
Learning to Generate Human-Human-Object Interactions from Textual Descriptions

—Supplementary Material—

Jeonghyeon Na*, Sangwon Baik*, Inhee Lee, Junyoung Lee, Hanbyul Joo†

Seoul National University

*Equal Contribution †Corresponding Author

{prom317, bsw1907, ininin0516, juncong, hbjoo}@snu.ac.kr

A Data Collection Details

A.1 Multiview HHOIs Data Capture System

To capture Human-Human-Object Interactions (HHOIs), we adopt a multi-camera setup inspired by Panoptic Studio [7]. As shown in Fig. 1, our system is composed of 36 synchronized RGB cameras to ensure accurate human pose estimation, even under severe occlusions during interaction. For tracking the 6-DoF pose of the interacting object, we attach ArUco markers [6] on the object surface, which provide robust and efficient object pose annotation throughout the interaction sequence.

We use DWPose [14] to extract 2D human keypoints from each camera view. To associate human detections across views, we incrementally cluster the detected people in 3D by minimizing the overall triangulation error, ensuring consistent identity assignment. Once the keypoints are clustered, we follow the preprocessing pipeline of PaMIR [16] to fit SMPL-X parameters for each individual. This involves optimizing body pose, scale, translation, and rotation with regularization from learned priors [12], resulting in plausible 3D human poses.

For object tracking, we begin by placing the target object at the center of the capture system to align its template mesh within the coordinate system of our multi-camera setup. After capture, we

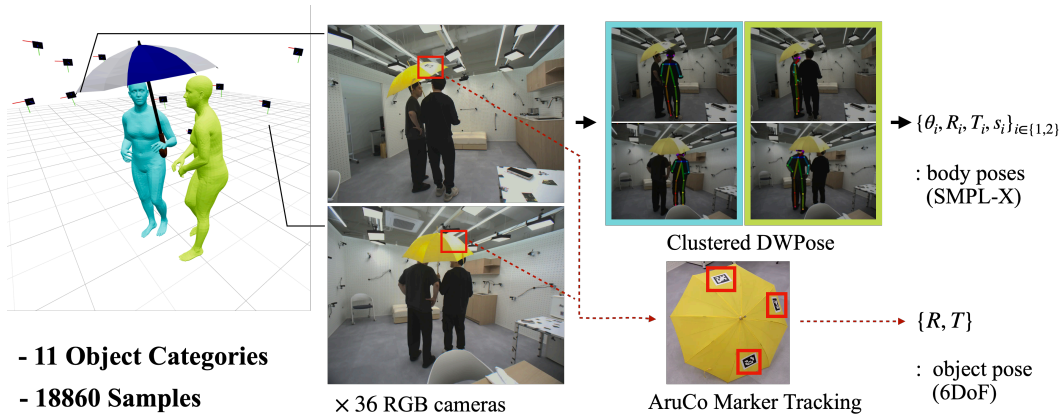


Figure 1: **HHOIs Capture System Overview.** We capture Human-Human-Object Interactions (HHOIs) with our multiple camera capture system. The object and human poses are tracked with ArUco markers [6] and DWPose [14] respectively.

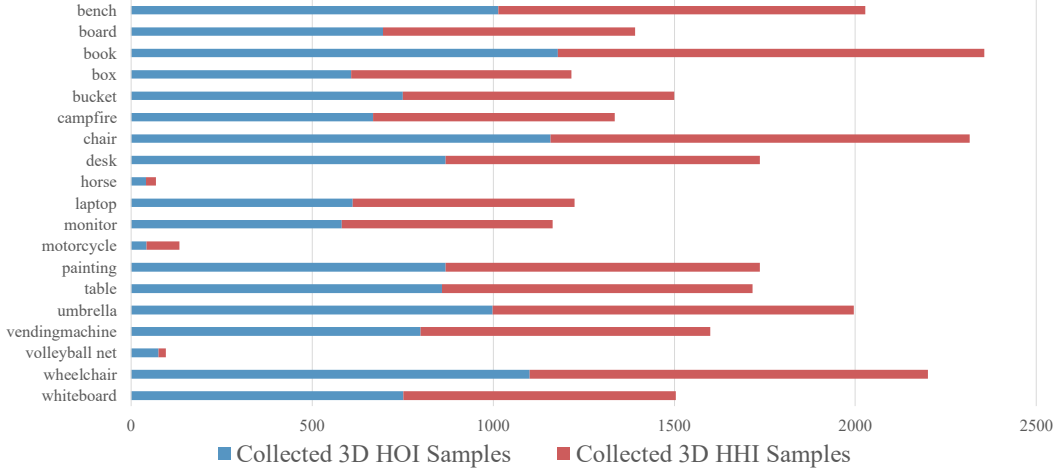


Figure 2: **Statistics on Collected Data Samples.** Our dataset is constructed by integrating data from CORE4D and our multiview capture system, alongside synthetic samples generated via our data generation pipeline.

