# FROM DIGITAL HUMAN INTERACTIONS TO PHYSICS-BASED HUMANOID SKILLS: PHYSICS-GROUNDED POST-TRAINING OF INTERACTION GENERATORS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent methods have made promising progress in generating interactions between two humanoids, largely relying on physics-based tracking policies to convert digital reference motions into executable trajectories. However, limited tracking capabilities restrict the range of reference motions that can be successfully executed, reducing data utilization. Moreover, even successful tracking does not guarantee physically plausible responses or faithful realization of the intended interactions. In this paper, we introduce DIGHT, a co-adaptive framework that couples a **D**igital human **I**nteraction **G**enerator with a **H**umanoid **T**racking policy. Our DIGHT first executes multiple text-conditioned interaction candidates in simulation using a fixed tracker. It then constructs physics-grounded preferences from the resulting rollouts, covering both general executability and interaction fidelity. Rather than collapsing these signals into a single scalar reward for candidate ranking, we align the pretrained generator using physics-decoupled diffusion direct preference optimization (DPO), preserving criterion-specific supervision without differentiating through the simulator. To improve executability, preference pairs are derived from tracking error, friction, and floating. Additionally, to improve interaction fidelity, we propose to incorporate force feedback from the simulator as a measure of contact fidelity and construct preferences over contact occurrence, location, duration, and force magnitude. The aligned generator then supplies reference motions for fine-tuning the tracker, improving compatibility between generation and physical execution. Extensive experiments demonstrate that our approach not only improves the physical plausibility of generated motions but also enables more reliable and faithful humanoid interactions in simulation.

## 1 INTRODUCTION

Physics-based interaction between two humanoids is an important capability for embodied agents, enabling them to coordinate with and physically respond to one another in a shared environment. It supports a wide range of applications, including collaborative agents, social robotics, character animation, and embodied simulation Sui et al. (2025); Mascaro et al. (2024); Zhang et al. (2023). In particular, generating such interactions from textual descriptions provides an intuitive way to specify diverse behaviors Li et al. (2026); Javed et al. (2024); Wang et al. (2025a) while requiring the resulting motions to be both semantically meaningful and physically realizable.

Relative to single-humanoid control, dual-humanoid coordination introduces heightened complexities Zhang et al. (2026a). In addition to ensuring individual stability and dynamic feasibility, the agents must synchronize their spatial configurations, motion timing, and physical interactions Zhang et al. (2023); Liu et al. (2026); Lionar & Lee (2026). Moreover, their executions are inherently coupled: the motion and contact forces of one humanoid directly affect the state and response of the other Shibata et al. (2026). Consequently, even small deviations in position or timing can lead to missed contacts, implausible forces, or unstable executions, ultimately altering the intended interaction Huang et al. (2026); Ugrinovic et al. (2024). Therefore, successful two-humanoid interaction requires not only *general executability* for each participant, but also *interaction fidelity* that preserves the intended contacts and coordination between them.



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.

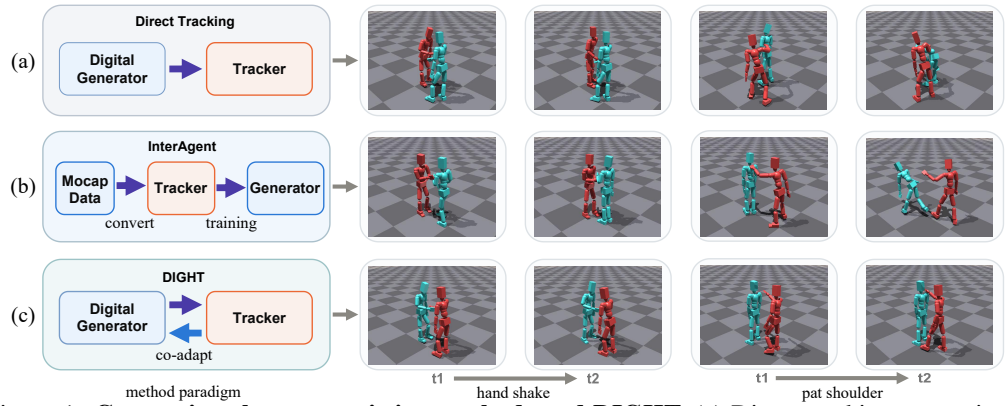



Figure 1: **Comparison between existing methods and DIGHT.** (a) Direct tracking may miss intended contacts or induce excessive forces despite successful pose tracking. (b) InterAgent produces physically implausible contact forces and unstable interactions due to the gap between the training and inference. (c) In contrast, DIGHT uses physical feedback from generated rollouts to co-adapt the interaction generator and tracking policy, preserving intended contacts while producing physically stable interactions.

Recently, many works have achieved humanoid interaction generation by leveraging humanoid tracking policies in two ways. Human-X Ji et al. (2025) first generates digital human motions and then executes them using a humanoid tracking policy for reaction generation. In contrast, PhysReaction Liu et al. (2024) and InterAgent Li et al. (2026) use tracking policies to convert MoCap data into executable humanoid trajectories, which serve as supervision for training interaction generators.

However, both paradigms inherit fundamental limitations from their reliance solely on tracking. As shown in Fig. 1(a), direct tracking primarily optimizes pose-level similarity with a generated reference rather than the physical fidelity of the interaction. Consequently, even a successfully tracked interaction may exhibit unrealistic contact behavior: the handshake may lose the intended hand-to-hand contact, while following the shoulder-patting reference may induce excessive contact forces that destabilize the receiving humanoid. On the other side, as shown in Fig. 1(b), although InterAgent trains its generator on trajectories obtained through humanoid tracking, it learns physical feasibility only indirectly by imitating these trajectories. This leaves a gap between training and generation, as the generated interactions are not subject to explicit physics-based constraints on their dynamics at inference time. Consequently, the generated patting shoulder interaction produces a physically implausible contact force that destabilizes the receiving humanoid.

In this paper, we introduce **DIGHT**, a physics-grounded co-adaptive framework that couples a **D**igital human **I**nteraction **G**enerator with a **H**umanoid **T**racking policy. The core idea behind DIGHT is to bridge digital generation and physical execution through physics-grounded co-adaptation: physical feedback from tracking rollouts informs the generator about the physical capabilities of target humanoids, while reference motions from the refined generator are used to adapt the tracking policy to the generator's motion distribution. By enabling each component to account for the other's capabilities, DIGHT narrows the gap between digital interaction generation and physical humanoid execution. Specifically, DIGHT samples multiple candidate interactions and executes them in simulation via the tracking policy. It then translates their relative performance in executability and interaction fidelity into criterion-specific preferences to optimize the generator. For improving executability, we use three general physical preferences including floating, friction, and tracking errors. To address missing and unrealistic contacts in humanoid interaction, we propose to incorporate force feedback from simulator to construct contact-related preferences that assess contact occurrence and interaction fidelity across contact locations, durations, and force magnitudes. To preserve distinct supervision signals from different physical preferences, we introduce physics-decoupled diffusion DPO rather than aggregating them into a single scalar reward. Through this optimization, the generator is progressively refined toward motions that are physically executable and faithful to intended contacts. Finally, the aligned generator provides reference motions for policy adaptation, establishing a mutually reinforcing co-adapted framework.

We evaluate DIGHT on two text-conditioned interaction datasets, InterHuman Liang et al. (2024) and Inter-X Xu et al. (2024), using two distinct diffusion backbones, InterGen Liang et al. (2024)

and TIMotion Wang et al. (2025a). Across all four dataset-backbone settings, DIGHT consistently reduces all five evaluated physical errors while largely preserving the motion quality and text alignment of the pretrained generators. Existing interaction datasets provide paired motions and textual descriptions, but lack explicit annotations of the contacts required by each interaction, making it difficult to directly evaluate whether a generated motion realizes the contacts implied by its text. To address this limitation, we establish a contact-aware benchmark based on the InterHuman test set, labeling whether an interaction requires physical contact and identifying the expected contacting body-region pairs. On this benchmark, DIGHT improves contact-topology precision from 38.29% to 46.77% and recall from 52.16% to 63.94% over the InterGen baseline. In addition, our method generalizes effectively to temporal compositional and multi-role interaction scenarios.

Our contributions can be summarized as follows: **1.** We introduce DIGHT, a physics-grounded co-adaptive framework that jointly adapts interaction generation and humanoid tracking through physical execution feedback. **2.** We propose physics-decoupled diffusion DPO to align the interaction generator with multiple physics-grounded preferences without collapsing all signals into a single scalar reward. **3.** We explicitly incorporate interaction forces to constructing contact-related preferences that evaluate contact occurrence and interaction fidelity. **4.** We develop a contact-annotated evaluation subset of InterHuman to directly measure the fidelity of intended physical contacts.

## 2 Related Work

**Text-conditioned Human-human Interaction Generation.** Text-conditioned human–human interaction generation aims to synthesize coordinated motions of two interacting individuals from textual descriptions. Early diffusion-based approaches such as InterGen Liang et al. (2024) jointly generate two motion sequences and model cross-participant dependencies through mutual attention. Subsequent methods improve interaction modeling through participant specific language conditions Ruiz-Ponce et al. (2024), collaborative masked prediction Javed et al. (2024), unified multi-person latent representations Li et al. (2025), temporal and role-evolving modeling Wang et al. (2025a), and efficient architectures Wu et al. (2025b). Other works support fine-grained control Wu et al. (2026); Wang et al. (2024), autoregressive interaction generation Ruiz-Ponce et al. (2026), or generation for arbitrary numbers of participants Fan et al. (2024); Yu et al. (2026). More recent approaches introduce physical priors into interaction generation. PhyInter Gao et al. (2025) incorporates physics-guided constraints into motion diffusion, while PhysiInter Yao et al. (2025) uses physical mapping to improve the fidelity of generated interactions. These works optimize their models primarily in kinematic motion space, without considering the physical realism of generated interactions when executed in a physics environment. In this work, we target on physically realistic humanoid interaction execution while simultaneously improving the kinematic interactions.

**Physics-based Humanoid Interaction.** Physics-based humanoid control learns policies that reproduce reference motions while satisfying embodiment and environmental dynamics. DeepMimic Peng et al. (2018) established reinforcement learning-based motion imitation for simulated characters. Subsequent methods learned reusable motion priors and skill representations from larger motion collections, including AMP Peng et al. (2021), ASE Peng et al. (2022), PHC Luo et al. (2023), PULSE Luo et al. (2024), and MoConVQ Yao et al. (2024). These advances enabled robust tracking and control across diverse motions rather than individual reference clips. Recent research has further extended physics-based control from individual characters to contact-rich interactions, developing interaction-aware objectives to preserve spatial coordination, contact relationships, and coupled dynamics in human-object and multi-character settings Zhang et al. (2023); Wang et al. (2023); Ugrinovic et al. (2024); Xu et al. (2025a); Wang et al. (2025b); Shibata et al. (2026).

Building on advances in humanoid control, some studies achieve conditioned humanoid generation in two paradigms. CLoSD Tevet et al. (2025)and Human-X Ji et al. (2025) first generate digital interaction motion, and then execute it using a humanoid tracking policy. In contrast, Uniphys Wu et al. (2025a), PhysReaction Liu et al. (2024) and InterAgent Li et al. (2026) use trajectories obtained through humanoid tracking as supervision for training humanoid generation models. These works demonstrate that physically humanoid execution can be realized by leveraging the capability of tracking policy. However, in complex two-humanoid interactions, relying solely on the tracking policy can lead to missing intended contacts and physically implausible interaction forces. To address this challenge, DIGHT introduces a co-adaptive framework that couples the interaction gen-

erator with the tracking policy through physical execution feedback and explicitly incorporates interaction forces into preference construction for generator alignment.

**Physical Feedback for Motion Model Alignment.** Physical feedback has recently been used to improve the executability of generated human motion. PhysDiff Yuan et al. (2023) and CLoSD Tevet et al. (2025) incorporate physics-based refinement or control during inference. RLPF Yue et al. (2025) and PhysMoDPO Zhang et al. (2026b) instead optimize a motion generator using the feedback from a frozen humanoid tracking policy, making the generated motion more physical plausible. In this work, we incorporate force feedback from the simulator to address physically unrealistic contacts in two-humanoid interactions.

