

Supplementary Material for GraspDiffusion: Synthesizing Realistic Whole-body Hand-Object Interaction

S1. Model Architecture

For the first stage of our pipeline, we use a diffusion model to synthesize a body pose grasping the input object. The model is trained to predict plausible body parameters (6 DoF body pose, global orientation), conditioned on the object’s relative location $t_{\text{obj}} \in \mathbb{R}^3$ and the target hand $c_{\text{left}}, c_{\text{right}} \in \{0, 1\}$. These conditions are transformed to a conditional embedding v_c , which is added to the timestep embedding e_t and passed to residual blocks within the model, following [6]. We used 3 ResNet blocks for the model, and adopted a cosine noise schedule in training. Additional grasping results are provided in Figure 4, and details on model parameters are provided in Table 1.

In Figure 1, we provide an example comparison between our method and previous grasping pose generation methods [9, 10, 12, 14]. When given a object with its location relative to the human body (left row), GOAL [9] and SAGA [12] tend to create distorted poses when the object is far away from the human, as they assume it to be in the same horizontal xy-plane. FLEX assumes a world-centric coordinate system, which leads to pose ambiguity for our scenario. While COOP has a similar objective, it focuses on various object heights, and requires an extensive test-time optimization of 5 different loss terms. We are the first to utilize a lightweight diffusion model in synthesizing body grasping poses.

For the second stage of our pipeline, we use a diffusion model based on the Latent Diffusion [8] architecture, and attach encoders [5] that receives spatial features from the synthesized body pose. Specifically, we first provide three spatial conditions from the full-body grasping pose; the hu-

Parameter	Diffusion Model (Body Pose)
Input Channels	132
Condition Channels	5
Model Channels	1024
ResBlock Number	3
Diffusion Steps	1000
Noise Scheduler	Cosine

Table 1. Model architecture for body pose generation diffusion model.

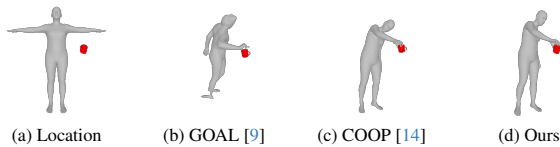


Figure 1. Grasp synthesis comparison with previous methods.



Figure 2. Synthesis results from a single image. We first synthesize a 3D Mesh from the image using TripoSR [11], InstantMesh [13], Real3D [4] and subsequently used the mesh as input.

man skeleton projection, joint depth map, and the occluded object with ambient lighting ($[s^i, d^i, o^i]$). Then we further refine the hand-object region by providing similar spatial

Parameter	Conditional Encoder(s)
Input Channels	3×64
Output Channels	[320, 640, 1280, 1280]
ResBlock Number	2
Kernel Size	1
Feature Weight (Body)	[1.0, 0.6, 1.0]
Feature Weight (Hand)	[1.0, 0.6, 1.0]
Parameter	Attention Injection
Human Strength	0.2
Object Strength	1.8
Negative Object Strength	-9.0
Weight coefficient (w')	0.4

Table 2. Model architecture for scene generation models, and inference parameters for attention injection.

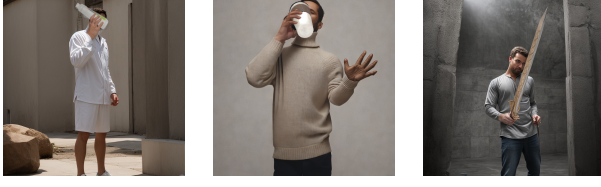


Figure 3. Failure cases for GraspDiffusion.

conditions, but centered on the hand region ($[s_h^i, d_h^i, o_h^i]$). For training, we only train the conditional encoders and fix the parameters for the original Stable Diffusion model, encouraging the encoders to be used with other diffusion models finetuned from Stable Diffusion, accounting to our pipeline’s style flexibility.

During inference, we further control the interaction by rectifying the cross-attention maps for the human and object. For the segmentation masks from the body pose (human : m^i , object : m_o^i , negative object : m_{no}^i), we assign different levels of strength for each maps to create an input attention matrix $A \in \mathbf{R}^{N_i \times N_t}$, where N_i and N_t are the number of image and text tokens. We assign a higher weight for the object masks due to their regional size differences. We then edit the cross attention layers so that it computes the output as $\text{softmax}(\frac{QK^T + wA}{\sqrt{d_k}})V$, where Q, K, V are the query, key and value embeddings, d_k is the dimensionality of Q and K , and w is a scalar weight that controls the total strength of user input attention. This encourages the image tokens in the segmented regions to adhere more to the corresponding text tokens, ensuring that the interaction captured by the body pose is well maintained. Following [2], we calculate w as

$$w = w' \cdot \log(1 + \sigma) \cdot \max(QK^T)$$

where w' is a user defined scalar. Details on model parameters and inference are provided in Table 2. Note that for the feature weights, we assigned a relatively low rate for the joint depth map, to ensure the result image doesn’t overfits to the SMPLX [7] mesh’s outline.

S2. Additional Results

We display failure cases for our pipeline in Figure. 3, where We note some failure cases, where the refined hand stands out from the image (left row), the hand shape tends to be uncanny (middle row), or where the complex object texture is not correctly preserved within the image (right row).

