Hand-Object Interaction Image Generation Supplementary Material

Hezhen Hu[1] **Weilun Wang** [1*] **Wengang Zhou**[1,2†] **Houqiang Li**[1,2†]
[1]CAS Key Laboratory of GIPAS, EEIS Department
University of Science and Technology of China (USTC)
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{alexhu, wwlustc}@mail.ustc.edu.cn
{zhwg, lihq}@ustc.edu.cn

This supplementary material provides more details which are not included in the main paper due to space limitations. In the following, we first discuss the broader impact of our work, demonstrate the impact of the random seed. Then we perform more discussion on the robustness and generation object attribute. Finally, we present more qualitative results. For more video results, please refer to the attached video file in "Video.mp4".

A. Broader Impact

Our research aims to benefit application scenarios like AR/VR and online shopping, both of which impact the daily life of millions of users. The capability of generating the realistic hand-object image under the given interaction condition better optimizes the user experience. Furthermore, the research advance in this challenging task can provide useful cues in solving other complex human-centric generation scenarios. On the other side, our framework involves generating non-existent images, which may be misused in fake content creation. This may spread misinformation and lead to other privacy issues, which is opposite to our intent. A variety of regulatory and technical measures have been proposed to address this issue.

B. Impact of the Random Seed

The impact of different random seeds on FID and LPIPS is shown in Figure 1. We repeat the experiments five times on HO3Dv3 and DexYCB datasets, respectively. It can be observed that the variance of these metrics is relatively small. Our method achieves stable performance on the testing set, which is relatively unaffected by the random seed.

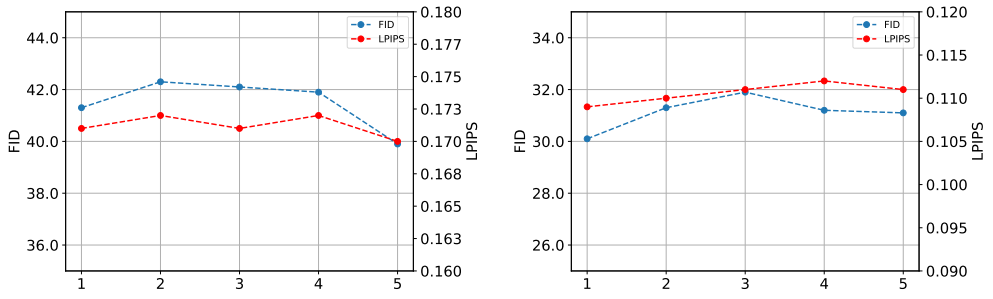
C. More Discussion on the Object Attribute

Symmetric characteristics. Our model does not contains the assumption on the symmetric characteristic of the object. Therefore, our model performs fairly well regardless of whether the object is symmetric. We visualize some samples in Figure 2 (a).

Object size. Since our model does not contain the assumption on the object size, our model performs consistently well independently of the object size. The drill and banana exhibit large differences in object size and appearance. As shown in Figure 2 (b), the hand-drill and hand-banana images generated by our model are consistent with their ground-truth images.

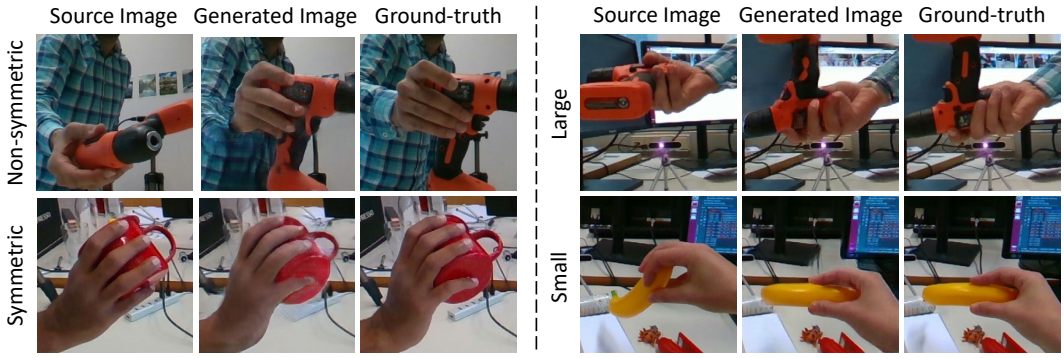
[*]Contribute equally with the first author.
[†]Corresponding authors: Wengang Zhou and Houqiang Li.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) HO3Dv3 dataset. (b) DexYCB dataset.

Figure 1: Impact of different random seeds on HO3Dv3 and DexYCB datasets. We demonstrate its impact on FID and LPIPS metrics. The experiments are repeated 5 times.



(a) Symmetric characteristics (b) Object size

Figure 2: Impact of different object attributes, *i.e.*, symmetry and object size. It demonstrates that our method performs well on object with different attributes.

