# Supplemental Materials for HIMO: A New Benchmark for Full-Body Human Interacting with Multiple Objects

Xintao Lv[1*], Liang Xu[1,2*], Yichao Yan[1†], Xin Jin[2], Congsheng Xu[1], Shuwen Wu[1], Yifan Liu[1], Lincheng Li[3], Mengxiao Bi[3], Wenjun Zeng[2], and Xiaokang Yang[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3] NetEase Fuxi AI Lab
https://lvxintao.github.io/himo

## A Loss Formulations of HIMO-Gen

In order to train the HIMO-Gen framework, we adopt four types of losses specifically designed for our task of text-driven human-object interaction (HOI) synthesis.

To generate authentic human motion, we adopt the widely used geometric loss in recent text-to-motion works [1,5,6], including joint position loss $\mathcal{L}_{pos}$ and joint velocity loss $\mathcal{L}_{vel}$. Formally, we denote the ground truth $j$-th joint position of human of the $n$-th frame as $P_n^j$, and the generated result as $\hat{P}_n^j$. The joint position loss $\mathcal{L}_{pos}$ is formulated as:

$$\mathcal{L}_{pos} = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J}\|P_n^j - \hat{P}_n^j\|_2^2, \tag{1}$$

where $N$ is the length of human motion and $J$ is the number of joints. Similarly, the joint velocity loss $\mathcal{L}_{vel}$ is formulated as:

$$\mathcal{L}_{vel} = \frac{1}{N-1}\sum_{n=1}^{N-1}\sum_{j=1}^{J}\|(P_{n+1}^j - P_n^j) - (\hat{P}_{n+1}^j - \hat{P}_n^j)\|_2^2. \tag{2}$$

Following [3], a signed distance field (SDF) based interpenetration loss $\mathcal{L}_{pen}$ is also adopted to prevent the collision between human and objects. Let $\phi$ be a modified SDF of the human body defined as follows:

$$\phi(x, y, z) = -\min(SDF(x, y, z), 0). \tag{3}$$

According the definition above, points inside the human body have positive values of $\phi$ proportional to the distance from the surface, otherwise it equals 0 outside of the human body. Here we define $\phi$ on a voxel grid of dimensions

$N_h \times N_h \times N_h$, where $N_h = 32$. For the object $o$, its $i$-th sampled point on the surface is denoted as $v_o^i$. Thus, the interpenetration loss of object $o$ colliding with the human is defined as :

$$P_o = \sum_{i=1}^{S} \tilde{\phi}(v_o^i), \tag{4}$$

where $S$ is the number of sampled points on the surface, and $\tilde{\phi}(v_o^i)$ samples the $\phi$ value for each 3D point $v_o^i$ in a differentiable way from the 3D grid via trilinear interpolation. Then the interpenetration loss for all objects is formulated as:

$$\mathcal{L}_{pen} = \sum_{o \in O} P_o, \tag{5}$$

where $O$ is the set of the involved objects.

To model the spatial relation between objects, we further use an object-pairwise loss $\mathcal{L}_{dis}$. Formally, we denote the sampled points set of object $i$ and $j$ during the whole HOI sequence as $V_i^{1:N}$ and $V_j^{1:N}$, and the points set transformed by the generated motion as $\hat{V}_i^{1:N}$ and $\hat{V}_j^{1:N}$. We then have the distance between the two objects as $\Delta V_{ij} = \|V_i^{1:N} - V_j^{1:N}\|_2^2$. Our insight is that this distance between them follows certain pattern, so that we adopt an $\mathbb{L}2$ loss to keep the consistency between the generated results and the ground truth. The object-pairwise loss is then formulated as :

$$\mathcal{L}_{dis} = \sum_{i \neq j} \|\Delta V_{ij} - \Delta \hat{V}_{ij}\|_2^2. \tag{6}$$

## B Implementation Details of Baselines

We re-implement MDM [6], PriorMDM [5] and IMoS [2] to support the condition input of of object geometries and initial states of human and objects. Below we detailed the implementation of each model.

**MDM.** [6] We extend the original feature dimensions of the input and output in MDM [6] from $D_h$ to $D_h + D_o$, where $D_h$ denotes the dimension of human motion representation and $D_o$ denotes that of object motion representation. To embed the condition input of object geometry, we feed it into a linear layer and concatenate it with initial pose of the object. Then all conditions are concatenated with the noised input into the motion embedding.