## 3 PRELIMINARIES

**Interaction Tracking Policy.** Following the previous works, we adopt reinforcement learning to learn a tracking policy for imitating motion capture data. Similar to PHC, we utilize a full SMPL humanoid controlled by PD controllers. The learned policy outputs target joint rotations as actions $a_t = \pi(s_t, s_{t+1}^{ref})$, where $s_t$ and $s_{t+1}^{ref}$ denote the current agent state and the future reference state, respectively. Beyond the tracking rewards used in PHC, we adopt an interaction graph reward Zhang et al. (2023); Li et al. (2026) to explicitly model the spatial relationships between agents, encouraging the policy to preserve the inter-agent configurations observed in the reference motions.

**Data Retarget.** To bridge the morphology gap between the tracking policy and the digital generator, we retarget the motion data to a unified humanoid morphology used in the simulation environment. Since retargeting may introduce contact misalignment, such as palm contacts shifting to the wrist or forearm Araujo et al. (2025), we further apply geometry-based contact refinement to recover the original contact relationships while preserving the overall motion structure. The detailed retargeting algorithm is provided in the Appendix H.

**Interaction Diffusion Model.** Text-conditioned human-human interaction models learn the distribution of coordinated motions from MoCap data. Given a clean two-person motion $x_0 = (x_0^A, x_0^B)$ and text description $c$, the forward diffusion process produces $q(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\, x_0, (1 - \bar{\alpha}_t)I)$, where $t$ is the diffusion timestep and $\bar{\alpha}_t$ is determined by a predefined noise schedule. The denoiser jointly processes the two participants and predicts the clean interaction, $\hat{x}_0 = G_\theta(x_t^A, x_t^B, t, c)$. Iterative reverse diffusion then defines the conditional motion distribution $p_\theta(x_0 \mid c)$. We use InterGen (Liang et al., 2024) and TIMotion (Wang et al., 2025a) as generator backbones, and both are optimized in the clean-motion prediction parameterization used by our preference objective.

## 4 METHOD

### 4.1 OVERVIEW

DIGHT couples a digital interaction generator $g_\theta(x \mid c)$ with a physics-based humanoid tracking policy $\pi_\phi$. Given text $c$, the generator generates a digital interaction motion $x = (x^A, x^B)$, and the tracker executes it in simulation to produce a physics-based humanoid interaction motion $\tau = \mathrm{Sim}(x; \pi_\phi)$. As shown in Fig. 2, DIGHT performs generator-tracker co-adaptation in two stages. In the first stage, the pretrained generator samples multiple candidate motions for each training prompt. The frozen pretrained tracker executes these candidates in simulation. The resulting rollouts are then evaluated to construct five physics-grounded preference sets based on tracking error, frictional dissipation, floating, contact occurrence, and contact quality, which are used to post-train the generator. Second, we freeze the aligned generator and use its sampled motions as references to adapt the tracker to its original tracking objective. The two components are therefore adapted using information induced by each other.

### 4.2 PHYSICS-GROUNDED PREFERENCE CONSTRUCTION

For each text-motion example $(c, x^{\mathrm{gt}})$, we sample $K$ candidates $x_i \sim g_\theta(x \mid c)$ and execute them with the same frozen tracker, obtaining $\tau_i = \mathrm{Sim}(x_i; \pi_\phi)$. We also execute $x^{\mathrm{gt}}$ to obtain $\tau^{\mathrm{gt}}$, which provides the reference contact pattern but is not automatically used as the preferred sample. We construct five physics-grounded preference pairs as follows.

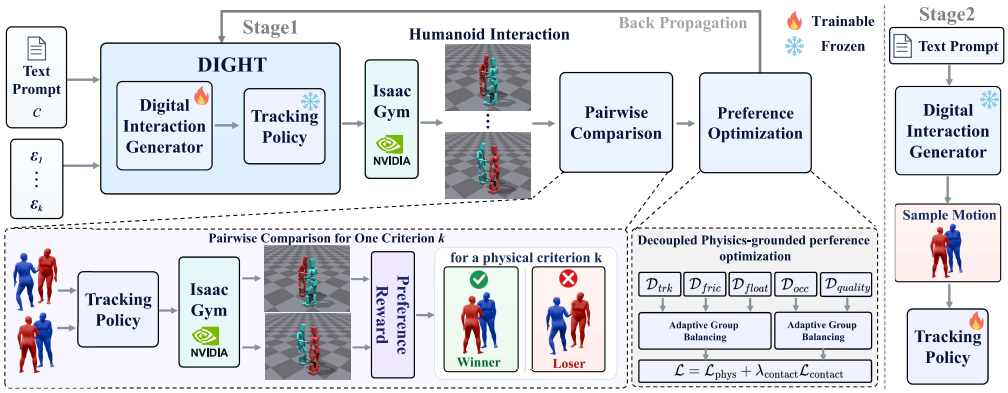



Figure 2: **Overview of DIGHT.** DIGHT consists of an interaction generator and a tracking policy, which are co-adapted in two stages. First, multiple interaction candidates produced by the generator are executed by the frozen tracker, whose rollouts are used to construct physics-grounded preferences for tracking, frictional dissipation, floating, contact occurrence, and contact quality. These preferences align the generator through physics-decoupled diffusion DPO. Second, motions from the aligned generator are used to adapt the tracking policy, improving their compatibility.

**General Physical Preferences.** For every rollout, we compute three general physical scores: mean body tracking deviation $E_{\text{trk}}$, frictional dissipation $E_{\text{fric}}$, and the float $E_{\text{float}}$. For each criterion $k \in \{\text{trk}, \text{fric}, \text{float}\}$, a valid winner must have a lower corresponding error than the loser:

$$E_k(\tau^+) < E_k(\tau^-), \qquad k \in \{\text{trk}, \text{fric}, \text{float}\}, \tag{1}$$

where $\tau^+$ and $\tau^-$ are the executions of the preferred and rejected candidates, respectively. The winner is additionally required to have better motion-text alignment than the loser.

**Contact Occurrence Preference.** Beyond general physical plausibility, two-person interactions differ from single-person in their need to preserve meaningful contact. To capture such contact, we leverage the force feedback provided by the physics simulator. We detect frame-level contacts between rigid bodies and then consolidate them into persistent contacts between semantic body regions. Specifically, each rigid body is assigned to one of six regions: the core, head, arms, hands, legs, or feet. At each frame, inter-person contact forces are grouped according to the corresponding pair of body regions. For every region pair, we slide a fixed-length temporal window over the sequence and compute the contact ratio, defined as the proportion of frames within the window in which contact forces are detected. A region pair is labeled as a persistent contact if its contact ratio reaches a predefined threshold in at least one window. This procedure filters out isolated collisions while remaining tolerant to brief interruptions in otherwise sustained contacts. We collect all persistent region pairs in $\mathcal{C}(\tau)$, which denotes the persistent contact topology of trajectory $\tau$. We use this topology to construct contact occurrence preferences. If $\mathcal{C}(\tau^{\text{gt}})$ is nonempty, the winner must reproduce at least one reference contact without introducing any persistent contact outside $\mathcal{C}(\tau^{\text{gt}})$, whereas the loser must contain no persistent contact. If $\mathcal{C}(\tau^{\text{gt}})$ is empty, the winner must remain contact-free, whereas the loser must contain at least one persistent contact.

**Contact Quality Preference.** For contact-positive motions, contact occurrence alone does not ensure interaction quality: the contact may be transient or poorly sustained, and even persistent contact may involve implausibly weak or excessive forces. We therefore measure three complementary aspects of contact quality relative to the $\tau^{\text{gt}}$: $E_{\text{topo}}$ measures the topology fidelity of persistent region-level contacts, $E_{\text{dur}}$ measures the duration fidelity of the longest episode for each corresponding contact edge, and $E_{\text{force}}$ measures its force fidelity in terms of the mean and maximum forces within that episode. A contact quality winner must be better in all component:

$$E_j(\tau^+) < E_j(\tau^-), \ \forall j, \tag{2}$$

where $j \in \{\text{topo}, \text{dur}, \text{force}\}$. These pairs are also constrained by the motion-text alignment. Finally, we construct the five independent preference sets $\mathcal{D}_k = \{(c, x^+, x^-) : x^+ \succ_k x^-\}$ for $k \in \mathcal{T} = \{\text{trk}, \text{fric}, \text{float}, \text{occ}, \text{quality}\}$. The detailed definitions of each preference score are provided in Appendix D.

### 4.3 PHYSICS-DECOUPLED DIFFUSION DPO

A motion can be physically preferable under one criterion while being inferior under another. Collapsing all physical signals into a single scalar reward for candidate ranking, discarding criterion-specific supervision. We therefore optimize a separate decoupled diffusion DPO objective for each preference set $\mathcal{D}_k$. During generator alignment, the trainable denoiser $\theta$ and the frozen reference denoiser $\theta_{\text{ref}}$ are initialized from the same pretrained generator. Since both backbones predict the clean motion, we define the relative denoising reward

$$r_\theta(x,t,c) = \|x - \hat{x}_{\theta_{\text{ref}}}(x_t,t,c)\|_2^2 - \|x - \hat{x}_\theta(x_t,t,c)\|_2^2, \tag{3}$$

where $x_t$ is the noisy motion at timestep $t$ and $\hat{x}_\theta$ denotes the predicted clean motion. For a preference pair $(c, x^+, x^-) \sim \mathcal{D}_k$, we define

$$R_k = r_\theta(x^+,t,c) - r_\theta(x^-,t,c), \qquad \mathcal{L}_k = -\mathbb{E}\left[\log \sigma(\beta R_k)\right], \tag{4}$$

where the expectation is over preference pairs, uniformly sampled diffusion timesteps, and independent forward-process noise for $x^+$ and $x^-$. Here, $\sigma$ is the sigmoid function and $\beta > 0$ controls the preference strength. A larger $R_k$ indicates that the trainable denoiser improves the preferred motion over the rejected one relative to the frozen reference denoiser.

We group tracking, friction, and floating into $\mathcal{G}_{\text{phys}}$, and occurrence and contact quality into $\mathcal{G}_{\text{contact}}$. Following BideDPO (Zhou et al., 2026), for either group $\mathcal{G} \in \{\mathcal{G}_{\text{phys}}, \mathcal{G}_{\text{contact}}\}$, their current mini-batch losses determine detached within-group weights:

$$\alpha_k^{(\mathcal{G})} = \text{sg}\left(\frac{\mathcal{L}_k}{\sum_{j\in\mathcal{G}} \mathcal{L}_j}\right), \qquad \mathcal{L}_{\mathcal{G}} = \sum_{k\in\mathcal{G}} \alpha_k^{(\mathcal{G})} \mathcal{L}_k, \qquad \mathcal{L} = \mathcal{L}_{\mathcal{G}_{\text{phys}}} + \lambda_{\text{contact}} \mathcal{L}_{\mathcal{G}_{\text{contact}}}, \tag{5}$$

where sg denotes stop-gradient and $\lambda_{\text{contact}}$ balances contact fidelity against general executability. Only the trainable denoiser $\theta$ is updated.

### 4.4 TRACKER ADAPTATION

After aligning the generator, we use it to provide reference motions for fine-tuning the tracker. Specifically, we freeze the aligned generator $g_{\theta_{\text{align}}}$, sample interaction motions $x \sim g_{\theta_{\text{align}}}(x \mid c)$, and use them as the reference motions to fine-tune the pretrained tracker with its original tracking objective. This stage adapts the tracker to the motion distribution of the aligned generator, thereby further improving compatibility between motion generation and physical execution. At inference time, $p_{\theta_{\text{align}}}$ generates the reference interaction and the adapted tracker $\pi_{\phi_{\text{adapt}}}$ executes it.

