporting, or direct
     interaction occurs.
  3. release: physical contact or interaction ends and one or both people withdraw
      or return to neutral.
  4. in-place: the interaction happens mostly without clear approach/contact/
     release progression,
     or the actors coordinate while staying near their positions.

Instructions:
1. Decompose the global prompt into as many temporally ordered phases as needed to
      capture its interaction progression.
2. Assign exactly one of the four phase types to each phase. Types may recur; they
      are not four mandatory steps or a fixed ordering.
3. For each phase, write one action description for P1 and one action description
     for P2.
4. The P1 and P2 actions must be partner-aware.
5. Preserve the semantics of the global prompt. Do not invent unrelated actions.
6. If the prompt implies asymmetric roles, make them explicit, such as initiator/
     receiver, attacker/defender, giver/taker, or supporter/assisted person.
7. Use in-place for coordination without a distinct approach, contact, or release
     stage.
8. Keep each action concise, physically plausible, and suitable for motion
     generation.

Output format:
Phase 1: (<phase_type>)
P1 action: <Person 1's action>
P2 action: <Person 2's action>
... (repeat the phase block as needed)
Phase K: (<phase_type>)
P1 action: <Person 1's action>
P2 action: <Person 2's action>
K is determined by the interaction, not by the number of phase types.

Global interaction prompt:
"<INPUT_GLOBAL_PROMPT>"
```

Figure 9: Inference-time prompt for LLM-based social structure planning. The four phase types form a reusable vocabulary; the number of phases depends on the interaction.

use and careful evaluation remain necessary when applying generated interaction motions in social or human-facing contexts.

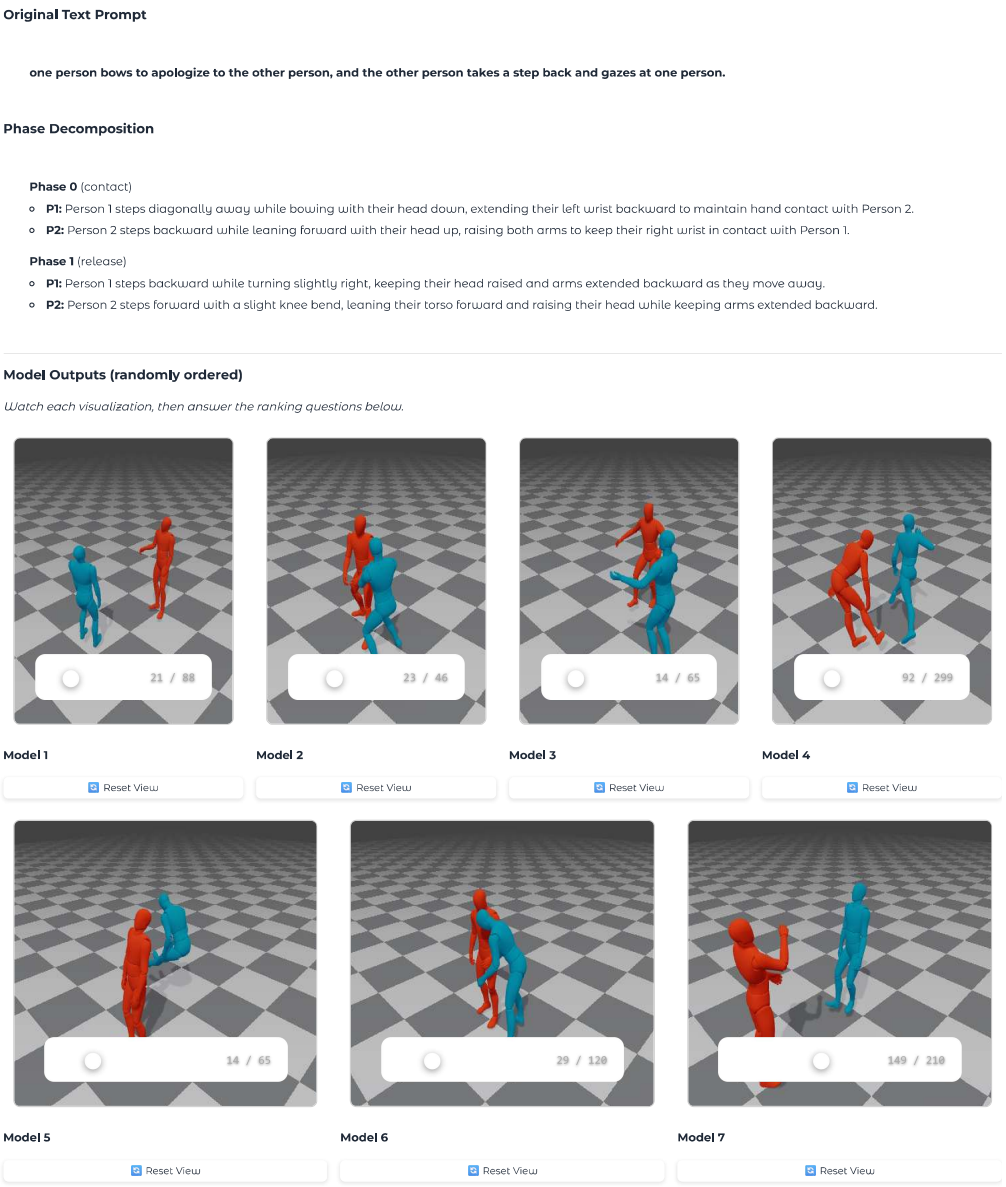

Figure 10: User study interface. Each case presents the original global text prompt, the LLM-decomposed phase-level prompts, and seven randomly ordered model outputs. Participants first evaluate the accuracy of the phase-level prompts, and then rank the generated motions according to global text-motion alignment, partner coordination, and phase-level alignment. The model order is randomized for each case to reduce ordering bias.

Table 5: Social structure prompts used for the qualitative examples in Fig. 5. The second column gives the global interaction prompt for each case. Planning + HYM1 and Ours use the phase-wise social structure text in the third column, while the other methods use the global prompt directly.

ID	Global Interaction Text	Social Structure Planning Result
a	The first one's right hand pushes the other person's left shoulder, and the other person blocks it with their hand.	**Phase 1: (contact)** **P1 action:** Person 1 pushes Person 2's left shoulder with their right hand while stepping back. **P2 action:** Person 2 blocks the push with their hand and steps back to create distance.
