InterAgent: Physics-based Multi-agent Command Execution via Diffusion on Interaction Graphs

Supplementary Material

8. Demo Video

Beyond the qualitative snapshots, we provide a demo video (attached in the zip) offering more detailed visualizations, further showcasing the effectiveness of our approach.

9. Additional Implementation Details

Interaction tracking policy. To ensure the collected data accurately captures the ongoing interactions among agents, we integrate the interaction graph reward [101] into the training of the tracking policy. Specifically, during the data collection phase, we define each edge of the interaction graph as $\boldsymbol{e}_{ij} = (\boldsymbol{p}_{ij}, \boldsymbol{v}_{ij})$, where $\boldsymbol{p}_{ij}$ and $\boldsymbol{v}_{ij}$ denote the relative position and relative velocity between joint i and joint j, respectively. Based on this edge representation, we further define the discrepancy for each pair of edges as:

$$d_{pos} = \sum_{i,j \in E} \omega_{ij} ||\boldsymbol{p}_{ij}^{ref} - \boldsymbol{p}_{ij}^{sim}||, \tag{7}$$

$$d_{vel} = \sum_{i,j \in E} \gamma_{ij} ||\boldsymbol{v}_{ij}^{ref} - \boldsymbol{v}_{ij}^{sim}||, \tag{8}$$

where the superscripts ref and sim correspond to the reference motion and simulation motion, ω_{ij} and γ_{ij} are edge weighting functions detailed in Zhang et al. [101]. Accordingly, we formulate the interaction graph reward as follows:

$$r_{ig} = \exp\left(-\lambda_{pos} * d_{pos} - \lambda_{vel} * d_{vel}\right), \tag{9}$$

The root reward is defined as:

$$r_{root} = \exp\left(-\lambda_{root} * ||\boldsymbol{x}_{root}^{ref} - \boldsymbol{x}_{root}^{sim}||\right), \tag{10}$$

where the λ_{pos}, λ_{vel} and λ_{root} denote the term sensitivities, $\boldsymbol{x}_{root}$ represents the position, velocity, orientation and angular velocity of the root joint. Therefore, our reward function is then calculated as:

$$r = r_{ig} * r_{root}. \tag{11}$$

Reactive humanoid control. To enable reactive behaviors, we embed an inpainting mechanism within the inference process. For a given command $\boldsymbol{c}$, we fix the behaviors of a humanoid via replaying its ground truth. At each denoising step t, upon the model predicting the clean behaviors $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_{p_1}, \hat{\boldsymbol{x}}_{e_1}, \hat{\boldsymbol{x}}_{a_1}, \hat{\boldsymbol{x}}_{p_2}, \hat{\boldsymbol{x}}_{e_2}, \hat{\boldsymbol{x}}_{a_2}]$, we substitute the generated proprioception of the fixed humanoid with its ground truth equivalents. For concreteness, we assume this substitution targets $\hat{\boldsymbol{x}}_{p_1}$ here, yielding $\hat{\boldsymbol{X}}' = [\boldsymbol{x}_{p_1}, \hat{\boldsymbol{x}}_{e_1}, \hat{\boldsymbol{x}}_{a_1}, \hat{\boldsymbol{x}}_{p_2}, \hat{\boldsymbol{x}}_{e_2}, \hat{\boldsymbol{x}}_{a_2}]$, which is then used to advance to the subsequent sampling process. This ensures that the generated interaction dynamics strictly constrained by the fixed humanoid's behaviors. Empirically, we refrain from replacing generated actions with ground truth counterparts, as the intrinsic randomness of simulation environment makes it difficult for the humanoid to maintain balance when the actions in collected dataset are directly deployed.

	Floating ↓ $[mm]$	Skating ↓ $[mm]$	Jerk ↓ $[mm/frame^3]$
Phys-GT	49.92	4.07×10^{-6}	16.24
InterGen++ [32]	53.35	1.24×10^{-3}	12.56
InterMask++ [27]	151.34	4.29×10^{-4}	39.15
PDP [67]	<u>49.84</u>	7.49×10^{-4}	**1.98**
CLoSD [66]	**48.64**	$\mathbf{2.62 \times 10^{-5}}$	29.07
InterAgent (Ours)	49.85	<u>2.81×10^{-4}</u>	<u>2.69</u>

Table 4. Quantitative evaluation of physical correctness. **Bold** and <u>underline</u> indicate the best and the second best result.

10. Additional Experimental Results

In this section, we introduce experimental results that are not included in the main paper due to space limit.

Qualitative results. As shown in Fig. 8, We show additional qualitative results generated by our model given various textual commands.