## 5 EXPERIMENTS

### 5.1 EXPERIMENT SETUP

In our experiments, we apply our approach to the commonly used model InterGen Liang et al. (2024) and the state-of-the-art model TIMotion Wang et al. (2025a). We evaluate both the quality of their generated digital motions and the quality of their physical execution by humanoid agents.

**Datasets and Evaluation Configurations.** We conduct experiments on two datasets, InterHuman Liang et al. (2024) and Inter-X Xu et al. (2024) datasets, which provide text annotations and are designed for evaluating multi-human interaction generation. Our evaluation has two parts: generator evaluation assesses the quality of the generated digital interaction motions, while humanoid execution evaluation assesses their physical execution by the target humanoid tracker in the simulation environment. The results of Inter-X are shown in Appendix B.

**Evaluation Metrics.** To comprehensively evaluate both the quality of generated interactions and their alignment with text descriptions, we adopt eight metrics: (1) R-Precision and MM-Dist for text-motion consistency; (2) FID for motion quality; (3) Float, Skate, Interpen, and Jerk for physical plausibility, including stability, foot contact, body penetration, and motion smoothness; and (4) Force Error for the interaction fidelity of humanoid execution by comparing the mean contact forces of the candidate and ground-truth executions. In Tab. 1, the first seven metrics are computed directly

Table 1: **Quantitative comparisons on InterHuman test set.** The best results are **bold**. DIGHT[1] and DIGHT[2] denote our framework initialized from InterGen and TIMotion, respectively. The Force Error is evaluated from simulated executions using the same pretrained tracker, while the other metrics evaluate the outputs of the generators.

| Methods | FID ↓ | MM Dist ↓ | R-Precision (Top-1) ↑ | Skate ↓ | Float ↓ | Interpen ↓ | Jerk ↓ | Force Error ↓ |
|---|---|---|---|---|---|---|---|---|
| InterGen | 1.539 | 0.954 | 0.474 | 7.27 | 22.40 | 10.69 | 148.96 | 577.41 |
| InterGen + SFT | 1.878 | 0.965 | 0.461 | 6.57 | 21.54 | 13.12 | 129.13 | 543.62 |
| InterGen + DPO | 1.501 | 0.955 | 0.472 | 6.04 | 20.96 | 9.85 | 134.82 | 491.40 |
| DIGHT[1] | **1.474** | **0.942** | **0.486** | **5.67** | **18.58** | **8.42** | **111.94** | **427.98** |
| TIMotion | 1.341 | 0.931 | 0.536 | 11.69 | 21.63 | 9.68 | 322.02 | 619.25 |
| DIGHT[2] | **1.297** | **0.927** | **0.543** | **8.53** | **18.28** | **7.83** | **199.76** | **451.33** |

Table 2: **Compared with humanoid generation on InterHuman test set.** The execution trajectory of DIGHT is executed by the adapted tracker.

| Methods | FID ↓ | MM Dist ↓ | R-Precision ↑ | | | Skate ↓ | Jerk ↓ | Force Error ↓ |
|---|---|---|---|---|---|---|---|---|
| | | | Top-1 | Top-2 | Top-3 | | | |
| InterAgent | 2.533 | 1.126 | 0.335 | 0.470 | 0.552 | 9.55 | 389.51 | 588.63 |
| DIGHT | **1.426** | **1.054** | **0.417** | **0.565** | **0.644** | **6.23** | **231.28** | **413.47** |

on the generated digital interactions, whereas Force Error is evaluated at the simulator level under the same pretrained tracker. Tab. 2 reports all metrics computed on simulated humanoid trajectories. To compute the text-motion consistency metrics, we separately train two motion-text encoder pairs with the contrastive loss Petrovich et al. (2023): one on text-motion pairs and the other on text-humanoid pairs. Full definitions and more details are provided in Appendix E and Appendix J.

**Implementation Details.** Following InterGen Liang et al. (2024) and TIMotion Wang et al. (2025a), we adopt their human–human interaction models as the diffusion backbones. For each training prompt in InterHuman and Inter-X, we sample $K = 12$ candidate motions from the corresponding diffusion model. We use NVIDIA Isaac Gym Makoviychuk et al. (2021) as the physics simulation environment, in which the candidates are evaluated through tracking rollouts to construct preference pairs. During training, DIGHT is optimized using the AdamW optimizer with a batch size of 32, and the learning rate is set to $2 \times 10^{-5}$. During inference, we generate motions over 50 steps using the DDIM strategy and classifier-free text guidance.

### 5.2 Generator Evaluation

**Quantitative Results.** Tab. 1 presents the quantitative results of the generator in DIGHT on the InterHuman. To demonstrate the generality of our framework, we instantiate DIGHT with two different motion generators, InterGen and TIMotion, denoted as DIGHT[1] and DIGHT[2], respectively. Across both backbones, our method consistently reduces all physical artifacts while largely preserving motion quality and text-motion alignment. DIGHT[1] and DIGHT[2] reduces the five physical metrics by averages of 22.1% and 25.3%, respectively. These results demonstrate that physics-grounded preferences introduce physical supervision into the digital generator, improving the physical plausibility of its generated results. The quantitative results of Inter-X are shown in Appendix B.

**Qualitative Results.** In Fig. 3, we present visual comparisons between DIGHT and the InterGen baseline. It can be seen that InterGen struggles to produce physically plausible interactions when executed in the simulation environment, often failing to establish intended contacts (e.g., during handshaking) or offer necessary physical force to support (e.g., when helping a falling person stand up). In contrast, our approach is able to generate motions that better preserve the intended contacts, while maintaining more natural and physically plausible interactions between the two humanoid. These observations demonstrate the superior ability of DIGHT to generate physically plausible human-human interactions. Please find more visual comparison results in the Appendix A.

### 5.3 Humanoid Execution Evaluation

**Comparison with Direct Humanoid Generation.** We further compare DIGHT with InterAgent Li et al. (2026), a state-of-the-art method for directly generating humanoid motions. As shown in

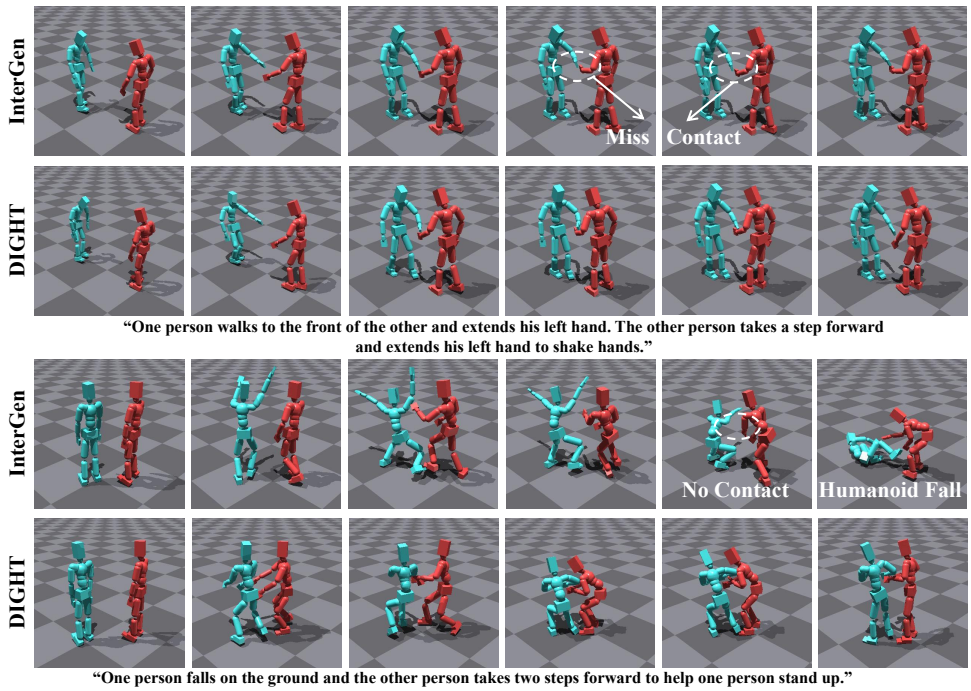



Figure 3: **Qualitative comparison.** We compare our DIGHT with InterGen under various text prompts. Our method generates more natural and physically plausible humanoid interactions, with more realistic and coherent physical contacts, while better aligning with the given text.

Table 3: **Ablation study of different preference combinations on InterHuman test set.**

| Reward | Skate | Float | Interpen | Jerk | Force Error |
|---|---|---|---|---|---|
| - | 7.27 | 22.40 | 10.69 | 148.96 | 577.41 |
| trk | 6.74 | 20.12 | 9.93 | 138.16 | 528.37 |
| + fric | 5.71 | 19.85 | 9.44 | 130.21 | 511.82 |
| + float | 5.93 | 18.81 | 9.51 | 124.57 | 496.49 |
| + contact | **5.67** | **18.58** | **8.42** | **111.94** | **427.98** |

Table 4: **Performance in the simulation environment.** DIGHT* denotes the result without fine-tuning the tracking policy.

| Method | Succ. (%) | Fall (%) | MBD (cm) |
|---|---|---|---|
| InterGen | 88.73 | 9.32 | 10.36 |
| DIGHT* | 91.45 | 6.28 | 8.22 |
| DIGHT | **92.19** | **5.77** | **7.86** |

Tab. 2, DIGHT consistently outperforms InterAgent across all evaluation metrics, achieving better motion quality and text alignment while producing more physically plausible interactions.

**Contact Fidelity.** We further evaluate DIGHT on text prompts requiring physical contact. To obtain the annotations required for this evaluation, we use GPT-5.6-sol to extend the InterHuman test set by determining from each textual description whether physical contact is required and which body-region pairs are expected to make contact. Detailed annotation procedures are provided in the Appendix G. As shown in Tab. 5, compared to baseline, DIGHT increases contact topology precision from 38.29% to 46.77% and recall from 52.16% to 63.94%. Meanwhile, the success tracking rate of contact motion increases from 84.70% to 89.23%. These improvements indicate that the proposed contact-related preference optimization enables more accurate and reliable contact interactions, rather than simply increasing unnecessary contacts.

Table 5: **Performance on text prompts requiring physical contact.**

| Method | Topo. Prec. | Topo. Recall | Succ. |
|---|---|---|---|
| InterGen | 38.29 | 52.16 | 84.70 |
| DIGHT | **46.77** | **63.94** | **89.23** |

**Effect of Generator-tracker Co-adaptation.** Tab. 4 evaluates the execution performance of the baseline, DIGHT without tracker adaptation, and the complete DIGHT. Aligning the generator through physics-grounded preference optimization raises execution success from 88.73% to 91.45%, reduces fall rate from 9.32% to 6.28%, and lowers tracking error (MBD) from 10.36 cm to 8.22 cm. Adapting the tracker to this updated generator provides further gains on all three metrics. These results show that physics-grounded preference optimization produces more physically plausible interactions that are easier for the humanoid tracker to execute, while adapting the tracker to the updated motion distribution further improves execution robustness and tracking accuracy. Notably, our framework naturally supports iterative generator-tracker co-adaptation: the adapted tracker can

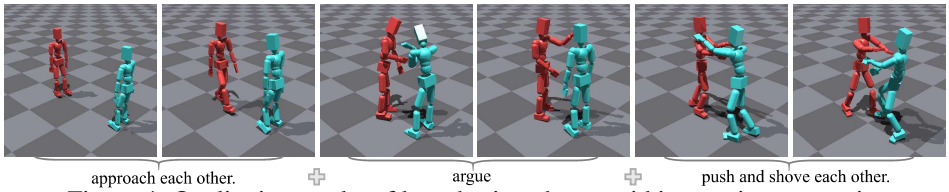
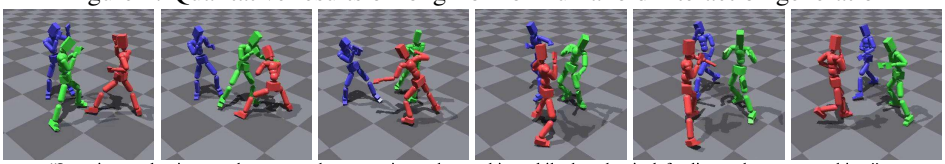



Figure 4: Qualitative results of long-horizon humanoid interaction generation



"In an intense boxing match, two people are continuously punching while the other is defending and counterattacking."