reconstruct a 3D Gaussian Splatting (3D-GS) [8] scene from the initial frames and manually align the template mesh to the physical object in the scene to determine its object pose. For dynamic objects, subsequent poses are computed by applying 6-DoF transformations from the ArUco markers.

During the data capture process, two actors are provided with high-level instructions for the interaction scenario while maintaining flexibility in execution to encourage natural behaviors and a diverse range of human poses. After recording, we post-process the captured sequences by categorizing them into fine-grained sub-scenarios based on interaction type and body configuration. To ensure data quality, we filter out frames where SMPL-X fitting is unreliable due to identity clustering failure, occlusion, or low keypoint confidence, resulting in a clean and robust dataset for downstream tasks.

A.2 Data Preprocessing

For the collected HHOI data from CORE4D [10] and our multiview capture system, each sample includes the object instance, as well as the SMPL-X translation, rotation, and body pose parameters for the two interacting individuals. To prepare the data for our model, we first normalize the object’s position and orientation to the origin and a canonical frame. The human poses are then transformed accordingly, such that the resulting SMPL-X parameters represent each individual’s body pose, translation, and rotation relative to the canonical object frame. These parameters are used as input to our HOI model.

For the HHI model, we normalize one individual’s position to the origin and its rotation to the identity (i.e., zero in the SMPL-X global rotation parameter), effectively expressing the second human’s translation and rotation relative to the first. We repeat this process by switching the reference human, resulting in two HHI data samples per HHOI frame. After preprocessing, the data is split into training and test sets with a 9:1 ratio. This test set serves as the ground-truth distribution for computing the Fréchet Distance in Sec. 5 of the main paper, enabling quantitative comparison of the realism of our generated results against the baseline methods.

For synthetic data, ComA directly generates HOI data in a format compatible with our model. Synthetic HHI samples are derived from the Human Mesh Recovery output with the above explained normalization procedure.

A.3 Data Statistics

In total, we collect 13,669 3D HOI samples and 13,650 3D HHI samples, spanning 19 object categories. Of these, 5 categories—board, box, bucket, chair, and desk—are obtained from the CORE-4D dataset. 11 categories—bench, book, campfire, laptop, monitor, painting, table, umbrella,

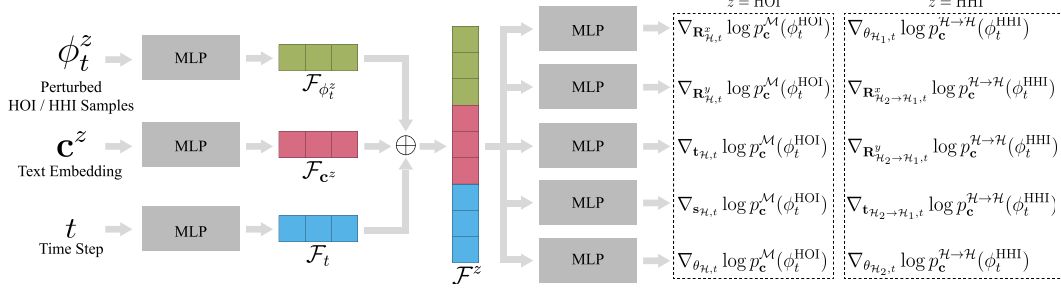


Figure 3: **HHOI Diffusion Architecture.** HHOI diffusion consists of two disjoint diffusion models: HOI diffusion and HHI diffusion. Although the overall structure of each diffusion model is the same, they are implemented with separate networks due to their different target distributions. Each network learns the score function of the HOI or HHI distribution, respectively.



