Supplementary Material for Generating Human Interaction Motions in Scenes with Text Control

Hongwei Yi[1,2], Justus Thies[2,3], Michael J. Black[2],
Xue Bin Peng[1,4], and Davis Rempe[1]

[1] NVIDIA
[2] Max Planck Institute for Intelligent Systems, Tubingen, Germany
[3] Technical University of Darmstadt
[4] Simon Fraser University

1 Implementation Details

Training. The scene-agnostic branch of our navigation model is trained on the 3D motions and text descriptions from the Loco-3D-FRONT dataset for 420k optimization steps. Subsequently, we freeze the base model weights and fine-tune the scene-aware branch, with additional 2D floor map inputs, for a further 20k steps. Similarly, the scene-agnostic base of our interaction model first trains on a mix of HumanML3D [2] and SAMP [3] data without objects for 400k steps. Then, the object-aware branch is fine-tuned on our text-annotated SAMP data with 3D object inputs for an additional 20k steps.

Test-time guidance. For the navigation model, we set the guidance weight α to 30 for goal-reaching guidance and 1000 for collision guidance. In the interaction model, we utilize weights of 1000 for goal-reaching loss and 10 for the collision SDF loss. To ensure smooth generation results, we exclude the inference guidance at the final time step of denoising. For a fair comparison with baselines, we do *not* use inference guidance unless explicitly stated in the experiment.

Evaluation Metrics For the full-body motion after in-painting, we use common metrics from prior work [2], **FID** measures the realism of the motion, **R-precision** (top-3) evaluates consistency between the text and motion, and **diversity** is computed based on the average pairwise distance between sampled motions. The **collision ratio**, the fraction of frames within generated trajectories where a collision occurs, evaluates the consistency of root motions with the environment. For the full-body motion after in-painting, we use common metrics from prior work [2], see more details in Sup. Mat. Additionally, the **foot skating** ratio [4] evaluates the physical plausibility of motion-ground interaction by the proportion of frames where either foot slides a distance greater than a specified threshold (2.5 cm) while in contact with the ground (foot height <5 cm).

The penetration value is the mean SDF value across all interpenetrated body vertices of the generated motions, while the ratio is the fraction of generated

poses containing penetrations (*i.e.*, SDF values < -3 cm) over all generated motion frames.

2 Ablation Study

Table 1: Ablation study comparing various full-body infilling methods and different representations of navigation motion generation using the Loco-3D-FRONT test set. (**Left**) For generated pelvis trajectories, our approach achieves the best goal-reaching accuracy with low collision rate. (**Right**) After in-painting the full-body motion, our method preserves diverse and realistic movements that align with the provided text prompt, much like the model employing an alternative OminiControl full-body in-painting technique. However, our approach distinctly outperforms the model utilizing full-body representation.

	Root trajectory evaluation				**Full-body motion evaluation**			
	Goal-reaching error ↓							
Method	Pos.	Orient.	Height	Collision ↓	FID ↓	R-precision ↑	Diversity ↑	Foot skating ↓
Ours (OmniControl [6] in-painting)	0.459	0.999	0.090	0.073	**17.927**	**0.396**	6.288	**0.0308**
Ours (full-body rep)	0.844	0.016	0.110	0.124	24.642	0.189	**6.967**	0.169
Ours	**0.169**	**0.119**	**0.008**	**0.031**	20.465	0.376	6.415	0.056

Alternative Full-Body In-painting Approach. While our root trajectory generation approach can integrate with several motion in-painting techniques, in the main paper we use PriorMDM [5]. As an alternative, we evaluate our method using OminiControl [6] for in-painting in Tab. 1. However, OmniControl overrides our generated dense pelvis trajectory and jointly generates full-body locomotion with a new pelvis trajectory. This severely degrades the goal-reaching ability (from 0.169 cm to 0.459 cm) as demonstrated in Table 1. Therefore, we choose to utilize PriorMDM as our body motion in-painting method. It aligns well with our generated trajectory, resulting in the generation of plausible locomotion while maintaining adherence to the goal position.