**PriorMDM.** [5] The original PriorMDM [5] is intended for dual-person motion generation with two branches of MDM and one singular ComMDM to coordinate the two branches. We modify the two human-motion branches into a human-motion branch and an object-motion branch. Also we place the ComMDM module after the 4-th transformer layer of each branch to enable the communication between the two branches.

**IMoS.** [2] IMoS [2] is an intent-driven HOI synthesis model with the architecture of VAE [4]. We modify the input action label into our text prompt and integrate the arm and body synthesis module into one. Additionally, the movement of

**Table A: The object categories of the HIMO dataset.**

| | | | | | | |
|---|---|---|---|---|---|---|
| 01. Plate | 02. Pan | 03. Teapot | 04. Washbasin | 05. Bowl | 06. Trashcan | 07. Dustpan |
| 08. Flowerpot | 09. Phoneholder | 10. Glasses case | 11. Knife | 12. Spatula | 13. Beer | 14. Spoon |
| 15. Knife board | 16. Broom | 17. Sprinkler | 18. Toothbrush | 19. Toothpaste | 20. Faucet | 21. Phone |
| 22. Bulb | 23. Lampholder | 24. Television | 25. Remote controller | 26. Hammer | 27. Notebook | 28. Pen |
| 29. Glasses | 30. Camera | 31. Laptop | 32. Mouse | 33. Headset | 34. Cube-M | 35. Cube-L |
| 36. Cylinder-M | 37. Cylinder-L | 38. Pyramid-M | 39. Pyramid-L | 40. Banana | 41. Apple | 42. Lemon |
| 43. Cabbage | 44. Pepper | 45. Teacup | 46. Mug | 47. Goblet | 48. Bottle | 49. Plane |
| 50. Piggybank | 51. Rubber duck | 52. Stanford rabbit | 53. Train | | | |

**Table B: The interaction settings of the HIMO dataset.**

| | | |
|---|---|---|
| 01. Put A(and B) into C | 02. Throw A(and B) into C | 03. Wash A(and B) under faucet |
| 04. Cut A on the knife board | 05. Pour water into A | 06. Transfer water between A and B |
| 07. Hit A with a hammer | 08. Play A with B | 09. Cook |
| 10. Have a meal | 11. Sweep the table | 12. Brush teeth |
| 13. Place A on B | 14. Use A and B | 15. Stack A(and B) on C |

objects is also generated through the VAE instead of the original optimization module.

## C Object Categories and Interaction Settings

### C.1 Object Categories

We list all 53 daily-life objects of our HIMO dataset in Tab. A.

### C.2 Interaction Settings

We list our interaction settings in Tab. B. Here A, B and C denote objects in Appendix C.1.

### C.3 Object Combinations

We visualize the distribution of object combinations in Fig. A

## D Generalization Experiments

**Unseen object geometries.** We experiment on the generalization ability to unseen object geometries of our `HIMO-Gen` model. We choose several object meshes that have the same category as our dataset but different geometries. We feed their geometries as input conditions to our model and synthesize corresponding human and objects motion. Visualization results are shown in Fig. B. We