Figure 5: Qualitative results of multi-humanoid interaction generation.

be used to collect updated physical feedback for the next round of generator optimization, followed by further tracker adaptation. This iterative strategy is optional. We use a single round co-adaptation in the main experiments and study the effect of iterative strategy in the Appendix C.

### 5.4 Discussion

**Effect of Different Physics-grounded Preferences.** To study the contribution of each preference criterion, we progressively incorporate tracking, friction, floating, and contact preferences into generator optimization. As shown in Tab. 3, the tracking preference improves all five physical metrics, while adding friction and floating preferences further reduces the corresponding skating and floating artifacts. Finally, adding the contact reward yields the best overall performance. These results demonstrate that the different physics-grounded preferences provide complementary supervision for improving both general executability and interaction fidelity.

**Effect of Physics-decoupled DPO Strategy.** To evaluate the effectiveness of our proposed physics-decoupled DPO strategy, we compare it against standard DPO and supervised fine-tuning (SFT). As shown in Tab 1, SFT leads to degradation in semantic alignment, motion quality and the physical metric of interpenetration. Standard DPO improves physical plausibility to some extent, but the gains remain relatively modest. In contrast, our method consistently outperforms both the SFT and the standard DPO in all evaluation metrics, demonstrating the effectiveness of the proposed physics-decoupled optimization strategy.

**Application Extensions.** We further evaluate DIGHT on two challenging applications: long-horizon temporal composition and multi-humanoid interaction. For long-horizon interaction, following PriorMDM Shafir et al. (2024), we apply temporal composition to the generator to synthesize a continuous two-person reference motion from multiple temporally ordered instructions, which is subsequently executed by the adapted humanoid tracker. As shown in Fig. 4, two people walk toward each other, meet face to face, argue, and eventually push one another. The physical execution preserves this temporal progression and produces a coherent long-horizon interaction. We also investigate an $N = 3$ scenario beyond the pairwise setting. Following Xu et al. Xu et al. (2025b), given the prompt "Three humanoids boxing", DIGHT produces three humanoids simultaneously participate in the interaction. As shown in Fig. 5, the humanoids change their relative configurations and perform distinct boxing-like motions toward different partners. These results demonstrate the extensibility of DIGHT to both temporally composed and multi-humanoid interactions.

## 6 Conclusion

In this paper, we propose DIGHT, a two-stage co-adaptive framework consisting of an interaction generator and a humanoid tracking policy. Using simulator feedback, DIGHT constructs physics-grounded preferences for general executability and interaction fidelity, with the latter explicitly modeling the contact occurrence, duration, and interaction forces. These preferences post-train the generator through physics-decoupled diffusion DPO. The aligned generator is subsequently used to adapt the tracking policy to the updated motion distribution. Extensive experiments demonstrate that DIGHT improves physical executability and interaction fidelity.

## AI USE STATEMENT

In this work, we used OpenAI's GPT model for two purposes. First, GPT-5.6-Sol was used to generate initial annotations for a dataset. All AI-generated annotations were subsequently reviewed item by item by human annotators and corrected where necessary. Second, we use the ChatGPT to polish the language of the manuscript, including its grammar, clarity, and readability. It was not used to generate or alter the research ideas, hypotheses, methodology, experimental design, results, interpretations, scientific claims, or conclusions of this work. Other tasks listed as requiring disclosure under the ICLR policy were either not performed with generative AI or were not applicable to this work. We reviewed all AI-assisted work and take full responsibility for the final content of the paper and the accompanying dataset.

## REFERENCES

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

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.

Dahua Gao, Wenlong Wang, Xinyu Liu, Yuxi Hu, and Danhua Liu. Physics-guided human interaction generation via motion diffusion model. *Computer Vision and Image Understanding*, pp. 104470, 2025.

Wei-Jin Huang, Yueyi Zhang, Yi-Lin Wei, Zhi-Wei Xia, Juantao Tan, Yuan-Ming Li, Zhi-Lin Zhao, and Wei-Shi Zheng. Learning whole-body human-humanoid interaction from human-human demonstrations. *ArXiv*, abs/2601.09518, 2026. URL `https://api.semanticscholar.org/CorpusID:284718143`.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. *arXiv preprint arXiv:2410.10010*, 2024.

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

Bin Li, Ruichi Zhang, Han Liang, Jingyan Zhang, Juze Zhang, Xin Chen, Lan Xu, Jingyi Yu, and Jingya Wang. Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15253–15265, 2026.

Boyuan Li, Xihua Wang, Ruihua Song, and Wenbing Huang. Two-in-one: Unified multi-person interactive motion generation by latent diffusion transformer. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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.

Stefan Lionar and Gim Hee Lee. Teamhoi: Learning a unified policy for cooperative human-object interactions with any team size. *arXiv preprint arXiv:2603.07988*, 2026.

Mengge Liu, Yan Di, Gu Wang, Yun Qu, Dekai Zhu, Yanyan Li, and Xiangyang Ji. Hint: Hierarchical interaction modeling for autoregressive multi-human motion generation. *arXiv preprint arXiv:2601.20383*, 2026.

Yunze Liu, Changxi Chen, Chenjing Ding, and Li Yi. Physreaction: Physically plausible real-time humanoid reaction synthesis via forward dynamics guided 4d imitation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3771–3780, 2024.

Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *International Conference on Learning Representations*, volume 2024, pp. 56766–56782, 2024.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Esteve Valls Mascaro, Yashuai Yan, and Dongheui Lee. Robot interaction behavior generation based on social motion forecasting for human-robot interaction. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 17264–17271, 2024. URL https://api.semanticscholar.org/CorpusID:267523356.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022.

Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9454–9463. IEEE, 2023.

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

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.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations*, volume 2024, pp. 8717–8733, 2024.

Yuto Shibata, Kashu Yamazaki, Lalit Jayanti, Yoshimitsu Aoki, Mariko Isogawa, and Katerina Fragkiadaki. Learning to assist: Physics-grounded human-human control via multi-agent reinforcement learning. *ArXiv*, abs/2603.11346, 2026. URL https://api.semanticscholar.org/CorpusID:286489400.

Kewei Sui, Anindita Ghosh, In Tae Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *International Journal of Computer Vision*, 134, 2025. URL https://api.semanticscholar.org/CorpusID:277065610.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Bermano, and Michiel Van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations*, volume 2025, pp. 46506–46520, 2025.

Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas J. Guibas. Multiphys: Multi-person physics-aware 3d motion estimation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2331–2340, 2024. URL https://api.semanticscholar.org/CorpusID:269214675.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7169–7178. IEEE, 2025a.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17540–17549. IEEE, 2025b.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. *Advances in Neural Information Processing Systems*, 37:105397–105424, 2024.

Qingxuan Wu, Zhiyang Dou, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, Lingjie Liu, et al. Text2interact: High-fidelity and diverse text-to-two-person interaction generation. In *International Conference on Learning Representations*, volume 2026, pp. 103856–103875, 2026.

Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. Uniphys: Unified planner and controller with diffusion for flexible physics-based character control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13214–13224. IEEE, 2025a.

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, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22260–22271. IEEE, 2024.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12266–12277. IEEE, 2025a.

Wenning Xu, Shiyu Fan, Paul Henderson, and Edmond SL Ho. Multi-person interaction generation from two-person motion priors. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025b.

Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Transactions on Graphics (TOG)*, 43(4):1–21, 2024.

Wei Yao, Yunlian Sun, Chang Liu, Hongwen Zhang, and Jinhui Tang. Physiinter: Integrating physical mapping for high-fidelity human interaction generation. *arXiv preprint arXiv:2506.07456*, 2025.

Heng Yu, Juze Zhang, Changan Chen, Tiange Xiang, Yusu Fang, Juan Carlos Niebles, and Ehsan Adeli. Socialgen: Modeling multi-human social interaction with language models. In *2026 International Conference on 3D Vision (3DV)*, pp. 1–17. IEEE, 2026.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15964–15975. IEEE, 2023.

Junpeng Yue, Zepeng Wang, Yuxuan Wang, Weishuai Zeng, Jiangxing Wang, Xinrun Xu, Yu Zhang, Sipeng Zheng, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion models with humanoid control. *arXiv preprint arXiv:2506.12769*, 2025.

Xiaotang Zhang, Ziyi Chang, Qianhui Men, and Hubert P. H. Shum. Physics-based motion tracking of contact-rich interacting characters. *Comput. Graph. Forum*, 45, 2026a. URL https://api.semanticscholar.org/CorpusID:286921609.

Yangsong Zhang, Anujith Muraleedharan, Rikhat Akizhanov, Abdul Ahad Butt, Gül Varol, Pascal Fua, Fabio Pizzati, and Ivan Laptev. Physmodpo: Physically-plausible humanoid motion with preference optimization. *arXiv preprint arXiv:2603.13228*, 2026b.

Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Dewei Zhou, Mingwei Li, Zongxin Yang, Yu Lu, Yunqiu Xu, Zhizhong Wang, Zeyi Huang, and Yi Yang. Bidedpo: Conditional image generation with simultaneous text and condition alignment. In *International Conference on Learning Representations*, volume 2026, pp. 54113–54126, 2026.

# APPENDIX

In this appendix, we present the following:

- Section A: More visualization results
- Section B: Results on Inter-X
- Section C: Iterative optimization strategy.
- Section D: Additional method details
- Section E: Full metrics definition.
- Section F: More details of datasets.
- Section G: Annotated contact on InterHuman testset.
- Section H: Motion retarget.
- Section I: Tracking policy.
- Section J: Text-motion evaluators.

## A MORE VISUALIZATION RESULTS

In this section, we provide additional visual comparisons, a long-horizon humanoid interaction sequence, multi-humanoid interaction results, a highly dynamic interaction result, and a rich contact interaction result. Please refer to *vis_page.html* in the supplementary material for video results and additional comparisons with InterGen.

**More Comparison Results.** As shown in Figs. 6, 7, and 8, we provide six additional sets of qualitative comparisons with the baseline. These examples further demonstrate that our method produces more physically plausible motions while better preserving the intended interactions.

**Long-horizon Humanoid Interaction Result.** As shown in Fig. 10, we present a long-horizon interaction sequence composed of multiple dynamically diverse actions. The resulting humanoid motion exhibits coherent transitions between the constituent actions and remains physically stable throughout the sequence.

**Multi-humanoid Interaction Results.** As shown in Fig. 9, we additionally present three interaction scenarios involving three or four humanoids. Our method successfully generates and executes these multi-humanoid interactions while maintaining coordinated behavior among all participants.

**Highly Dynamic and Contact-rich Interaction Results.** As shown in Figs. 11 and 12, we additionally present a boxing sequence with sustained contact and a highly dynamic kicking sequence. The boxing example demonstrates coordinated motion during prolonged close-range contact, while the kicking example involves rapid limb movements, substantial shifts in balance, and dynamic contact. In both cases, our method produces physically stable motions while preserving the intended interactions.