Figure 4: **Capsule-based Approximation of SMPL-X Human.** Our capsule-based human approximation models SMPL-X humans well, enabling a computationally cheap collision loss formulation.

vending machine, wheelchair, whiteboard—are captured using our multi-view capture system. The remaining 3 categories—horse, motorcycle, and volleyball net—are generated using our synthetic data pipeline. 27,020 samples are collected from CORE4D and capture system, and 299 samples are generated from synthetic pipeline. Detailed statistics for each category are shown in Fig. 2.

B Implementation Details

B.1 HHOI Diffusion Architecture

We adopt the MLP-based score network from GenPose [15], which is originally designed to predict 6D poses of point clouds, as the backbone of our HHOI diffusion model. We once again emphasize that HHOI diffusion is composed of two disjoint diffusion models: HOI diffusion and HHI diffusion. These two diffusion models do not share parameters and are trained separately. However, as shown in Fig. 3, the overall architecture of the two diffusion models is the same. First, separate MLPs are used as feature extractors for the HOI or HHI sample, the CLIP text embedding, and the time step, respectively. The resulting features are denoted as $\mathcal{F}_{\phi_t^z} \in \mathbb{R}^{256}$ for the HOI or HHI sample, $\mathcal{F}_{c^z} \in \mathbb{R}^{128}$ for the CLIP text embedding, and $\mathcal{F}_t \in \mathbb{R}^{128}$ for the time step. The features are concatenated into a single vector $\mathcal{F}^z \in \mathbb{R}^{512}$, which is then passed to other MLPs to predict the score functions. Each diffusion model employs distinct MLPs to predict the score functions corresponding to each component in ϕ_t^z . In particular, following GenPose, 6D rotations are decomposed into two 3D vectors, and separate score functions are learned for each. The total number of trainable parameters in the HHOI diffusion model is 10.2M, with approximately 5.1M in each of the HOI and HHI diffusion models.

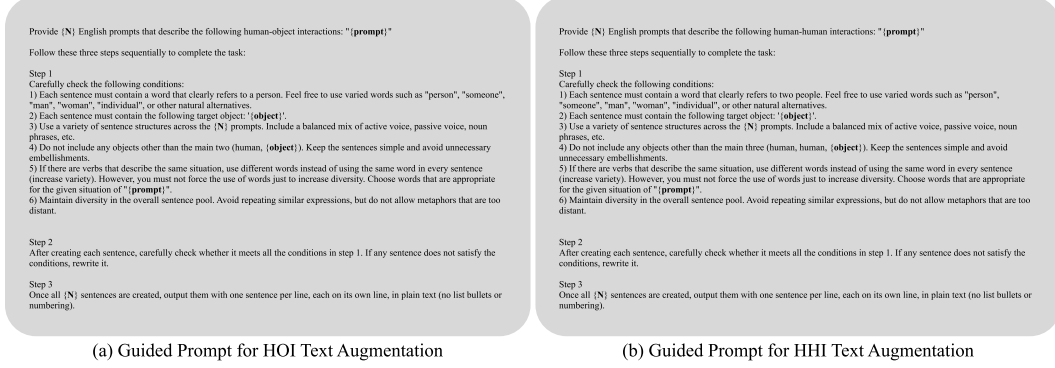


Figure 5: **Guided Prompt Provided to LLM for Text Augmentation.** (a) Guided prompt for HOI text prompt augmentation. (b) Guided prompt for HHI text prompt augmentation.

B.2 Capsule-based Human for Collision Loss

In Sec. 3.3 of the main paper, we approximate each human with a 24-capsule proxy to define a computationally efficient collision loss. A capsule is a set of points that are equidistant from a line segment called an axis segment. Consequently, collisions between two capsules are determined simply by checking whether the minimum distance between their axis segments is smaller than the sum of their radii.

Given a body pose embedding $\theta_{\mathcal{H}}$, we construct the capsule proxy through the following steps: (1) Decode $\theta_{\mathcal{H}}$ with a body pose decoder to obtain the actual $21 \times 6D$ body pose, and apply forward kinematics to obtain the positions of the 22 human joints including root; (2) Consider the 21 bones that connect each joint to its parent joint as the axis segment of each capsule; (3) Define 3 additional axis segments for the capsules corresponding to the hands and head, utilizing the two wrists and head joints; (4) Apply pre-learned per-capsule radii to obtain the 24-capsule proxy. Fig. 4 shows that our capsule-based human approximation works well.