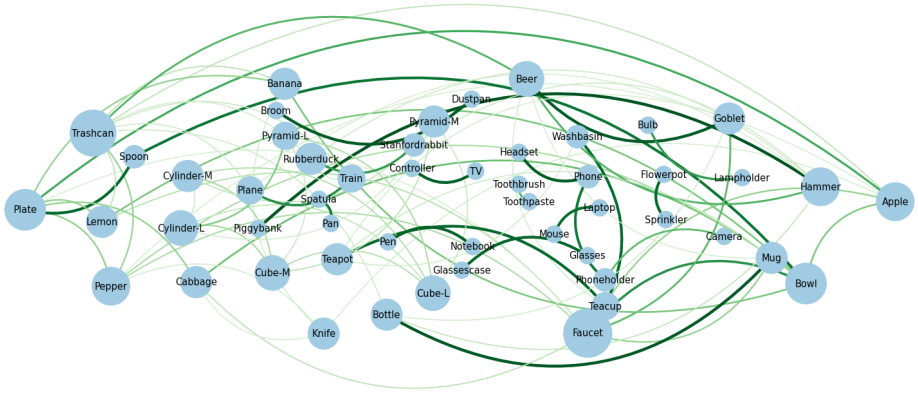


**Fig. A: Visualization of the distribution of object combinations.**

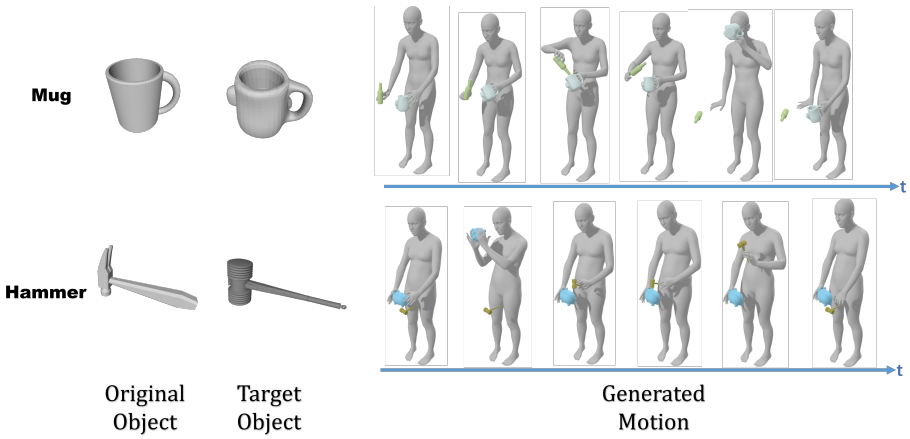


**Fig. B: Generalization experiment on unseen object meshes.**

can observe that, to some extent, our model can generalize to unseen object meshes despite of some flaws of human-object contact, which may result from the deficiency of their geometric information in our model.

**Novel HOI compositions.** We experiment on the ability of our `HIMO-SegGen` to generate *novel* HOI compositions. We choose several HOI combinations that never appear in our training set, for instance, "Knocks on the beer with a hammer" → "Grabs the beer to observe" → "Puts them down on the table". Then we feed the texts to our `HIMO-SegGen` consecutively to auto-regressively generate HOI clips. Visualization results are depicted in Fig. C.

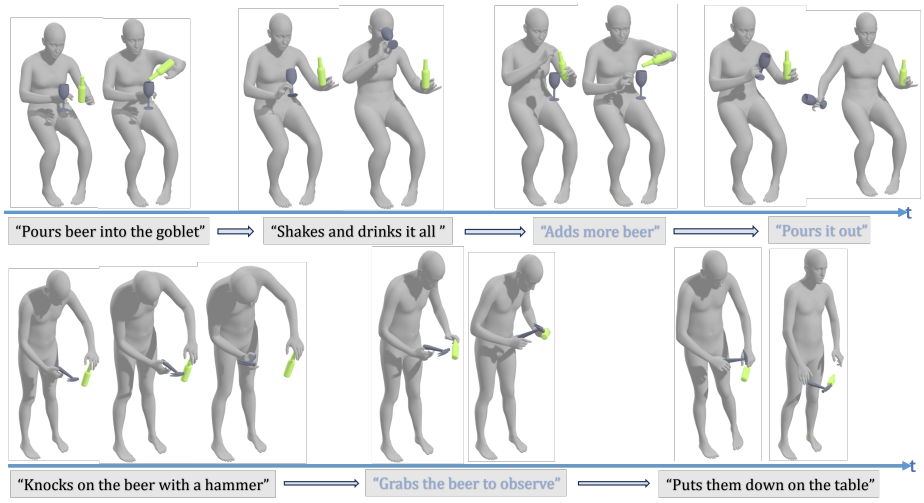


**Fig. C: Generalization experiment on novel HOI compositions.** The text in blue denotes the *novel* HOI action.

## E More Visualization Results

**Visualization of Dataset.** We present more samples from our HIMO dataset in Fig. D and Fig. E. Note that we segment both the motion and text descriptions into several clips. Additional visualization results are presented in the supplementary video.