## B RESULTS ON INTER-X

To further demonstrate the generalizability of our framework, we conduct experiments on the Inter-X Xu et al. (2024) dataset using both InterGen Liang et al. (2024) and TIMotion Wang et al. (2025a)

Table 6: **Quantitative comparisons on Inter-X test set.** The best and second best results are **bold** and underlined. DIGHT[1] and DIGHT[2] denote our framework initialized from InterGen and TIMotion, respectively. The Force Error is evaluated from simulated executions using the same pretrained tracker, while the other metrics evaluate the outputs of the generators.

| Methods | FID ↓ | MM Dist ↓ | R-Precision (Top-1) ↑ | Skate ↓ | Float ↓ | Interpen ↓ | Jerk ↓ | Force Error ↓ |
|---|---|---|---|---|---|---|---|---|
| InterGen | 1.162 | 0.884 | 0.578 | 4.42 | 14.43 | 16.01 | 63.04 | 236.54 |
| InterGen + SFT | 1.639 | 0.897 | 0.570 | 3.89 | 14.35 | 21.60 | 59.47 | 347.07 |
| InterGen + DPO | 1.125 | 0.886 | 0.578 | 4.17 | 12.66 | 14.97 | 57.33 | 223.68 |
| DIGHT[1] | 1.107 | 0.883 | 0.581 | **3.62** | **10.87** | **13.94** | **52.88** | **209.92** |
| TIMotion | 1.054 | 0.863 | 0.599 | 7.65 | 15.39 | 13.77 | 97.52 | 268.13 |
| DIGHT[2] | 1.013 | 0.865 | 0.594 | 5.97 | 12.42 | 11.59 | 76.80 | 232.86 |

as generator backbones. As shown in Tab. 6, DIGHT consistently improves all five physical metrics across both backbones while largely preserving motion quality and text-motion alignment. When initialized from InterGen, DIGHT reduces the five physical errors by an average of 16.6%. Similar improvements are observed with TIMotion, where the average relative reduction reaches 18.3%. These consistent gains across a different interaction dataset and two distinct generator architectures demonstrate that our physics-grounded alignment framework generalizes beyond a specific dataset or backbone.

## C Iterative optimization strategy.

In main paper, we update the generator from the physical feedback of tracker and then adapts the tracker once, but the same two-stage procedure can optionally be repeated. As shown in Tab. 7, increasing the iterations from 1 to 3 improves the success rate from 92.19% to 93.05% and reduces the fall rate from 5.77% to 4.96%, while MBD decreases from 7.86 cm to 7.41 cm. Further increasing the iterations to 4 slightly degrades the performance. Given that the single iteration already achieves strong results, while additional iterations provide comparatively modest gains and require renewed simulation and optimization, we use the single iteration in our main experiments and treat iteration refinement as an optional extension.

Table 7: **Effect of iterative optimization.**

| Iter. | Succ. (%) | Fall (%) | MBD (cm) |
|---|---|---|---|
| 1 | 92.19 | 5.77 | 7.86 |
| 2 | 92.88 | 5.31 | **7.32** |
| 3 | **93.05** | **4.96** | 7.41 |
| 4 | 92.51 | 5.48 | 7.68 |

## D Additional Method Details

### D.1 General Physical Preferences

For each execution trajectory $\tau$. Let $q_t^{p,b}$ and $\bar{q}_t^{p,b}$ denote the executed and reference positions of body $b \in \mathcal{B}$ for participant $p \in \{A, B\}$ at frame $t$. The tracking error is the mean body deviation

$$E_{\mathrm{trk}}(\tau) = \frac{1}{2T|\mathcal{B}|} \sum_{p\in\{A,B\}} \sum_{t=1}^{T} \sum_{b\in\mathcal{B}} \left\| q_t^{p,b} - \bar{q}_t^{p,b} \right\|_2 . \tag{6}$$

Let $D_t^{p,b}$ be the tangential contact dissipation reported by the simulator for an active ground-contact body $b \in \mathcal{F}$. We aggregate this physical loss as

$$E_{\mathrm{fric}}(\tau) = \frac{1}{2T} \sum_{p\in\{A,B\}} \sum_{t=1}^{T} \sum_{b\in\mathcal{F}} D_t^{p,b}. \tag{7}$$

Let $z_t^{p,b}$ be the body height in the ground coordinate at frame $t$, the Float metric is

$$E_{\mathrm{float}} = \frac{1}{2T} \sum_{p,u} \max\left( \min_{j\in\mathcal{J}} z_{u,j}^{p} - z_{\mathrm{floor}}, 0 \right). \tag{8}$$

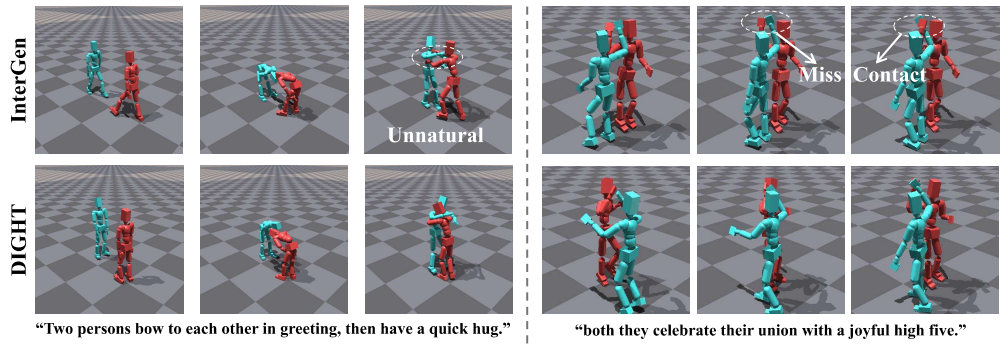



Figure 6: Additional qualitative comparisons between our method and the baseline.

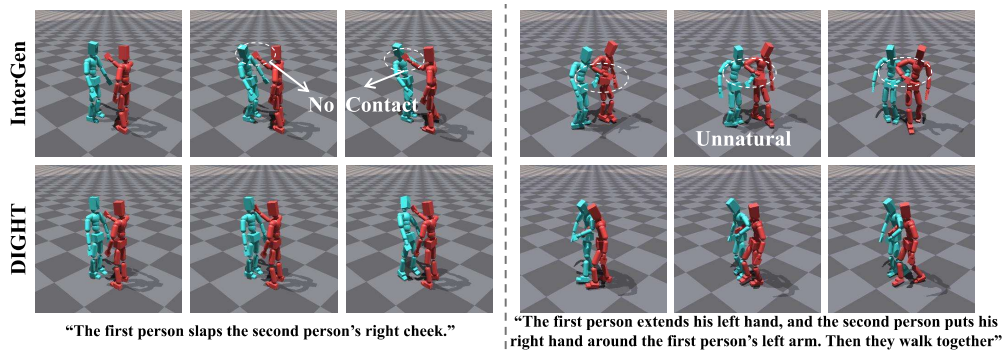



Figure 7: Additional qualitative comparisons between our method and the baseline.

We also measure text-motion distance $d_(x, c)$ between the generated motion $x$ and corresponding text $c$ using the frozen text feature and motion feature encoders $f_{\text{motion}}$ and $f_{\text{text}}$,

$$d_(x, c) = \|f_{\text{motion}}(x) - f_{\text{text}}(c)\|_2 \,. \tag{9}$$

For target criterion $k$, a pair $(i, j)$ is valid only if $E_k(\tau_i) < E_k(\tau_j)$ and $d(x_i, c) < d_(x_j, c)$.

### D.2 Contact Representation and Comparison

As defined in the main paper, we use $C(\tau)$ denoting the persistent region pair set for trajectory $\tau$. To calculate the contact quality, we define the normalized topology discrepancy between region pair sets $A$ and $B$ as:

$$d_{\text{topo}}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \tag{10}$$

with value zero when both sets are empty. And define the bounded log error as:

$$d_{\log}(a, b; \epsilon) = \frac{|\log(a + \epsilon) - \log(b + \epsilon)|}{|\log(a + \epsilon) - \log(b + \epsilon)| + \log 2}. \tag{11}$$

**Contact-topology Error.** For each humanoid trajectory $\tau$, its contact-topology error is therefore:

$$E_{\text{topo}}(\tau) = d_{\text{topo}}\big(\mathcal{C}(\tau), \mathcal{C}(\tau^{\text{gt}})\big)\,, \tag{12}$$

where $\mathcal{C}(\tau)$ as defined in the main paper denotes the persistent region pairs set.

**Edge-wise Contact Duration and Force.** Let $\mathcal{R} = \{\text{core}, \text{head}, \text{arm}, \text{hand}, \text{leg}, \text{foot}\}$ denote the six body regions. Each region pair $e = (r^A, r^B) \in \mathcal{R} \times \mathcal{R}$ corresponds to one entry of the $6 \times 6$ contact graph. For each edge $e$, we group its active frames into contact episodes. Before extracting the episodes, we fill inactive gaps of fewer than four frames; thus, two contact segments separated

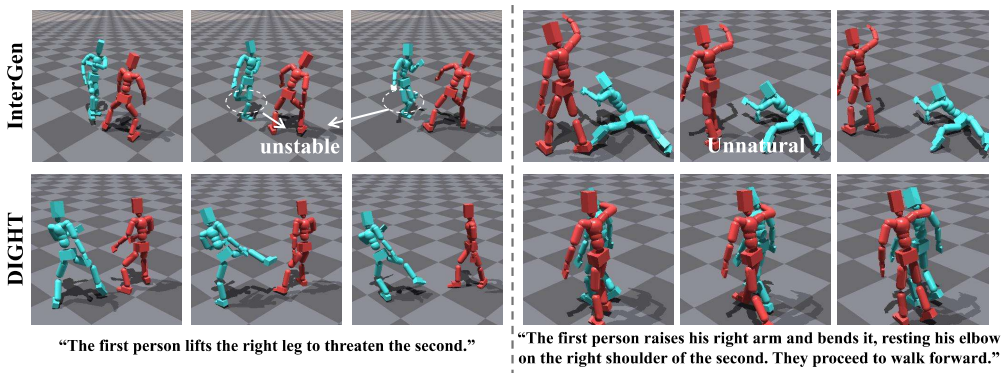



Figure 8: Additional qualitative comparisons between our method and the baseline.

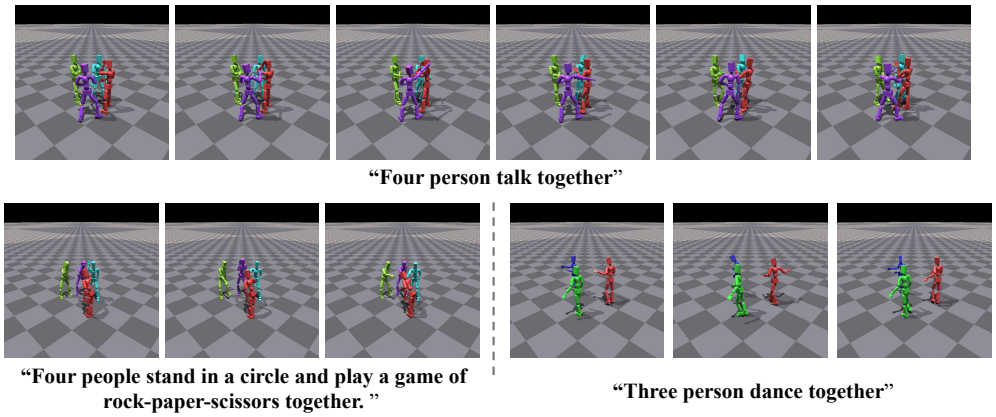



Figure 9: Additional multi-humanoid interaction results.

by one to three inactive frames are merged into the same episode. We denote the longest resulting episode in rollout $\tau$ by $I_e^*(\tau)$. Its contact duration is

$$D_e(\tau) = |I_e^*(\tau)|. \tag{13}$$

We set $D_e(\tau) = 0$ if edge $e$ has no contact episode.

Let $\mathcal{P}_{e,t}$ denote the active rigid-body pairs mapped to region edge $e$ at frame $t$. The aggregate force of this region edge is

$$F_{e,t}(\tau) = \sum_{(i,j)\in\mathcal{P}_{e,t}} \left\| \sum_{m\in\mathcal{M}_{i,j,t}} \mathbf{f}_{m,t} \right\|_2, \tag{14}$$

where $\mathcal{M}_{i,j,t}$ contains the contact manifolds between rigid bodies $i$ and $j$. Each contact force is included only once. We characterize the force of edge $e$ by its mean and maximum values within its longest contact episode:

$$\bar{F}_e(\tau) = \frac{1}{|I_e^*(\tau)|} \sum_{t\in I_e^*(\tau)} F_{e,t}(\tau), \qquad F_e^{\max}(\tau) = \max_{t\in I_e^*(\tau)} F_{e,t}(\tau). \tag{15}$$

Both statistics are set to zero if $I_e^*(\tau)$ is empty.