To obtain per-capsule radii, we reuse the 922K human body pose samples used to train the body pose encoder-decoder described in Sec. 3.1 of the main paper. For each pose sample, we generate the corresponding SMPL-X mesh and its capsule-based approximation, uniformly sample 5000 surface points from each, and then optimize a set of per-capsule radii by minimizing the Chamfer distance between the two point sets.

B.3 Hyper Parameter Settings

We train both the HOI and HHI diffusion models for 20,000 epochs with a batch size of 500. The learning rate is initialized at $1e-2$ and decays by a factor of 0.999 at each step, with a lower bound of $7e-4$. For guided HHOI sampling, we set $\lambda_1 = \min(100000, \frac{100}{t^2})$ and $\lambda_2 = \min(1600000, \frac{1600}{t^2})$, which correspond to the weight terms defined in Eq. 15 of the main paper. To ensure that the inconsistency loss and collision loss have meaningful effects, we apply these losses starting from the time step $t = 0.5$. Consequently, HHOI samples are obtained by solving the following PF ODE:

$$\begin{aligned} \phi_1^{z,i} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I}_z), \\ \frac{d\phi_t^{z,i}}{dt} &= -\sigma(t)\dot{\sigma}(t)\Psi_{\Theta^z}(\phi_t^{z,i}, t|\mathbf{c}^z, z), \quad t \in (0.5, 1.0], \\ \frac{d\phi_t^{z,i}}{dt} &= -\sigma(t)\dot{\sigma}(t)\Psi_{\Theta^z}(\phi_t^{z,i}, t|\mathbf{c}^z, z) + \lambda_1 \nabla_{\phi_t^{z,i}} \mathcal{L}_{\text{inc}}(\Phi_t) + \lambda_2 \nabla_{\phi_t^{z,i}} \mathcal{L}_{\text{col}}(\Phi_t), \quad t \in [\epsilon, 0.5]. \end{aligned} \quad (1)$$

B.4 Text Augmentation for Training

We adopt the LLM-based text augmentation method proposed in [1] to train our text-conditioned HHOI diffusion model. Fig. 5 shows the detailed guided prompts to be provided to LLM [11]. The guided prompt instructs the LLM to produce diverse sentence structures, varied verbs, and concise

sentences without flowery language. We demonstrate in Sec. D that our model generalizes to unseen text.

C Experiments Details

C.1 Baseline Methods

GenZI We modify GenZI [9] method to generate multiple number of humans in the given scene, instead of one. We follow GenZI’s pipeline in selecting the views to render and inpaint the scene. In the inpainting process, we change the input text prompt to the inpainting pipeline to suit our desired number of humans, and correspondingly set the tokens used in dynamic masking. AlphaPose [5] detects the 2D joint positions of generated humans in each image. Given that each image contains multiple individuals, establishing cross-view correspondence for each human is needed to reconstruct 3D humans from the multi-view images. We estimate the correspondence by evaluating possible correspondence combinations and selecting the set that minimizes the reprojection error. Subsequently, the original objective function of GenZI is calculated independently for each individual, and the final loss is obtained by summing these individual losses. For the iterative refinement, we load all the generated humans and render the silhouette to use as the inpainting mask.

Depth Opt. For a simple but effective baseline, we employ Human Mesh Recovery and Depth Estimation method to generate 3D humans from text prompt and scene. Following the GenZI framework, we render a single-view image of the scene and inpaint multiple individuals based on a given text prompt. Then, we reconstruct the 3D human poses from the generated image using multi-HMR [2]. With the generated SMPL-X humans and the scene, we render to obtain the image and depth. Then, we use Depth-Pro [3] to estimate the depth from the image. The image is segmented into scene-only and human-only pixels. For the scene-only regions, we compute the difference between the estimated depth and the ground-truth depth obtained from the renderer. This difference is averaged across all scene-only pixels and used as an offset, which is then applied to the estimated depths in the human-only regions. Finally, we compute the ratio between the offset-corrected estimated depth and the rendered depth for the human-only pixels, and apply this ratio to scale the generated 3D human reconstructions accordingly.

C.2 User Study

We conduct a user study to evaluate how effectively our model captures human-human and human-object interactions. We collect responses from 97 participants via CloudResearch [4], offering a reward credit of \$1.50 per participant. The study comprises 30 questions covering 10 objects, each associated with three randomly ordered interaction scenarios—one from our model and two from baseline methods. For each question, participants are presented with 10 multi-view renderings of the HHOI scene, showing the humans, object, and environment as generated by each method. They are asked to select the rendering that best illustrates the given text prompt. The structure of the questionnaire is shown in Fig. 6.