## F Societal impacts and responsibility to human subjects.

Our dataset can be leveraged to generate plausible human interactions with multiple objects, which may lead to the creation and spread of misinformation. For privacy concerns, all performers signed the agreement on the release of their motion data for research purposes. We focus on the pure motion rather than RGB, thus no RGB videos are released and their identity will not be leaked.
**Visualization of HIMO-SegGen.** The generated results of `HIMO-SegGen` are shown in Fig. F. Each text prompt is sent to `HIMO-SegGen` and generates motion clips auto-regressively conditioned on the last 10 frames of the previous clip.
**Visualization of HIMO-Gen.** We show qualitative comparisons of our `HIMO-Gen` methods with baseline methods in Fig. G. From the figure we can see that our methods keep both consistency and semantics over the baseline methods.

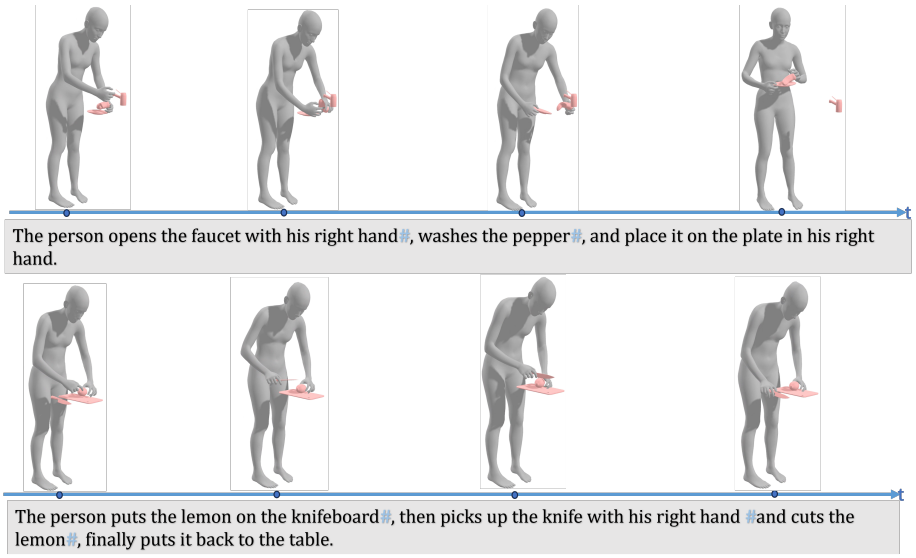