Let $\mathcal{C}^{\mathrm{gt}} = \mathcal{C}(\tau^{\mathrm{gt}})$ denote the persistent contact edges of the reference rollout. The edge-wise duration error is

$$E_{\mathrm{dur}}(\tau) = \frac{1}{|\mathcal{C}^{\mathrm{gt}}|} \sum_{e\in\mathcal{C}^{\mathrm{gt}}} d_{\log}\left(D_e(\tau), D_e(\tau^{\mathrm{gt}}); \epsilon_d\right). \tag{16}$$

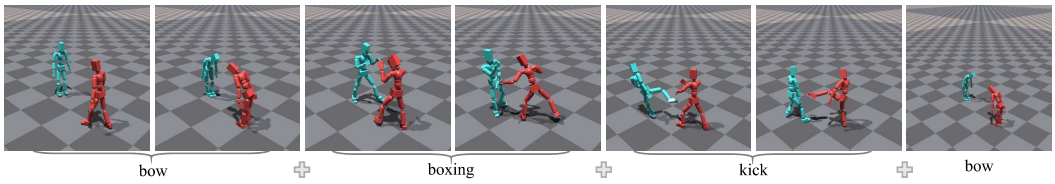



Figure 10: Additional long-horizon humanoid interaction result.

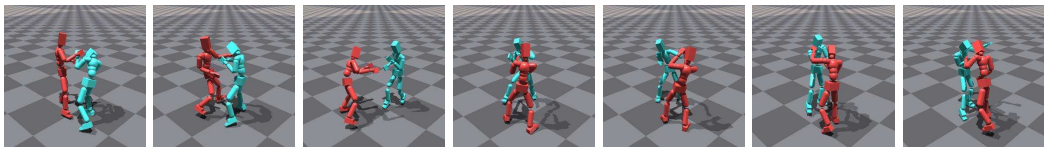


Figure 11: An appealing humanoid interaction with sustained contact (boxing). Please refer to the supplementary video.

The edge-wise force error is

$$E_{\text{force}}(\tau) = \frac{1}{2|\mathcal{C}^{\text{gt}}|} \sum_{e \in \mathcal{C}^{\text{gt}}} \left[ d_{\log}\left(\bar{F}_e(\tau), \bar{F}_e(\tau^{\text{gt}}); \epsilon_f\right) + d_{\log}\left(F_e^{\max}(\tau), F_e^{\max}(\tau^{\text{gt}}); \epsilon_f\right) \right]. \quad (17)$$

Extra contact edges generated by the candidate are penalized separately by the topology error. Candidates that fall during simulation are excluded from all preference sets except the trk preference set

### D.3 Diffusion Parameterization

Given a clean two-person interaction $x_0 = (x_0^A, x_0^B)$, the forward process samples

$$q(x_t \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0, (1 - \bar{\alpha}_t) I\big). \quad (18)$$

Both generator backbones predict the clean interaction, $\hat{x}_0 = G_\theta(x_t^A, x_t^B, t, c)$, and iterative reverse diffusion defines $p_\theta(x_0 \mid c)$. For a DPO pair, the preferred and rejected motions share the sampled timestep $t$ and are independently perturbed as

$$x_t^{\pm} = \sqrt{\bar{\alpha}_t}\, x^{\pm} + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon^{\pm}, \qquad \varepsilon^{\pm} \sim \mathcal{N}(0, I). \quad (19)$$

## E Full Metrics Definitions

### E.1 Motion Quality.

We follow the evaluation protocols established in prior work Liang et al. (2024); Li et al. (2026), and train evaluators as described in Appendix J to assess the alignment between the generated stylized motions and the input content text. Specifically, we use Frechet Inception Distance (FID) to assess the visual quality of generated motions, Retrieval Precision (R-Precision) and Multi-Modal Distance (MM-Dist) to evaluate the alignment between motions and the input text.

**FID**. Fréchet Inception Distance (FID) is employed to evaluate the quality of generated motions by comparing the distribution of motion features between generated and real motion data. Lower FID values indicate that the generated motions are closer to real samples in terms of visual realism and temporal coherence. We extract motion features using the same motion encoder trained in App. J. We obtain a set of motion features $X$ from the generated motion data, and another set of features $X'$ from the test dataset. The formulation of FID is defined as follows:

$$\text{FID} = ||\mu_1 - \mu_2||^2 - \text{Trace}(\sigma_1 + \sigma_2 - 2(\sigma_1\sigma_2)^{\frac{1}{2}}), \quad (20)$$

where $\mu_1$ and $\mu_2$ represent the mean feature vectors of $X$ and $X'$, respectively, while $\sigma_1$ and $\sigma_2$ denote their corresponding covariance matrices.

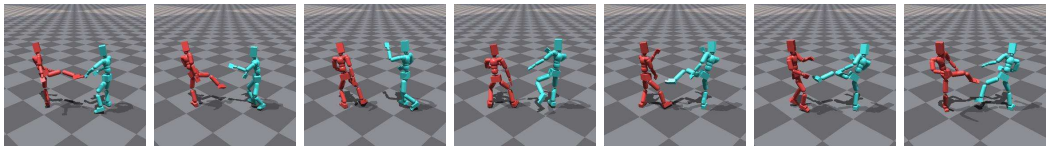


Figure 12: An appealing humanoid interaction result featuring a highly dynamic motion (kick).

**MM-Dist**. Multi-Modal Distance (MM-Dist) quantifies the semantic distance between the text and the generated motion in a multi-modal embedding space. Lower MM-Dist values imply better alignment between the motion and the content text prompt. We use the same evaluator to extract the text feature $t_i$ and motion feature $m_i$ for all the text prompts and generated motion results. The formulation of MM-Dist as follows:

$$\text{MM Dist} = \frac{1}{N_m} \sum_{i=1}^{N_m} ||t_i - m_i||, \tag{21}$$

where $N_m$ is the total number of generated motions.

**R-Precision.** R-Precision is employed to evaluate the semantic alignment between generated motions and their corresponding textual descriptions. For each generated motion, we form a candidate pool that includes its paired textual description along with 95 distractor descriptions randomly sampled from the test set. Both the motion and textual descriptions are embedded into a shared feature space, and their pairwise cosine distances are computed. We rank the candidate descriptions according to these distances, and a retrieval is counted as correct if the paired description is found within the top-$k$ positions.

### E.2 Physical Quality.

**Float**. Float measures the vertical clearance between the executed human motion and the ground plane. A large Float value indicates that the character tends to hover above the ground, while a lower value indicates better ground alignment. We compute this metric for digital generated motions. For each frame, we select the lowest vertical position among the 22 body joints, excluding the two palm joints, and clamp below-ground values to zero. The Float metric is defined as:

$$h_{i,t,a} = \max\left(\min_{j \in \mathcal{J}_{22}} z_{i,t,a,j}, 0\right), \tag{22}$$

$$\text{Float} = \frac{1000}{N_m} \sum_{i=1}^{N_m} \frac{1}{2T_i} \sum_{a=1}^{2} \sum_{t=1}^{T_i} h_{i,t,a}, \tag{23}$$

where $z_{i,t,a,j}$ denotes the vertical position, in meters, of body joint $j$ of person $a$ at frame $t$ in motion $i$. $\mathcal{J}_{22}$ denotes the set of 22 body joints excluding the two palm joints, $T_i$ denotes the valid length of motion $i$, and $N_m$ is the total number of evaluated motions. The factor of 1000 converts the result from meters to millimeters. Below-ground positions do not produce negative Float values and are evaluated separately by the interpenetration metric. Lower Float values indicate better physical plausibility.

**Skate.** Skate measures the horizontal displacement of the feet while they are close to the ground. A physically plausible motion should exhibit little horizontal foot movement during ground contact, whereas large values indicate undesirable foot sliding. Let $\mathcal{F}$ denote the four foot bodies, including the left and right ankles and toes. Here, $i$ indexes the interaction sequence, $a$ indexes the person in sequence $i$, $t$ indexes the frame, and $T_i$ denotes the number of frames in sequence $i$. A foot is considered close to the ground if its vertical position is below 5 cm:

$$g_{i,t,a,f} = \mathbb{I}\left[z_{i,t,a,f} < 0.05\right], \tag{24}$$

where $f \in \mathcal{F}$ indexes the foot body, $z_{i,t,a,f}$ denotes its vertical position in meters, and $\mathbb{I}[\cdot]$ is the indicator function. The horizontal displacement of foot $f$ between two consecutive frames is

$$d_{i,t,a,f} = \left\| \mathbf{p}^{xy}_{i,t+1,a,f} - \mathbf{p}^{xy}_{i,t,a,f} \right\|_2. \tag{25}$$

where $\mathbf{p}^{xy}_{i,t,a,f} \in \mathbb{R}^2$ denotes the horizontal position of foot body $f$ belonging to person $a$ at frame $t$ of sequence $i$. We first compute the average grounded-foot displacement for each person:

$$s_{i,a} = \frac{\sum_{t=1}^{T_i-1} \sum_{f \in \mathcal{F}} g_{i,t,a,f}\, d_{i,t,a,f}}{\max\left(1, \sum_{t=1}^{T_i-1} \sum_{f \in \mathcal{F}} g_{i,t,a,f}\right)}. \tag{26}$$

where $s_{i,a}$ denotes the average horizontal displacement of the grounded foot bodies of person $a$ in sequence $i$. The maximum operation in the denominator prevents division by zero when no foot body is detected near the ground. The metric is reported in millimeters; therefore, $s_{i,a}$ is multiplied by $1000$ when the foot positions are represented in meters. No division by the frame interval is applied. Lower Skate values indicate less foot sliding and better contact consistency.

**Interpen.** Interpenetration measures the geometric overlap between the two interaction. Each body joint is approximated by a sphere with radius $r_b = 0.06$ m. For every frame, we compute all pairwise distances between the body joints of the two people:

$$D_{i,t,j,k} = \left\| \mathbf{p}^{\text{gen}}_{i,t,1,j} - \mathbf{p}^{\text{gen}}_{i,t,2,k} \right\|_2, \tag{27}$$

where $\mathbf{p}^{\text{gen}}_{i,t,a,j} \in \mathbb{R}^3$ denotes the generated position of joint $j$ of person $a$ at frame $t$ in motion $i$. The minimum cross-person joint distance at frame $t$ is:

$$d^{\min}_{i,t} = \min_{j \in \mathcal{J}_{24},\, k \in \mathcal{J}_{24}} D_{i,t,j,k}, \tag{28}$$

where $\mathcal{J}_{24}$ denotes the set of 24 body joints. The frame-level interpenetration proxy is defined as:

$$I_{i,t} = \max\left(2r_b - d^{\min}_{i,t}, 0\right) = \max\left(0.12 - d^{\min}_{i,t}, 0\right). \tag{29}$$

The motion-level Interpenetration is computed by averaging over all valid frames:

$$I_i = \frac{1000}{T_i} \sum_{t=1}^{T_i} I_{i,t}, \tag{30}$$

where $T_i$ is the valid length of motion $i$, and the factor of $1000$ converts meters to millimeters. The final Interpenetration metric is obtained using a motion-equal average:

$$\text{Interpen} = \frac{1}{N_m} \sum_{i=1}^{N_m} I_i, \tag{31}$$

where $N_m$ is the total number of generated motions. Lower Interpenetration values indicate less cross-person body overlap and better kinematic plausibility.