D Additional Qualitative Results

Fig. 7 shows the HHOI results obtained using the advanced sampling introduced in Sec. 3.3 of the main paper. Our method performs well even in multi-human cases involving more than two individuals across diverse categories.

We additionally provide ablation results for our guided HHOI sampling. Fig. 8 presents the HHOI sampling results for three versions: our full method, a version without collision loss, and a version without both collision and inconsistency losses. Without the collision loss, in cases involving more than three individuals, collisions are more likely to occur between humans whose HHI relationships are not explicitly specified. Removing the inconsistency loss disrupts the integration of HOI and HHI, producing less plausible HHOI results.

Finally, to demonstrate the effectiveness of our text augmentation, we present in Fig. 9 the HHOI sampling results for unseen text prompts.

COMMON GUIDELINES

Your task is to **select the image set that best reflects the described human-human and human-object interaction**. Please base your judgment **only** on how well the spatial relationships and interactions match the given prompt.

Each set is a 360-degree rendering of the scene. **Some views may not clearly show the interaction due to limited field of view—please ignore those views when making your decision.** Do **not** consider other factors such as texture quality, lighting, or scene mesh quality.

This study takes approximately **5 minutes** to complete. Please answer each question carefully. Your feedback is vital for our research on interaction understanding.

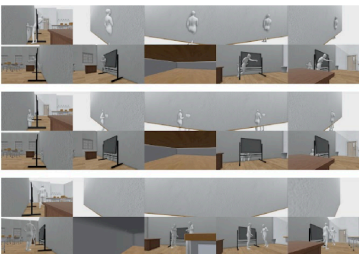
Which set of multi-view images best illustrates the human-human and human-object relationship described in the following text prompt: "Two women are discussing with each other in front of the whiteboard?" } Prompt

Please select the set that most accurately represents the **human-human and human-object interaction** described in the given text prompt.

☐ A

☐ B

☒ C



} Renders

30 questions covering various scenarios →

Figure 6: **Questionnaire for User Study.** Participants select the multi-view image that best depicts the human-human and human-object relationship.

E Licenses and Rights

The following 3D models were used in this work for dataset and rendering. All assets are publicly available under the Creative Commons Attribution (CC-BY 4.0 or CC BY-NC 4.0) license or cited from academic datasets. Attribution is provided below where required.

- **Bench**
Source: <https://fab.com/s/5cfe864b44e1>
License: CC BY 4.0
Attribution: "Bench" is licensed under CC BY 4.0.
- **Book**
Title: Book opened
Source: <https://skfb.ly/6s9uU>
License: CC BY 4.0
Attribution: "Book opened" by Jiří Kuba is licensed under CC BY 4.0.
- **Campfire**
Title: Campfire Wood Survival Warm and Light
Source: <https://skfb.ly/6QYoY>
License: CC BY 4.0
Attribution: "Campfire Wood Survival Warm and Light" by digrafstudio is licensed under CC BY 4.0.
- **Horse**
Title: Horse basemesh
Source: <https://skfb.ly/oZoW8>
License: CC BY 4.0
Attribution: "Horse basemesh" by Sid Ahearne is licensed under CC BY 4.0
- **Laptop**
Dataset: SAPIEN (Object ID: 10211) [13]
- **Monitor**
Title: Desktop Computer
Source: <https://skfb.ly/6RCNx>
License: CC BY 4.0
Attribution: "Desktop Computer" by Tytan is licensed under CC BY 4.0.
- **Motorcycle**
Title: Motorcycle
Source: <https://skfb.ly/6oEPN>
License: CC BY 4.0
Attribution: "Motorcycle" by Silas6 is licensed under CC BY 4.0

