HOIDiffusion: Generating Realistic 3D Hand-Object Interaction Data

Supplementary Material

A. Architecture Details

There are two components in HOIDiffusion: the main diffusion model and the condition model. For the main branch, we base our diffusion model on Stable Diffusion [3] with changes of taking condition embeddings from three physical sources: normal map, hand skeleton projection, and segmentation. The conditions are added into the U-Net encoder, to shift feature maps of each encoding channel, similar to the method adopted in Adapter [2]. During training, we turn on the training of pretrained Stable Diffusion's decoder, with other parameters fixed (including other modules in U-Net and the text encoder). Model parameter details are provided in Table 1.

For condition models, we encode the condition images into different embedding levels for all feature channels in the diffusion model encoder. We adopt ResNet blocks to obtain the embeddings for each channel. Since there are three structural physical sources, we adopt the same model architecture for different conditions and utilize a weighted sum as the final condition outputs. Detailed information is shown in Table 2.

Parameter	Diffusion Model (512×512)
Latent Shape	4×64×64
Channels	320
Channels Multiple	[1, 2, 4, 4]
ResBlock Number	2
Context Dimension	768
Batch Size	8
Diffusion Steps	1000
Noise Scheduler	Linear
Learning Rate	10^{-5}
Optimizer	Adam

Table 1. Model architecture and training scheme (decoder) for main diffusion models. Input images are all resized to 512×512.

Parameter	Condition Model (512×512)
Input Channels	3×64
Output Channels	[320, 640, 1280, 1280]
ResBlock Number	2
Feature Weight (h,n,s)	[1,1,1]
Kernel Size	1
Batch Size	8
Learning Rate	10^{-5}
Optimizer	Adam

Table 2. Model architecture and training hyperparameters for condition model. (h,n,s) in Feature Weight represents summation weights for three conditions: (hand projection, normal map, segmentation).

B. More Results on HOI Generation

We provide more results on realistic hand-object-interaction image generation, with more diverse hand poses, object shapes, and object categories in Figure 3.

C. Object Appearance Control

In this section, we further explore the ability of HOIDiffusion to control the appearance of objects, with different colors or styles unseen in training data or descriptions. The results are shown in Figure 1. The results can demonstrate that our proposed HOIDiffusion is able to utilize the knowledge from the pretrained models, hence, when a novel appearance is provided, our model is still able to generate expected objects.

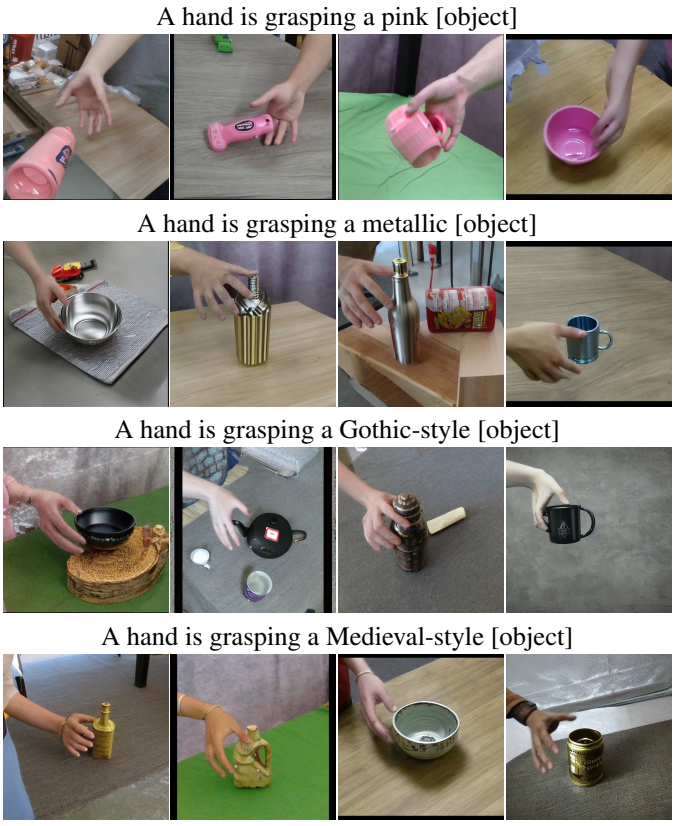

Figure 1. Generated images using different style texts to control object appearance.