**Jerk.** Jerk measures the temporal smoothness of the motion by quantifying the rate of change of body acceleration. A large Jerk value indicates abrupt changes in body movement and physically implausible high-frequency oscillations, while a lower value indicates smoother and more natural motion. Let $\mathbf{p}_{i,t,a,j} \in \mathbb{R}^3$ denote the position, in meters, of body joint $j$ of person $a$ at frame $t$ in motion $i$. The third-order finite difference of the body position is defined as:

$$\Delta^3 \mathbf{p}_{i,t,a,j} = \mathbf{p}_{i,t+3,a,j} - 3\mathbf{p}_{i,t+2,a,j} + 3\mathbf{p}_{i,t+1,a,j} - \mathbf{p}_{i,t,a,j}. \tag{32}$$

Given a frame rate of $f$ Hz, the discrete jerk vector is computed as:

$$\mathbf{j}_{i,t,a,j} = f^3 \Delta^3 \mathbf{p}_{i,t,a,j}. \tag{33}$$

We compute the motion-level Jerk by averaging the jerk magnitude over all valid temporal positions, both people, and all 24 rigid bodies:

$$J_i = \frac{1}{2N_J(T_i-3)} \sum_{a=1}^{2} \sum_{j=1}^{N_J} \sum_{t=1}^{T_i-3} \left\| \mathbf{j}_{i,t,a,j} \right\|_2, \tag{34}$$

where $N_J = 24$ is the number of rigid bodies and $T_i$ is the valid executed length of motion $i$. The final Jerk metric is obtained using a motion-equal average:

$$\text{Jerk} = \frac{1}{N_m} \sum_{i=1}^{N_m} J_i, \tag{35}$$

where $N_m$ is the total number of evaluated motions. The metric is reported in $\mathrm{m/s^3}$, and lower Jerk values indicate smoother and more physically plausible motion.

**Force Error.** Force Error evaluates whether the intensity of inter-person physical contacts in the generated motions is consistent with that of the corresponding ground-truth motions. We obtain exact rigid-body contacts from the PhysX simulator after each physics step. For each contact point $c$ associated with an inter-person body pair $e$, PhysX provides a normal contact magnitude $\lambda_c$ and a world-space contact normal $\mathbf{n}_c$. All contact normals are consistently oriented from the first person to the second person. The resultant normal-contact magnitude for body pair $e$ at physics query $q$ is:

$$r_{i,q,e} = \left\| \sum_{c \in \mathcal{C}_{i,q,e}} \lambda_c \mathbf{n}_c \right\|_2, \tag{36}$$

where $\mathcal{C}_{i,q,e}$ is the set of contact manifold points belonging to body pair $e$. A body pair is considered contact-active when its resultant magnitude exceeds the contact threshold $\tau = 1.0$. The total contact force at physics query $q$ is:

$$L_{i,q} = \sum_{e} \mathbb{I}\left[r_{i,q,e} > \tau\right] r_{i,q,e}. \tag{37}$$

Let $\mathcal{Q}_i^+ = \{q \mid L_{i,q} > 0\}$ denote the set of contact-active physics queries. The mean contact force of motion $i$ is:

$$\mu_i = \frac{1}{|\mathcal{Q}_i^+|} \sum_{q \in \mathcal{Q}_i^+} L_{i,q}. \tag{38}$$

We compare the generated and ground-truth mean force to calculate the force error:

$$e_i^{\text{force}} = \left| \mu_i^{\text{gen}} - \mu_i^{\text{GT}} \right|. \tag{39}$$

For multiple generated motions associated with the same text prompt, we first average the errors within each prompt and then average equally over all prompts:

$$\text{Force Error} = \frac{1}{N_p} \sum_{p=1}^{N_p} \frac{1}{|\mathcal{S}_p|} \sum_{i \in \mathcal{S}_p} e_i^{\text{force}}, \tag{40}$$

where $\mathcal{S}_p$ denotes the set of valid generated motions associated with prompt $p$, and $N_p$ is the number of evaluated prompts. The final Force Error is measured in Newtons (N), and lower values indicate that the generated contact intensity is closer to that of the tracker-executed ground-truth motion.

**Success Rate (Succ.).** Success Rate (Succ.) measures the fraction of generated two-person motion sequences that can be tracked with sufficiently low average error while remaining stable over the entire sequence. For each generated interaction sequence $i$, we compute the sequence-averaged mean per-joint tracking error of person $p$ as

$$d_{i,p} = \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{1}{J} \sum_{j=1}^{J} \left\| \mathbf{x}_{i,p,t,j} - \mathbf{x}_{i,p,t,j}^{\text{ref}} \right\|_2, \tag{41}$$

where $\mathbf{x}_{i,p,t,j}$ and $\mathbf{x}_{i,p,t,j}^{\text{ref}}$ denote the global positions of the $j$-th body joint in the simulated and reference motions, respectively. Here, $p \in \{1, 2\}$ denotes the person index, $T_i$ is the number of evaluated frames in the $i$-th motion pair, and $J = 24$ is the number of body joints.

A motion pair is considered successful if the sequence-averaged tracking error of each person does not exceed the tracking threshold $\tau_d$ and neither person's pelvis height falls below the stability threshold $\tau_h$ at any frame:

$$s_i = \mathbb{I}\left[ \max_{p \in \{1,2\}} d_{i,p} \le \tau_d \ \wedge \ \min_{\substack{p \in \{1,2\} \\ 1 \le t \le T_i}} z_{i,p,t}^{\text{pelvis}} \ge \tau_h \right], \tag{42}$$

where $\mathbb{I}[\cdot]$ is the indicator function. The Success Rate is then computed as

$$\text{Succ.} = \frac{1}{N_m} \sum_{i=1}^{N_m} s_i, \tag{43}$$

where $N_m$ is the total number of evaluated two-person motion pairs. Higher Succ. values indicate better full-sequence tracking performance. In our evaluation, we set $\tau_d = 0.25\,\text{m}$ and $\tau_h = 0.3\,\text{m}$.

**MBD.** Mean Body Deviation (MBD) measures the average global positional discrepancy between the simulated and reference body motions. For each person in a motion pair, we first average the body-joint distances over all joints and all frames of that motion. We then average the resulting motion-level distances over both persons and all evaluated motion pairs. The formulation of MBD is as follows:

$$\text{MBD} = \frac{1}{N_m} \sum_{i=1}^{N_m} \left[ \frac{1}{2} \sum_{p=1}^{2} \left( \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{1}{J} \sum_{j=1}^{J} \left\| \mathbf{x}_{i,p,t,j} - \mathbf{x}^{\text{ref}}_{i,p,t,j} \right\|_2 \right) \right], \tag{44}$$

where $N_m$ is the total number of evaluated two-person motion pairs, $T_i$ is the motion length, and $J = 24$ is the number of body joints. Lower MBD values indicate more accurate motion tracking.

We adopt motion-level averaging, such that each motion pair contributes equally to the final MBD regardless of its sequence length.

## F More Details of Datasets

Our experiments are conducted on two datasets: InterHuman Liang et al. (2024) and Inter-X Xu et al. (2024).

**InterHuman.** InterHuman is a standard benchmark for text-conditioned two-person motion generation. It contains 6,022 interaction sequences paired with 16,756 natural-language descriptions, amounting to 6.56 hours of motion. Each sequence represents the coordinated motions of two participants in a shared coordinate frame, enabling the evaluation of both individual motion quality and inter-person coordination.

**Inter-X.** Inter-X is a large-scale dataset for versatile human–human interaction analysis. It contains 11,388 SMPL-X interaction sequences and 34,164 natural-language descriptions, spanning 40 interaction categories and more than 8.1 million motion frames. Compared with InterHuman, Inter-X provides broader interaction coverage and more detailed body and hand motions, offering a complementary benchmark for evaluating generalization across interaction distributions.

For both datasets, we follow the official train, validation, and test splits and the preprocessing protocols of the corresponding interaction-generation backbones. To ensure a consistent embodiment between motion generation and humanoid tracking, we retarget all interaction sequences to the neutral SMPL body with zero shape parameters. The retarget details are shown in Appendix H.

## G Annotated Contact on InterHuman Testset

InterHuman provides paired motions and natural-language descriptions, but it does not specify which cross-person body contacts are semantically required by each description. To support contact-sensitive evaluation, we construct a structured annotation subset of the official InterHuman test set. The annotations describe the contact topology entailed by a caption which are not frame-level contact labels extracted from the ground-truth motion. This distinction is important because the evaluation asks whether a generated physical interaction realizes the contact semantics expressed in the text, while allowing different valid motion realizations.

We process the captions associated with valid motions in the official test split. Motion arrays are used only to establish test-set membership and validity and are never provided to the annotation model.

**Structured Contact Specification.** Each caption is mapped to a typed contact specification over the 24 rigid bodies of the simulated humanoid. In addition to exact body names, the ontology provides semantic selectors such as `Hand`, `Arm`, `Foot`, `Leg`, `Trunk`, and `Any_Body`, each of which expands deterministically to a set of exact bodies. A specification contains: (i) *required* contact clauses that must be realized; (ii) *allowed* contacts that are explicitly compatible with the caption but not mandatory; and (iii) *forbidden* contacts explicitly ruled out by the caption. Each required clause may contain multiple `any_of` alternatives, allowing several body-pair realizations of the same semantic contact, together with a `distinct_count` when the caption requires multiple distinct contacts. This representation avoids treating one arbitrarily selected body pair as the unique realization of an ambiguous instruction.

The annotation contract is deliberately conservative. It requires evidence quoted verbatim from the caption and prohibits inferring contact merely from proximity, synchronized motion, reaching, attempted actions, or object transfer. It also records controlled ambiguity flags, an abstention decision, a confidence score, and whether contacts not mentioned by the caption should be ignored or treated as disallowed. Participant labels are directional, but the complete specification is canonicalized under one globally consistent $A/B$ transpose; individual contact edges cannot be swapped independently.

**Multi-pass Annotation and Strict Filtering.** We annotate each exact caption in two separate text-only passes using GPT-5.6-sol. The first pass proposes a contact specification, whereas the second uses a skeptical prompt designed to reject contacts that are plausible in a typical execution but are not entailed by the text. Both outputs must satisfy the same JSON schema and are additionally checked by a local semantic validator. The validator verifies the selector vocabulary, required/allowed/forbidden consistency, valid multiplicities, confidence range, and exact matching between every evidence span and the input caption. A caption is included in the final subset only when the two passes agree and satisfy our confidence and quality requirements. Selected disagreement or high-risk cases undergo blind third-pass adjudication and are included only when consensus is reached under the same strict criteria. The final strict subset contains 503 contact-needed captions from 223 motions and specifies 631 required contact slots. It comprises 445 annotations accepted by two-pass consensus and 58 additional annotations resolved by the blind third pass. We further manually reviewed all annotations in the final strict subset to ensure their accuracy. All remaining captions are withheld from strict contact evaluation.

**Core Annotation Prompt.** The complete prompt supplies the body ontology and output schema. Its central instruction is:

> *Given exactly one interaction caption, identify only cross-person body contacts that are explicitly entailed by the text. Do not infer contact from a typical execution, proximity, synchronized motion, reaching, attempted actions, or object transfer. Represent mandatory contacts as required clauses; use* `any_of` *when the caption permits multiple body-pair realizations and* `distinct_count` *when it explicitly requires multiple distinct contacts. Add allowed or forbidden contacts only when supported by the caption. If the contact cannot be determined safely, abstain. For every non-abstaining decision, quote an exact supporting span from the caption and return the result using the provided schema.*

**Evaluation Metrics.** The annotations remain in the 24-body ontology to preserve left/right and fine-grained body identities. For the evaluation in Tab. 5, Topology Recall is the fraction of required slots matched, while Topology Precision measures the fraction of evaluated predicted region pairs that are required or allowed by the specification. Contact success indicates whether the rollout completes without humanoid falling and the mean body deviation is less than $\tau_d$.

## H MOTION RETARGET

The source interaction data and the simulated humanoids have different body proportions and kinematic structures. We first use General Motion Retargeting (GMR) (Araujo et al., 2025) to map each participant to the target humanoid. The two participants are retargeted in the same world coordinate frame so that their relative placement and timing are retained. Denote the resulting target configurations by $\widetilde{q}_t^p$, where $p \in \{A, B\}$ and $t \in \{1, \ldots, T\}$. Although GMR provides a strong whole-body

initialization, morphology differences can still perturb fine-grained inter-person contacts at the limb end effectors. An intended hand or foot contact in the source motion may become separated after retargeting or displaced to a nearby joint on the target humanoid.