One-stage Navigation Motion Generation. To evaluate the efficacy of our two-stage navigation model design, we compare to a single-stage full-body motion generation ablation of our model. This model operates on the same input data but directly generates full-body locomotion. However, as shown in Tab. 1, this approach limits goal-reaching ability and does not produce motion styles that align with the input text. The local poses are somewhat dissociated from the global pelvis trajectories, allowing for trajectory variations while maintaining the same motion style. For instance, individuals can walk along different paths while maintaining consistency in their motion style.

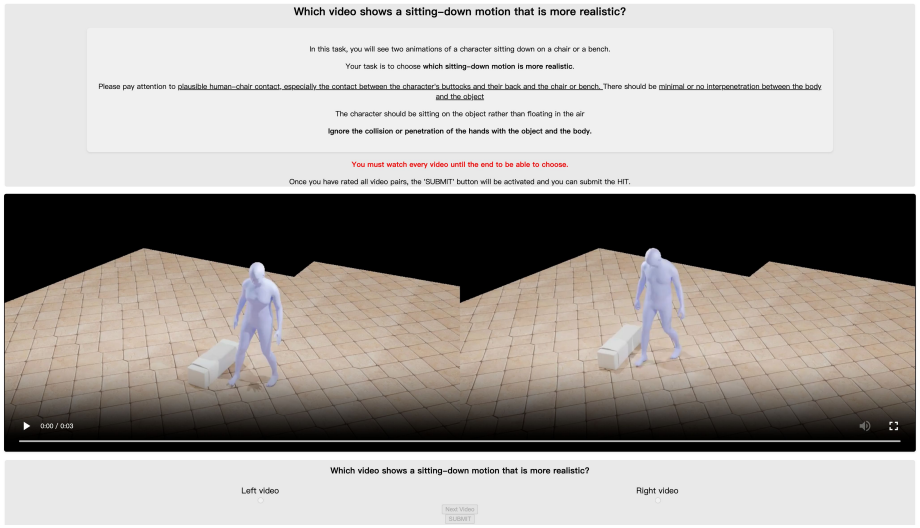

Fig. 1: The layout of our perceptual study for evaluating the plausibility of human-object interaction.

3 Details on User Study for Interaction Motions

To evaluate the plausibility of human-object interaction, we perform a user study to compare our method and DIMOS [7]. We employ Amazon Mechanical Turk (AMT) [1] to solicit assessments from 30 individuals. Raters are presented with two side-by-side videos depicting generated interactions and asked to determine which appeared more realistic, particularly focusing on the contact between the character's buttocks and their back with the chair or bench, and the presence of minimal or no interpenetration between the body and the object. We present 70 test videos with the positions of our generated videos and DIMO's results randomly shuffled horizontally. In order to filter out poor responses, we duplicated our 5 test examples where clear preferences between two video results were evident, serving as catch trials. Ultimately, we obtained 65 useful responses out of 70 raters. The full survey page is illustrated in Fig. 1. The user study reveals a distinct preference for motions generated by our approach (preferred 71.9%) over those produced by DIMOS.

4 Details on Collision Guidance Used in Interaction Motion Generation

At test time, a collision objective is used to discourage penetrations between humans and objects. Remarkably, our interaction motion generation model outputs 3D joint positions. We then link randomly sampled vertices on the SMPL mesh surfaces with the 3D skeletons in an A-pose, allowing us to obtain the posed

sampled vertices for each new pose. This is defined as $\mathcal{J}_c = \text{SDF}(\hat{\mathbf{x}}_0, \mathcal{S}_{\mathcal{O}})$ where SDF calculates the SDF volume of the object O, then queries the sign distance value at each time step of the body vertices. Positive distances, indicating body vertices inside the interactive object, are averaged to get the final loss.

References

1. Amazon Web Services, Inc.: Amazon mechanical turk (Accessed 2024), https://www.mturk.com/ 3
2. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022) 1
3. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11374–11384 (October 2021) 1
4. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023) 1
5. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 2
6. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023) 2
7. Zhao, K., Zhang, Y., Wang, S., Beeler, T., , Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: International conference on computer vision (ICCV) (2023) 3