D. Background Control

In this section, we present more generated hand-object-interaction images with various background prompts, from

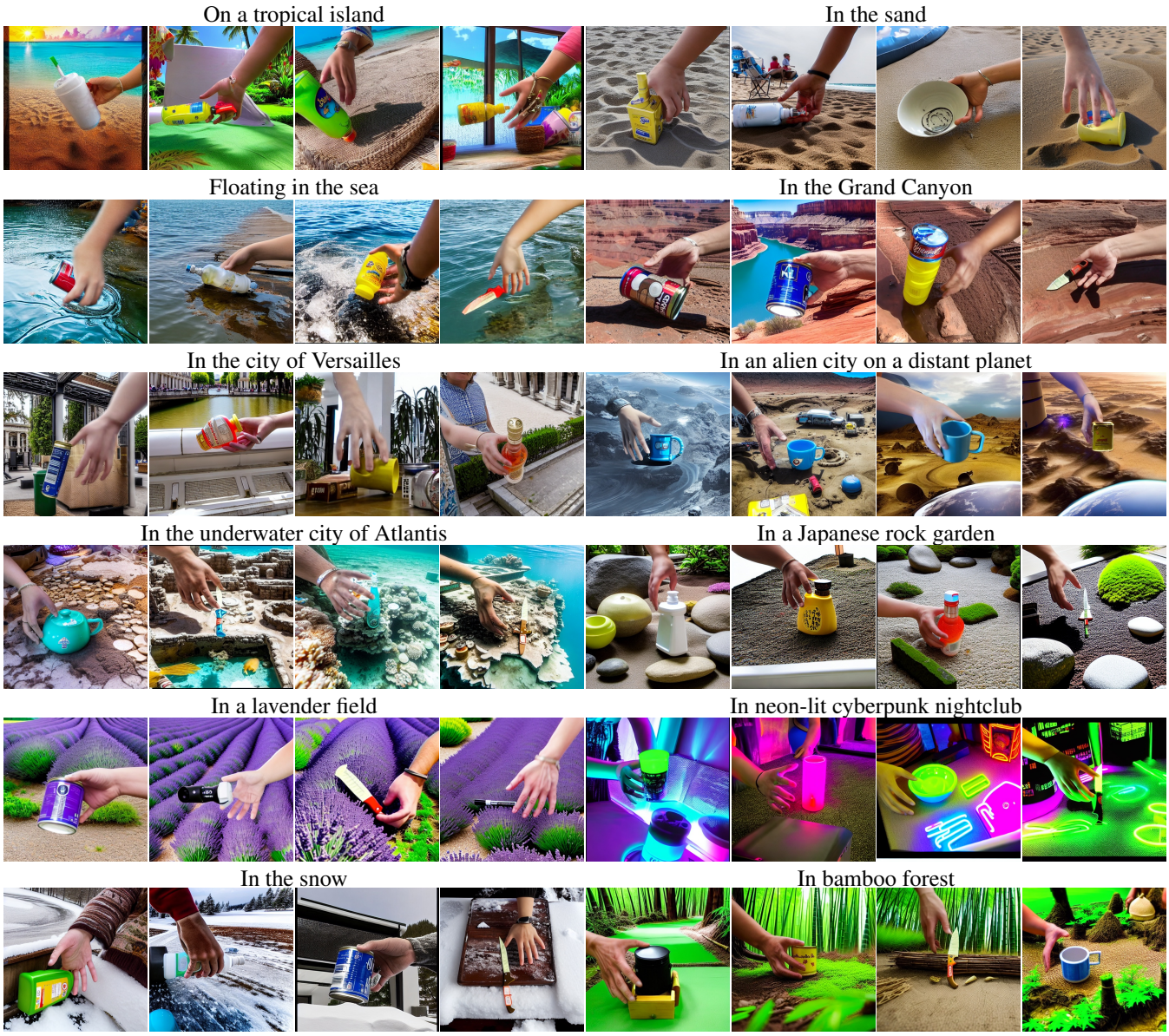

Figure 2. Generated images with more text prompts ranging from daily landscape to virtual scene.

everyday scenarios to special contexts. We refer readers to Figure 2.

The apparent differences in styles between Figure 3 and Figure 2 come from finetuning and regularization. In detail, during finetuning, the diffusion model quickly converges to the styles of the training dataset, which is more realistic. Prompts in HOI datasets (in our case, DexYCB [1]) lack detailed descriptions of background or in most cases, in a laboratory or the studio environment. Therefore, during inference, if we leave background descriptions vacant or use "on the table", the generated images are much closer to training cases, a bit blurred. To prevent the pretrained model from overfitting, we introduce the regularization module. This allows the model to utilize embedded knowledge from previous scale training, and thus when novel background prompts are provided, our HOIDiffusion is still able to depict diverse images as expected.

E. Social Impact

During the training process, we use the public well-known hand-object-interaction dataset DexYCB to supervise our training, which is licensed under CC BY-NC 4.0. Our proposed method generates human hands in images. The attributes of generated hands come from the learning knowledge from DexYCB and the pretrained model using LAION [4], which can be viewed as accumulated average results. Hence, Our synthesized HOI images don't incorporate any personal information or privacy.

Figure 3. More results on synthesized images with diverse object shapes and hand poses.

References

[1] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2

[2] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[4] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2