Quantitative physical analysis. To evaluate the physical correctness, we analyze three widely used metrics [81, 98]: *Floating* (measuring vertical drift), *Skating* (indicating horizontal sliding), and *Jerk* (quantifying the smoothness of the motion). Penetration is not reported herein, as it is negligible across all physics-based methods. As shown in Tab. 4, InterMask++ [27] exhibits significantly poor performance in floating and jerk metrics. Such deficiencies make it inherently challenging for humanoids to maintain balance in physical environments, let alone execute complex interactive behaviors. Consequently, such physical inconsistencies directly contribute to its poor FID and R-precision scores. In contrast, Our InterAgent demonstrates competitive performance across all three physical metrics, closely approaching the ground truth (Phys-GT) and outperforming most baseline methods. It strikes a strong balance between miniming floating, controlling skating, and reducing jerk, making it robust for text-driven physics-based multi-agent humanoid control.

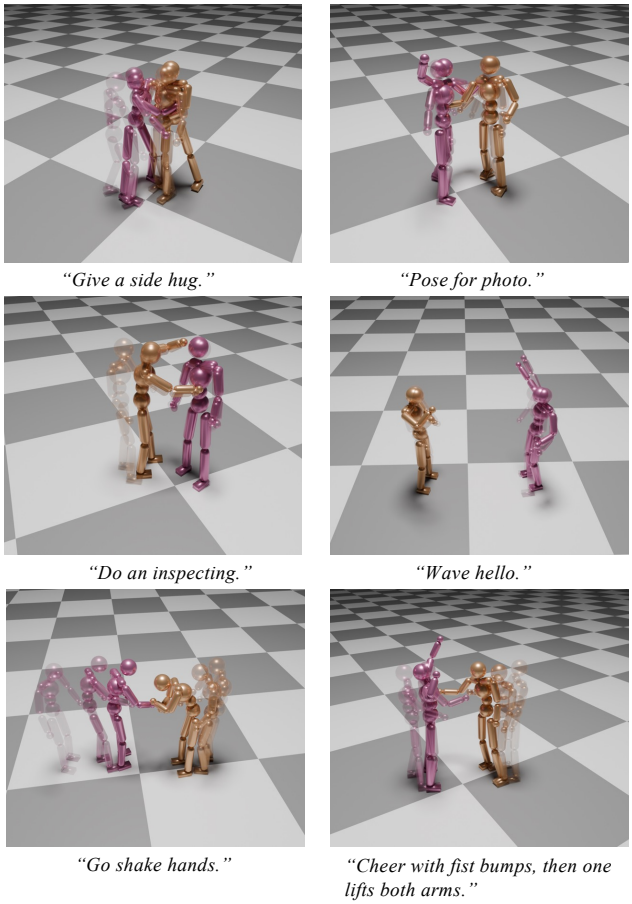

Figure 8. **Qualitative results.**

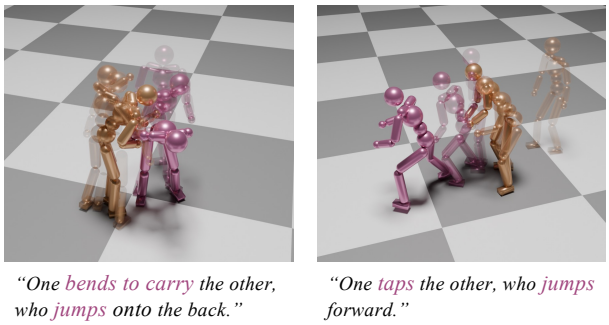

Figure 9. **Failure cases.** Our method struggles with highly dynamic behaviors like jumping.

11. Discussion

Limitations and failure cases. As shown in Fig. 9, our method performs well for most cases but still struggles with more dynamic behaviors like jumping. This issue mainly duo to model's bias toward smooth transitions, which conflicts with the significant instantaneous dynamics of jumping (explosive push-off, mid-air balance, landing impact). A potential solution could involve a dynamic physical constraint module tailored to high-dynamic behaviors. Additionally, augmenting training data with annotated high-dynamic sequences may further enhance the model's adaptability while preserving its steady-state performance.

Discussion and future work. While InterAgent demonstrates strong capability in generating coordinated, physically plausible, and text-consistent multi-agent behaviors, several aspects of the framework present valuable opportunities for further advancement. Firstly, the current text interface describes high-level motion intent but does not explicitly reason about long-horizon task structure, role assignment, or interactive strategies. Integrating a higher-level planning module, or combining large language models with our physics-based controller, could enable richer collaborative behaviors and more natural multi-turn interactions. Secondly, our model is trained with a fixed number of agents, and the computational cost grows with the number of entities due to pairwise relational modeling. Extending the framework toward scalable multi-agent coordination—possibly through hierarchical grouping, clustering of interaction patterns, or dynamic neighborhood selection—will be an essential step toward deployment in complex multi-agent interaction environments. Finally, deploying InterAgent beyond simulation remains an exciting challenge, applying our framework to real humanoid robots or VR/AR avatars will require addressing sim-to-real transfer, real-time inference, and robust perception under noisy sensory inputs. We envision InterAgent serving as a building block for future interactive embodied systems, enabling fluid multi-agent coordination in entertainment, robotics, and immersive virtual environments.