**Fig. D: More visualization samples of our HIMO dataset.** # is the separator of our temporal segments.

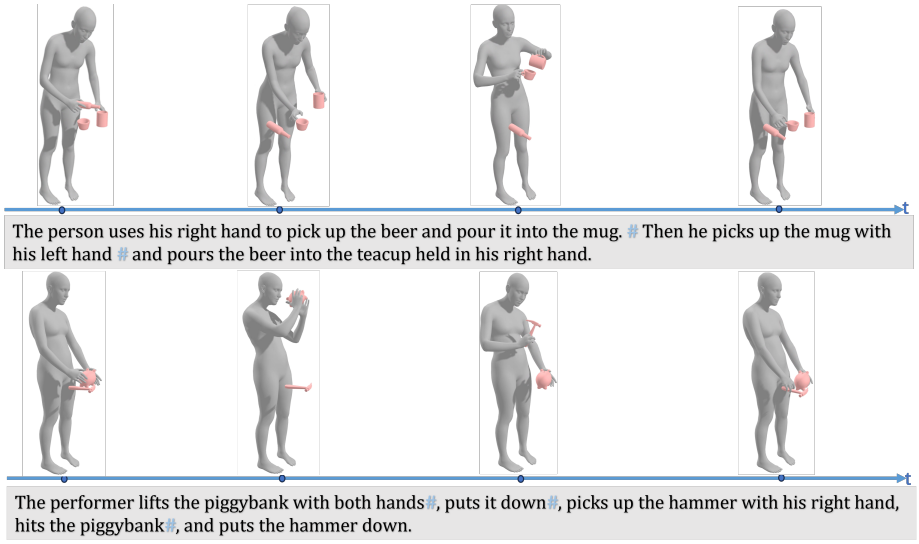


**Fig. E: More visualization samples of our HIMO dataset.** # is the separator of our temporal segments.

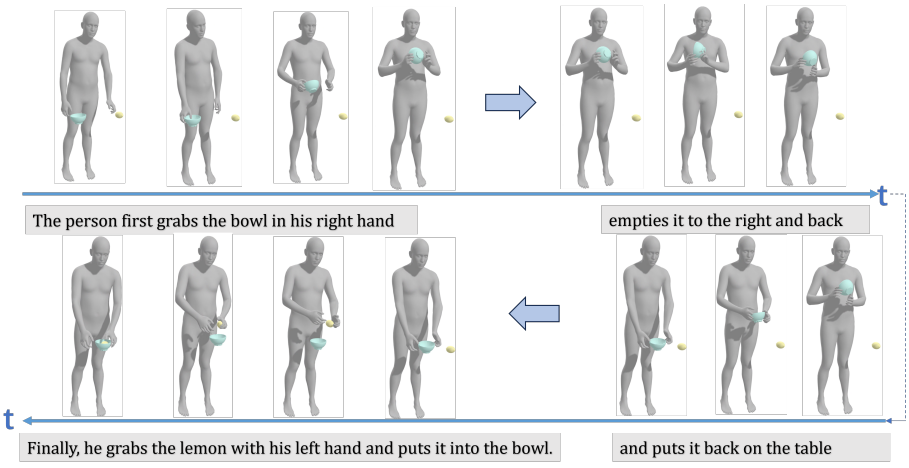


**Fig. F: Visualization results of `HIMO-SegGen`.**

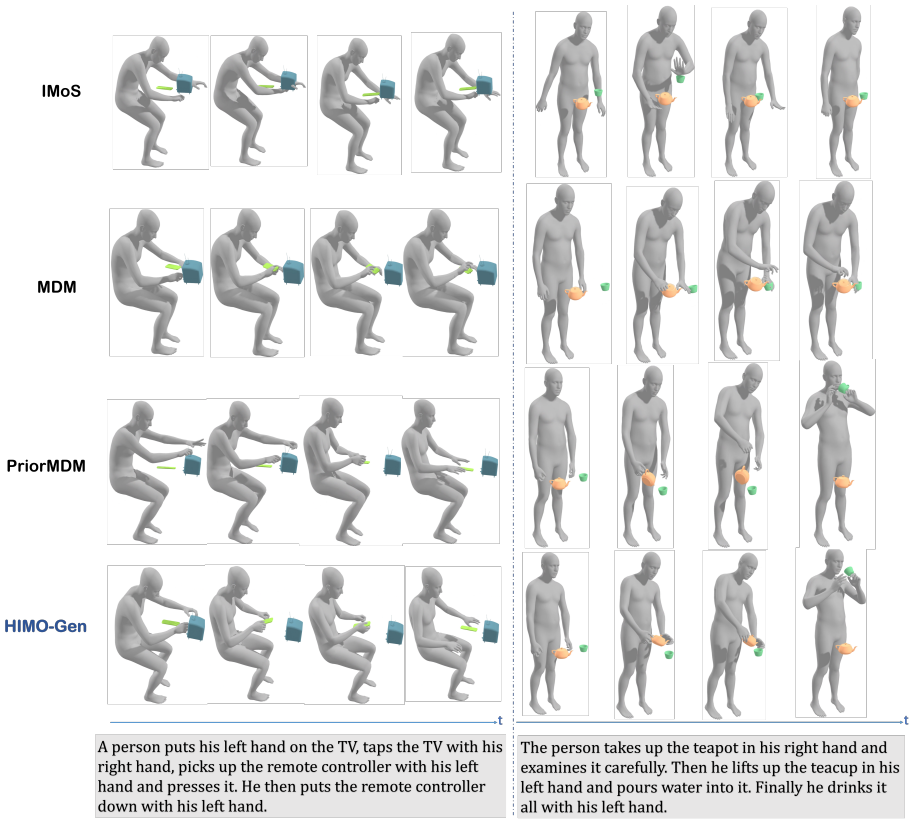


**Fig. G: Visualization comparisons of `HIMO-Gen` and baselines.**

## References

1. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR. pp. 18000–18010 (2023)
2. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Imos: Intent-driven full-body motion synthesis for human-object interactions. In: Eurographics (2023)
3. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020)
4. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)
5. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
6. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=SJ1kSyO2jwu`