- **Painting**
Title: Birthplace Painting
Source: <https://skfb.ly/6zNAL>
License: CC BY 4.0
Attribution: "Birthplace Painting" by Znyth Technologies is licensed under CC BY 4.0.
- **Table**
Title: Kitchen table
Source: <https://skfb.ly/6Goxs>
License: CC BY 4.0
Attribution: This work is based on "Kitchen table" by tahax licensed under CC BY 4.0.
- **Umbrella**
Title: Umbrella
Source: <https://skfb.ly/6YpNI>
License: CC BY 4.0
Attribution: "Umbrella" by Diccbudd is licensed under CC BY 4.0.
- **Vending Machine**
Title: Vending Machine
Source: <https://skfb.ly/6ZtVQ>
License: CC BY 4.0
Attribution: "Vending Machine" by yashwanthantony9542 is licensed under CC BY 4.0
- **Volleyball Net**
Title: volleyball net
Source: <https://skfb.ly/oYMHq>
License: CC BY 4.0
Attribution: "volleyball net" by otyken is licensed under CC BY 4.0
- **Wheelchair**
Title: Wheelchair
Source: <https://skfb.ly/ouV8F>
License: CC BY 4.0
Attribution: "Wheelchair" by Fine_poultry is licensed under CC BY 4.0.
- **Whiteboard**
Source: <https://www.fab.com/listings/398f36d0-bd12-4d93-b348-0f02f1677eae>
License: CC BY 4.0
Attribution: "Whiteboard" is licensed under CC BY 4.0.

The following 3D models were used in this work as a background mesh to render our generation result, and for the baseline model implementation. Attribution is provided below where required.

- **Low Poly Farm V2**
Source: <https://skfb.ly/6QYJI>
License: CC BY 4.0
Attribution: "Low Poly Farm V2" by EdwiixGG is licensed under CC BY 4.0
- **Bangkok City Scene**
Source: <https://skfb.ly/6GxU0>
License: CC BY 4.0
Attribution: "Bangkok City Scene" by ArneDC is licensed under CC BY 4.0
- **Living Room**
Source: <https://skfb.ly/6wYHE>
License: CC BY 4.0
Attribution: "Living Room" by Taranpreet is licensed under CC BY 4.0
- **Camp Scene, Free Download**
Source: <https://skfb.ly/6SnYV>
License: CC BY NC-4.0
Attribution: "Camp Scene, Free Download" by Bento is licensed under CC BY NC-4.0
- **Low Poly Simple Hallway Room**
Source: <https://skfb.ly/oyLtI>

License: CC BY 4.0

Attribution: "Low Poly Simple Hallway Room" by jimbogies is licensed under CC BY 4.0

- **Livingroom**

Source: <https://skfb.ly/6RBx8>

License: CC BY 4.0

Attribution: "Livingroom" by Amy is licensed under CC BY 4.0

- **Gallery Museum Showroom Banquet Hall**

Source: <https://skfb.ly/otq0Y>

License: CC BY 4.0

Attribution: "Gallery Museum Showroom Banquet Hall" by jimbogies is licensed under CC BY 4.0

- **Venice city scene 1DAE08 Aaron Ongena**

Source: <https://skfb.ly/6TptH>

License: CC BY 4.0

Attribution: "Venice city scene 1DAE08 Aaron Ongena" by AaronOngena is licensed under CC BY 4.0

- **Basic Classroom**

Source: <https://skfb.ly/ovnsE>

License: CC BY 4.0

Attribution: "Basic Classroom" by Kibele is licensed under CC BY 4.0

- **Tennis Court**

Source: <https://skfb.ly/6YKBs>

License: CC BY 4.0

Attribution: "Tennis Court" by Spark Games is licensed under CC BY 4.0

References

- [1] S. Baik, H. Kim, and H. Joo. Learning 3d object spatial relationships from pre-trained 2d diffusion models. *ICCV*, 2025. 4
- [2] F. Baradel*, M. Armando, S. Galaaoui, R. Brégier, P. Weinzaepfel, G. Rogez, and T. Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. *ECCV*, 2024. 5
- [3] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun. Depth pro: Sharp monocular metric depth in less than a second. *ICLR*, 2025. 5
- [4] CloudResearch. Connect cloud research. URL <https://connect.cloudresearch.com/researcher/>. 5
- [5] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *TPAMI*, 2022. 5
- [6] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 2014. 1
- [7] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. *CVPR*, 2015. 1
- [8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *SIGGRAPH*, 2023. 2
- [9] L. Li and A. Dai. GenZI: Zero-shot 3D human-scene interaction generation. *CVPR*, 2024. 5
- [10] Y. Liu, C. Zhang, R. Xing, B. Tang, B. Yang, and L. Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353*, 2024. 2
- [11] OpenAI. Chatgpt: Optimizing language models for dialogue, 2023. URL <https://openai.com/blog/chatgpt>. 4
- [12] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. *CVPR*, 2019. 1

- [13] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, et al. Sapient: A simulated part-based interactive environment. *CVPR*, 2020. 6
- [14] Z. Yang, A. Zeng, C. Yuan, and Y. Li. Effective whole-body pose estimation with two-stages distillation. *ICCV*, 2023. 1
- [15] J. Zhang, M. Wu, and H. Dong. Generative category-level object pose estimation via diffusion models. *NeurIPS*, 2024. 3
- [16] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 2021. 1



“People are having a discussion in front of a whiteboard.”