**Source End-effector Contacts.** We infer the intended end-effector contacts before retargeting, rather than treating every pair of nearby joints as a contact constraint. Let $j_{t,i}^{p,\mathrm{src}} \in \mathbb{R}^3$ denote the position of joint $i$ for participant $p \in \{A, B\}$ in the source motion. We define the limb end-effector set $\mathcal{J}_{\mathrm{ee}}$, containing the left and right hand and foot joints, together with a selected set of potential counterpart joints $\mathcal{J}_{\mathrm{c}}$. Candidate pairs are restricted to

$$\mathcal{P}_{\mathrm{ee}} = (\mathcal{J}_{\mathrm{ee}} \times \mathcal{J}_{\mathrm{c}}) \cup (\mathcal{J}_{\mathrm{c}} \times \mathcal{J}_{\mathrm{ee}}), \tag{45}$$

where the first and second elements index participants $A$ and $B$, respectively. Thus, every optimized contact involves at least one limb end effector. For each $(i, j) \in \mathcal{P}_{\mathrm{ee}}$, we compute

$$d_{t,i,j}^{\mathrm{src}} = \left\| j_{t,i}^{A,\mathrm{src}} - j_{t,j}^{B,\mathrm{src}} \right\|_2 . \tag{46}$$

A candidate pair is considered to be in contact whenever its distance is below the source detection threshold $\tau_{\mathrm{src}}$:

$$\mathcal{C}_t^{\mathrm{src}} = \left\{ (i, j) \in \mathcal{P}_{\mathrm{ee}} \;\middle|\; d_{t,i,j}^{\mathrm{src}} < \tau_{\mathrm{src}} \right\}. \tag{47}$$

Consecutive detections of the same joint pair are grouped into contact episodes, and isolated detections shorter than a minimum duration are discarded. Restricting the candidate set is important for interactions such as hugging, in which many upper-body joints may be spatially close but should not all be pulled together. We retain the selected joint identities throughout retargeting using the source-to-target joint correspondence provided to GMR. Consequently, a selected hand or foot contact must be restored using the corresponding end-effector joint and cannot be replaced by bringing a nearby wrist, forearm, ankle, or lower-leg joint close to the other participant.

**Contact-aware Local Refinement.** Starting from the GMR solution, we jointly refine the two target configurations using forward kinematics. Let $\phi(i)$ denote the target joint corresponding to source joint $i$, and let $J_{\phi(i)}^{p}(q_t^p)$ be its world position under target configuration $q_t^p$. For each retained source contact $(i, j)$, the corresponding target-joint distance is

$$d_{t,i,j}^{\mathrm{tar}} = \left\| J_{\phi(i)}^{A}(q_t^A) - J_{\phi(j)}^{B}(q_t^B) \right\|_2 . \tag{48}$$

We use a target tolerance $\tau_{\mathrm{tar}}$, separate from the more permissive source detection threshold $\tau_{\mathrm{src}}$, and minimize

$$\mathcal{L}_{\mathrm{contact}} = \sum_t \sum_{(i,j) \in \mathcal{C}_t^{\mathrm{src}}} \left[ \max\left( d_{t,i,j}^{\mathrm{tar}} - \tau_{\mathrm{tar}}, 0 \right) \right]^2 . \tag{49}$$

Because $\mathcal{C}_t^{\mathrm{src}}$ contains only end-effector-involving pairs, the refinement targets hand and foot contacts rather than indiscriminately attracting all nearby body joints. The joint correspondence $\phi$ also prevents a selected contact from being satisfied by switching the optimized target from the designated end effector to a nearby joint. The one-sided loss restores a missing contact but applies no additional attraction after the designated joints are sufficiently close.

For efficiency and locality, we optimize only temporally padded windows around the detected contact episodes. For an end-effector–body contact, we update the kinematic chain of the contacting limb while the counterpart body joint remains on its GMR trajectory. This corresponds primarily to the shoulder–elbow–wrist–hand chain for a hand contact and the hip–knee–ankle–foot chain for a foot contact. When both members of a pair are end effectors, we update the corresponding limb chains of both participants. All unrelated joint trajectories remain fixed to the GMR result. Let $\mathcal{Q}_\Omega$ collect these selected joint variables within an optimized window $\Omega$. We regularize them toward the GMR solution and preserve its temporal changes:

$$\mathcal{L}_{\mathrm{pose}} = \sum_{p \in \{A,B\}} \sum_{t \in \Omega} \| q_t^p - \widetilde{q}_t^p \|_{W_q}^2 , \tag{50}$$

$$\mathcal{L}_{\mathrm{temp}} = \sum_{p \in \{A,B\}} \sum_{t, t-1 \in \Omega} \left\| (q_t^p - q_{t-1}^p) - (\widetilde{q}_t^p - \widetilde{q}_{t-1}^p) \right\|_2^2 . \tag{51}$$

The root and joints outside the active limb chains receive substantially larger entries in $W_q$, preventing the optimizer from recovering a local end-effector contact by translating the full character or disturbing unrelated parts of its support configuration. The configurations at the boundaries of each padded window are fixed to the GMR trajectory. The final refinement solves

$$\min_{\mathcal{Q}_\Omega} \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}, \tag{52}$$

subject to the target humanoid's joint limits. Because the source contact pairs are detected only once, each optimization step evaluates only the forward-kinematic positions of a small set of target joints. We use the refined trajectories as the unified references for training and evaluating the interaction tracking policy.

# I TRACKING POLICY

**Humanoid Control and Observation.** Following PHC Luo et al. (2023), we represent each participant by a 24-body neutral-SMPL humanoid with 23 actuated non-root joints. The policy operates at 30 Hz and outputs a 69-dimensional action specifying three-dimensional target rotations for these joints; the simulator runs at 60 Hz and applies the targets through PD controllers. The normalized policy output $a_t^p$ for participant $p$ is converted to a generalized-position target as

$$q_{t,\text{PD}}^p = q_{\text{offset}} + q_{\text{scale}} \odot a_t^p, \tag{53}$$

where the offset and scale are determined by the joint limits.

For each participant, the observation contains both self and partner information. The current-state descriptor includes root height and the heading-local body positions, six-dimensional rotations, linear velocities, and angular velocities. The reference descriptor encodes the corresponding differences from the next reference frame, together with its body poses in the same local representation. Denoting the other participant by $\bar{p}$, the policy is conditioned on

$$o_t^p = \left[o_{t,\text{self}}^p, o_{t,\text{ref}}^p, o_{t,\text{self}}^{\bar{p}}, o_{t,\text{ref}}^{\bar{p}}\right]. \tag{54}$$

Conditioning on the complete partner observation allows the tracker to respond to the partner's realized state rather than independently tracking the two reference motions.

**Cross-attention Policy.** The self and partner descriptors are processed by separate MLP encoders, producing features $f_t^p, f_t^{\bar{p}} \in \mathbb{R}^{512}$. Each feature is reshaped into four 128-dimensional tokens. We use the self tokens as queries and the partner tokens as keys and values in a four-head cross-attention layer:

$$h_t^p = \text{LN}\big(f_t^p + \text{vec}\big[\text{MHA}\left(T_t^p, T_t^{\bar{p}}, T_t^{\bar{p}}\right)\big]\big). \tag{55}$$

The residual feature $h_t^p$ is passed to an MLP policy head that predicts the action distribution, and we use its mean action for deterministic rollout. The critic uses the same interaction-aware encoding structure. In contrast to independently controlled trackers, this design lets each humanoid modulate its action according to the current and desired motion of its partner.

**Interaction-aware Reward.** The base imitation objective follows PHC and measures errors in body positions, rotations, linear velocities, and angular velocities. Writing these errors as $E_p, E_R, E_v$, and $E_\omega$, respectively, the imitation reward is

$$r_{\text{imit}} = 0.5e^{-100E_p} + 0.3e^{-10E_R} + 0.1e^{-0.1E_v} + 0.1e^{-0.1E_\omega}. \tag{56}$$

We further introduce an interaction-graph reward Zhang et al. (2023); Li et al. (2026) to preserve the spatial and dynamic relationships between the two humanoids. For every cross-person body pair $(i, j)$, let

$$d_{ij} = \left\|p_i^A - p_j^B\right\|_2, \qquad u_{ij} = v_j^B - v_i^A \tag{57}$$

be its distance and relative velocity, with $\bar{d}_{ij}$ and $\bar{u}_{ij}$ denoting the corresponding reference quantities. Using proximity-aware edge weights $\alpha_{ij}$, we compute

$$E_{\mathrm{G}}^{p} = \sum_{i,j} \alpha_{ij} (d_{ij} - \bar{d}_{ij})^2, \tag{58}$$

$$E_{\mathrm{G}}^{v} = \sum_{i,j} \alpha_{ij} \left\| u_{ij} - \bar{u}_{ij} \right\|_2^2, \tag{59}$$

$$r_{\mathrm{G}} = 0.8 e^{-20 E_{\mathrm{G}}^{p}} + 0.2 e^{-0.5 E_{\mathrm{G}}^{v}}. \tag{60}$$

The task reward combines individual motion tracking and interaction preservation as

$$r_{\mathrm{task}} = 0.6 \left( r_{\mathrm{imit}} + r_{\mathrm{power}} \right) + 0.4 r_{\mathrm{G}}, \tag{61}$$

where $r_{\mathrm{power}}$ penalizes excessive joint power. Inter-person collision is enabled during training and execution. We optimize the actor and critic using PPO with the adversarial motion-prior objective retained from PHC.

## J TEXT-MOTION EVALUATORS

We train separate text-motion evaluators InterCLIP and InterCLIP-Phys for the digital motion domain and the physically executed humanoid domain, respectively. InterCLIP evaluates generated digital motion before execution and is trained on the InterHuman and Inter-X training sets using source-balanced sampling. Following the standard interaction representation, it encodes joint positions, frame displacements, and local joint rotations for both participants, while excluding the foot-contact indicators. InterCLIP-Phys evaluates the tracker-executed trajectories. The metrics related to text-motion consistency in Tab. 2 are computed by it. It encodes the positions, orientations, linear velocities, and angular velocities of the 24 rigid bodies of both humanoids, expressed in an actor-centric heading-local frame. Both evaluators adopt the TMR-style retrieval architecture Petrovich et al. (2023). A frozen CLIP ViT-L/14 text tower followed by a two-layer trainable adapter encodes captions, while an eight-layer temporal Transformer encodes motion sequences. The two branches produce $\ell_2$-normalized 512-dimensional embeddings. Let $z_i^m$ denote the embedding of motion $i$, $z_j^t$ the embedding of caption $j$, and $P_i$ the captions associated with motion $i$. Their similarity is

$$s_{ij} = \frac{(z_i^m)^\top z_j^t}{\tau}, \tag{62}$$

where $\tau$ is a temperature hyperparameter that controls the concentration of the text–motion similarity distribution. We optimize the symmetric multi-positive contrastive objective

$$\mathcal{L}_{m \to t} = -\frac{1}{B} \sum_{i=1}^{B} \frac{1}{|P_i|} \sum_{j \in P_i} \log \frac{\exp(s_{ij})}{\sum_{k \in V_i} \exp(s_{ik})}, \tag{63}$$

$$\mathcal{L}_{\mathrm{eval}} = \frac{1}{2} \left( \mathcal{L}_{m \to t} + \mathcal{L}_{t \to m} \right), \tag{64}$$

where $V_i$ is the set of valid caption candidates and $\mathcal{L}_{t \to m}$ is defined analogously. All captions describing the same motion are treated as positives. Identical normalized captions associated with different motion identities are removed from the negative set to avoid explicit false negatives.

**Physical Evaluator Data.** Following the evaluation protocol of InterAgent Li et al. (2026), InterCLIP-Phys is trained exclusively on physical executions of InterHuman training motions. We remove duplicated two-person motions across splits, including participant-swapped duplicates, and execute the remaining reference interactions with a frozen two-person tracking policy. To ensure a fair comparison, this policy is distinct from the pretrained tracker used by DIGHT, preventing the evaluator construction from favoring our method. We retain complete, finite, non-falling rollouts with mean body tracking distance below $0.25$ m. During training, we sample motion identities uniformly, select one accepted execution, apply temporal cropping, and randomly exchange the two participants.