b	The two are photographing themselves together.	**Phase 1: (in-place)** **P1 action:** Person 1 holds or positions the camera/phone in front of both people and poses for the photo. **P2 action:** Person 2 stays close to Person 1 and poses together for the photo.
c	One retrieves a paper from one's pocket and passes it to the other.	**Phase 1: (approach)** **P1 action:** Person 1 walks forward and extends an arm to retrieve a paper from their pocket. **P2 action:** Person 2 stands still facing Person 1, waiting for the paper to be passed. **Phase 2: (contact)** **P1 action:** Person 1 passes the paper into Person 2's hand. **P2 action:** Person 2 grasps and receives the paper from Person 1.
d	The first person raises his/her hands, reaches forward, and pinches both cheeks of the second person.	**Phase 1: (approach)** **P1 action:** Person 1 raises both hands and extends both arms forward toward Person 2's face. **P2 action:** Person 2 remains mostly still, facing Person 1 and waiting for the interaction. **Phase 2: (contact)** **P1 action:** Person 1 places both hands on Person 2's cheeks and gently pinches both sides of the face. **P2 action:** Person 2 stays in place while receiving the cheek-pinching interaction, with only minor head or upper-body response. **Phase 3: (release)** **P1 action:** Person 1 releases Person 2's cheeks and slightly withdraws both hands from Person 2's face. **P2 action:** Person 2 remains facing Person 1 and returns to a neutral posture after the contact ends.
e	The first person advances and launches an attack with the right hand while the other person retreats.	**Phase 1: (approach)** **P1 action:** Person 1 advances toward Person 2 with a raised right hand, preparing to strike. **P2 action:** Person 2 retreats while facing Person 1, maintaining a defensive posture without contact.
f	Two people reach out their right hands to greet one another.	**Phase 1: (contact)** **P1 action:** Person 1 walks forward extending their right hand to initiate a greeting. **P2 action:** Person 2 moves slightly forward extending their right hand to meet Person 1.
g	The other person assists one person by lifting them up.	**Phase 1: (contact)** **P1 action:** Person 1 holds or supports Person 2 and lifts them upward. **P2 action:** Person 2 is physically supported by Person 1 and is lifted upward with assistance. **Phase 2: (release)** **P1 action:** Person 1 gradually reduces support and releases Person 2 after the lift. **P2 action:** Person 2 returns to a stable standing posture after being lifted.
h	One of the persons suddenly stands up while both were sitting.	**Phase 1: (in-place)** **P1 action:** Person 1 suddenly rises from a sitting position to a standing posture. **P2 action:** Person 2 remains seated and stays mostly still while Person 1 stands up.