“People are sitting at a desk.”



“People are facing a monitor while seated.”



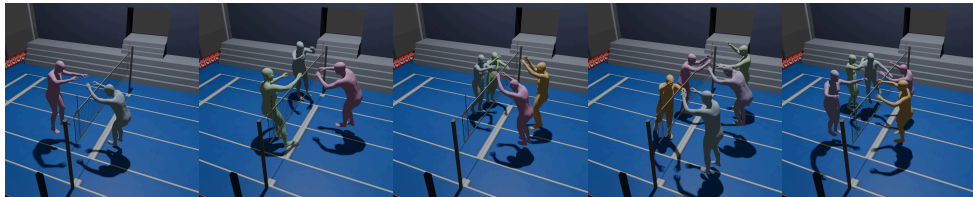
“People are reading a book while seated.”



“People are looking at a laptop while seated.”



“People are viewing a painting.”



“People are jumping in front of a volleyball net.”

Figure 7: Additional HHOI Qualitative Results.

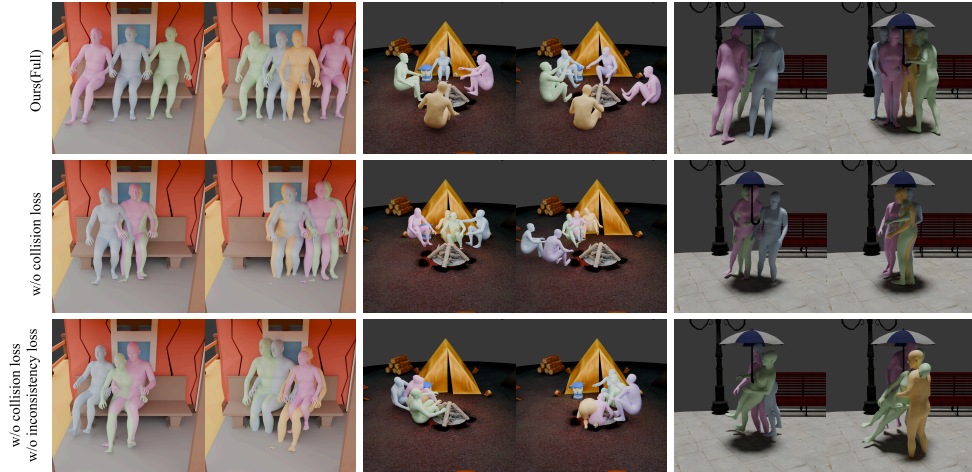


Figure 8: Ablation Study for Guided HHOI Sampling.

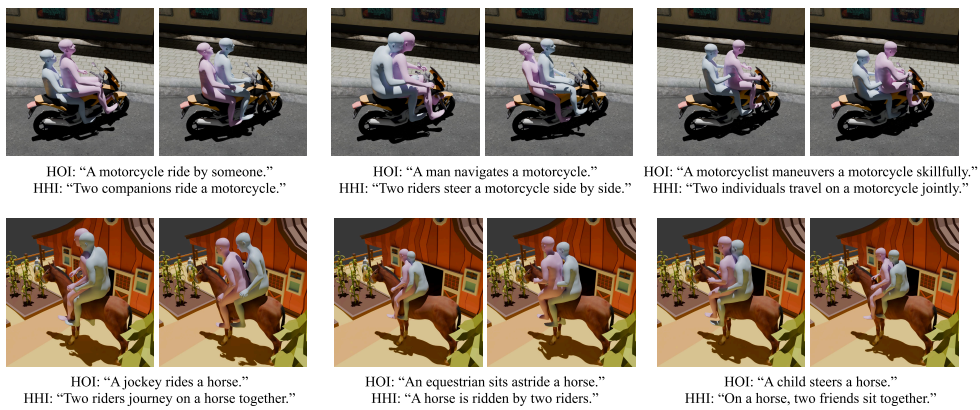


Figure 9: Qualitative Results of HHOI Sampling for Unseen Text Prompts