We also provide additional results for realistic, full-bodied human object interaction image generation in Figure. 4. We display that our model is capable of producing images of realistic humans interacting with the given object, with high diversity over human identity, body pose, camera angle, background, and other relevant scene context. The

results demonstrate our pipeline’s capability in creating realistic grasps for unseen objects.

References

- [1] Sketchfab. <https://sketchfab.com/>. Accessed: 2024-09-03. 3
- [2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2
- [3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021. 3
- [4] Hanwen Jiang, Qixing Huang, and Georgios Pavlakos. Real3d: Scaling up large reconstruction models with real-world images. *arXiv preprint arXiv:2406.08479*, 2024. 1
- [5] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhong-gang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *ArXiv*, abs/2302.08453, 2023. 1
- [6] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1
- [7] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2
- [8] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1
- [9] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13253–13263, 2021. 1
- [10] Purva Tendulkar, D’idac Sur’is, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21179–21189, 2022. 1
- [11] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 1
- [12] Y. Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-

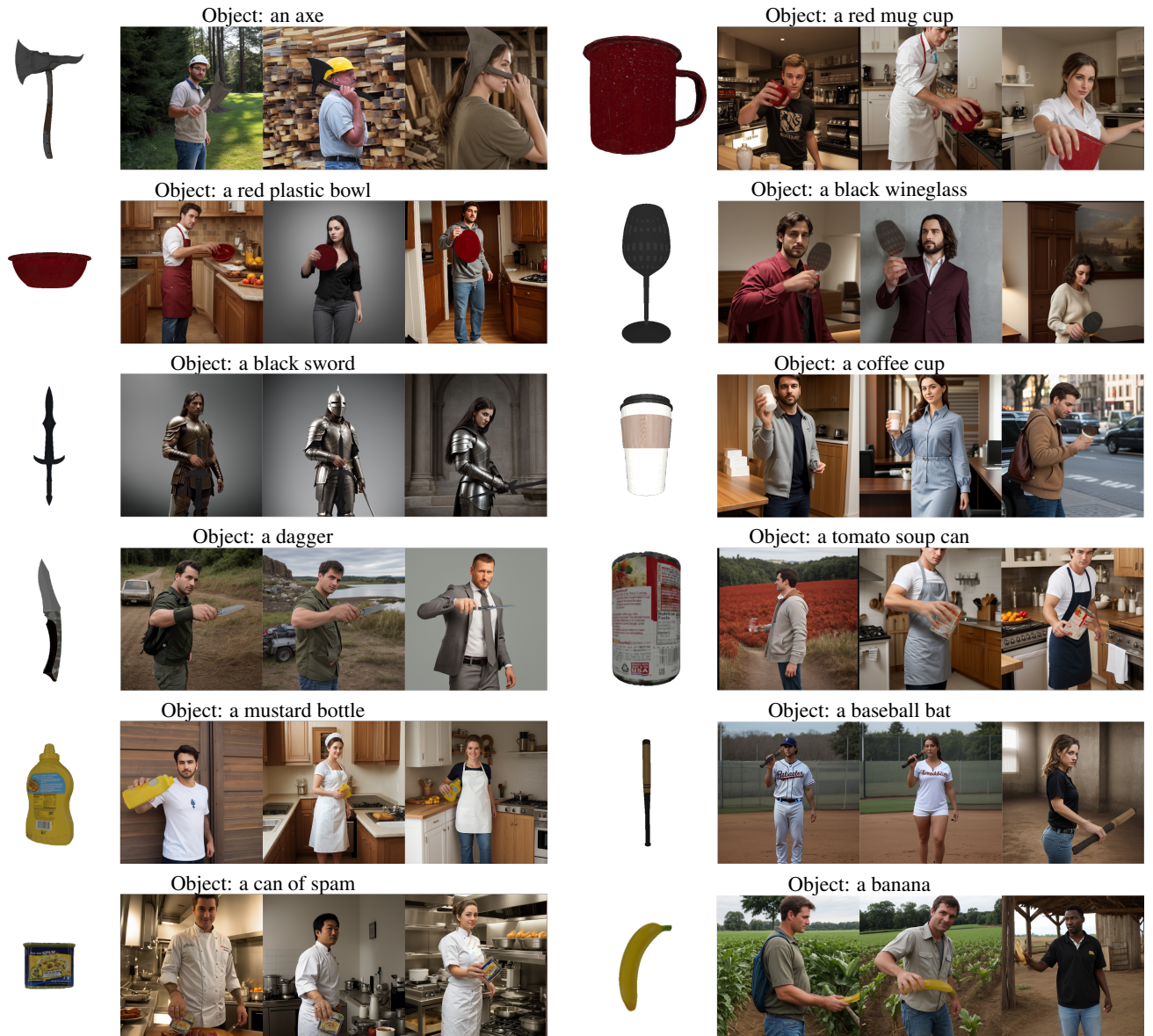


Figure 4. More results on synthesized images from a given object. Objects were gathered from the DexYCB dataset [3] and SketchFab [1]

body grasping with contact. In *European Conference on Computer Vision*, 2021. 1

- [13] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1
- [14] Yanzhao Zheng, Yunzhou Shi, Yuhao Cui, Zhongzhou Zhao, Zhiling Luo, and Wei Zhou. Coop: Decoupling and coupling of whole-body grasping pose generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2163–2173, 2023. 1