D. More Qualitative Results

We further visualize more generated results on HO3Dv3 and DexYCB datasets. As shown in Figure 3, it can be observed that our method generates the hand-object image with better consistency with the target image. Specifically, the generated hand exhibits correct structure and keeps the same identity as the source, while the generated object demonstrates vivid fine-grained details.

Besides, we also demonstrate another application, *i.e.,* image animation. Its goal is to generate realistic videos under a sequence of target postures while preserving the source identity. For better clarification, please refer to "Video.mp4" for more details. Previous methods usually suffer flickering or failure on preserving object details. In contrast, our framework can generate more fluent video with fine-grained details.

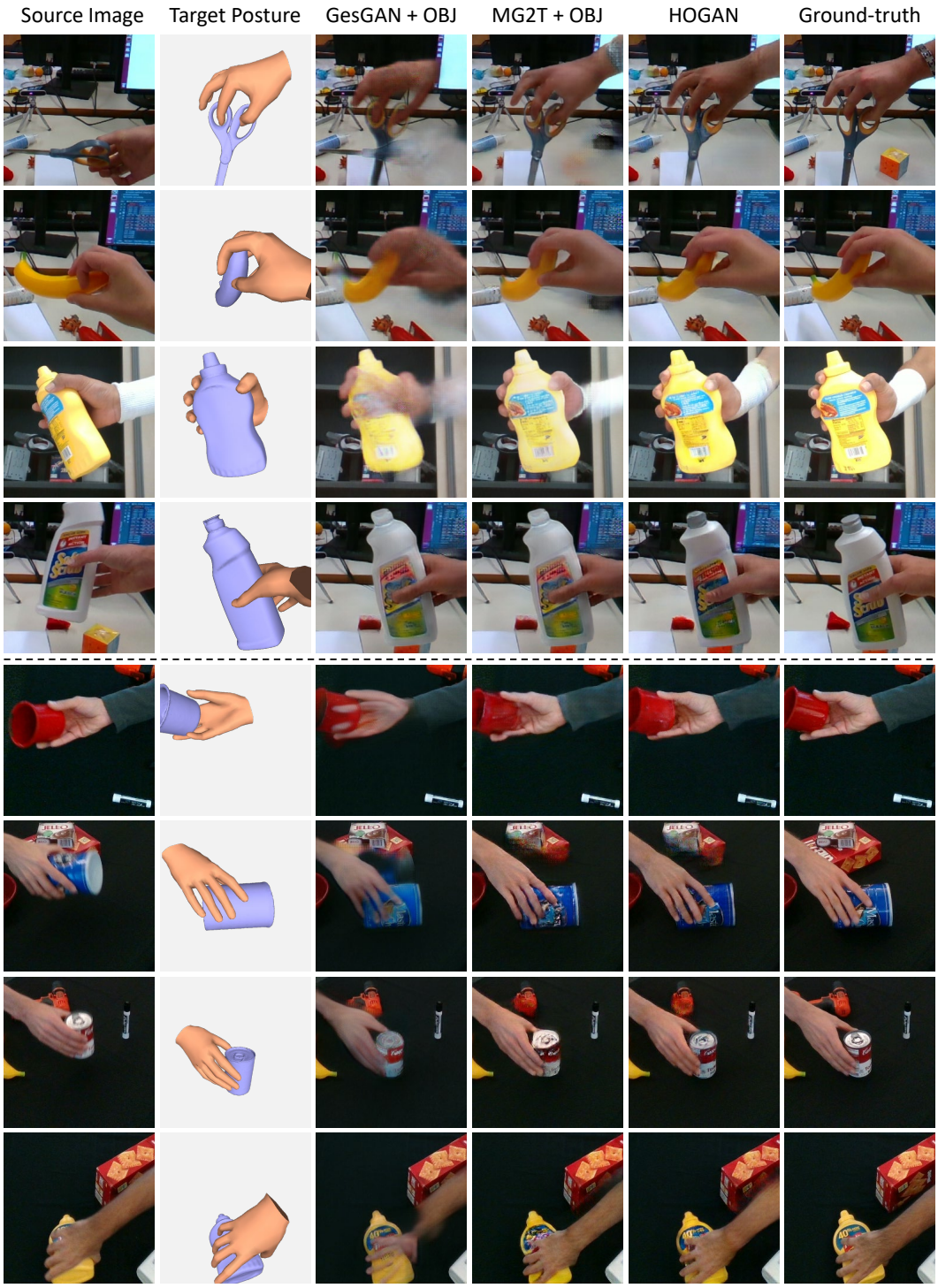


Figure 3: More qualitative comparison with baselines, including GestureGAN + OBJ and MG2T + OBJ, on the HO3Dv3 (top) and DexYCB (bottom) datasets.