GraspDiff: Grasping Generation for Hand-Object Interaction with Multimodal Guided Diffusion —Supplementary Material—

Binghui Zuo, Zimeng Zhao, Wenqian Sun, Xiaohan Yuan, Zhipeng Yu, and Yangang Wang, *Member, IEEE*

A Network Architecture

A.1 Encoders for Multimodal Conditions

We first provide details about the multimodal encoders for various modalities. In our proposed GraspDiff, all 3D object point clouds, contact maps, part maps, and 2D images are considered as the guidance for grasping generation. For the vanilla version, *i.e*, given known object point clouds, we hope that GraspDiff could generate multiple candidates while ensuring that the obtained interactions are reasonable and plausible. For this, we devise a PointNet-based encoder that embeds the 3D point clouds into latent vectors. Putting these extracted condition features into the denoising model, they are regraded as constraints in the denoising process. Besides this basic guidance, for the modality of contact maps and part maps, considering both of them are represented as vectors with the consistent number of point clouds, we also devise similar PointNet-based encoders to extract their information. For the 2D images, a trained ResNet18 [1] followed by a linear layer is used to extract corresponding information and align feature dimensions respectively. It should be noted that except for point clouds, the other three modalities are selective, which are treated as an additional condition to guide the generation. The representation of these modalities and the adopted feature encoders are illustrated in Fig. 1 (a).

A.2 Denoising Model

We employ a transformer-based model to obtain denoising targets. As stated in the main paper, we directly estimate the desired latent vector z_0 from the noisy z_t, which is sampled from the normal distribution. The architecture of this block is shown in Fig. 1 (b), consisting of attention and MLP modules. Generally speaking, after passing the inputs through the self-attention layer, both the obtained intermediate feature F and extracted condition information c are fed into a cross-attention module. Similarly to the original transformer structure [2], an MLP-based feedforward module is also integrated into each transformer block. In our experiments, we totally adopt 6 transformer blocks, *i.e* $N_{attn} = 6$. At the end of the estimator, we connect a linear layer to ensure that the dimension of estimated z_0 is consistent with the latent space. Fig. 1 (c) further illustrates the implementation of Equation (7) in the main paper, where

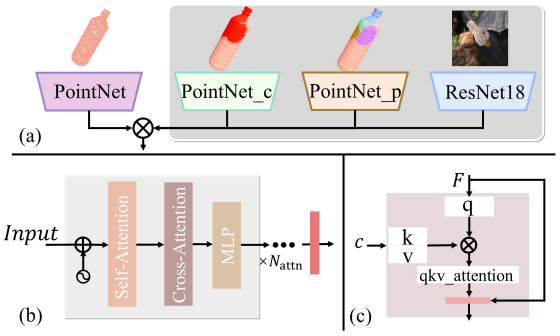

Fig. 1. Network architecture of GraspDiff. (a) Encoders for multimodal conditions; (b) Attention module in denoising network; (c) Cross-attention mechanism.

the key and value are obtained from condition features, while the query is obtained from intermediate denoising features F.

B Training Details

In this section, we provide more training details, including the detailed process for preparing multimodal conditions and the loss functions involved in Equation (5).

B.1 Training Data

Our framework is compatible with multimodal controllability. In addition to the training dataset discussed in the main paper, we provide more details on how to prepare these conditions from the dataset: **i) Contact maps**: Inspired by [3], we use a virtual capsule at each object point to model contact. If any hand vertex lies inside the capsule, the object point will be labeled as in contact. In this way, we obtain contact labels for each object point, which are used to represent contact maps. More details can be found in [3]. **ii) Part maps**: The K-Nearest Neighbors are performed between the point clouds of the hand and the object to find the nearest hand vertex for each object point. As the whole hand is divided into 16 separate parts, we can further infer which part of the hand is closest to each object point. We use a 16-dimensional one-hot vector to store this corresponding information. Considering that the part maps would be meaningful within the contact

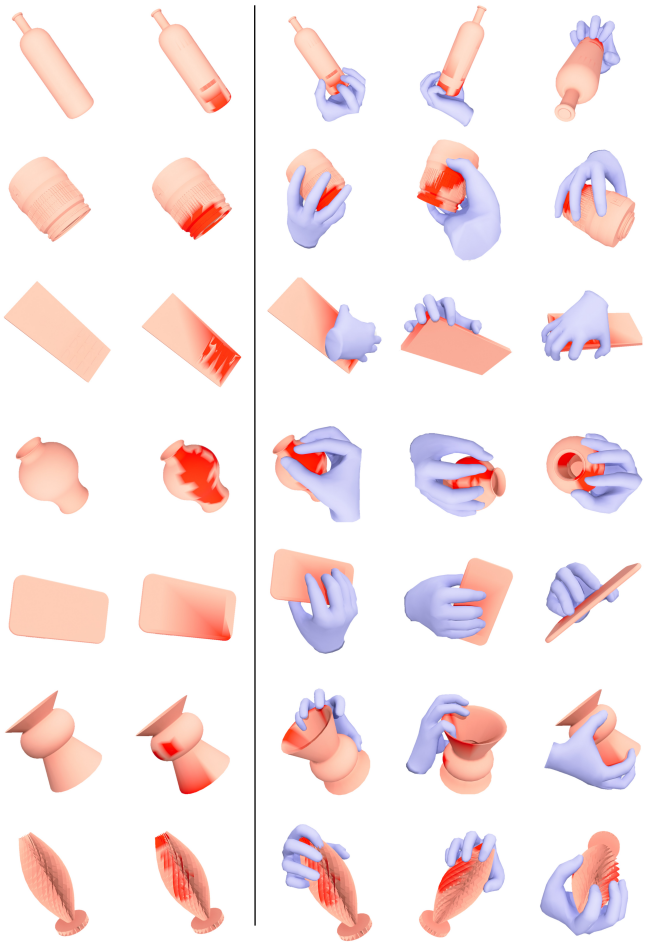

Fig. 2. Visualizations of the generation that are jointly guided by object point clouds and contact maps. We display three views for each generated hand grasp.

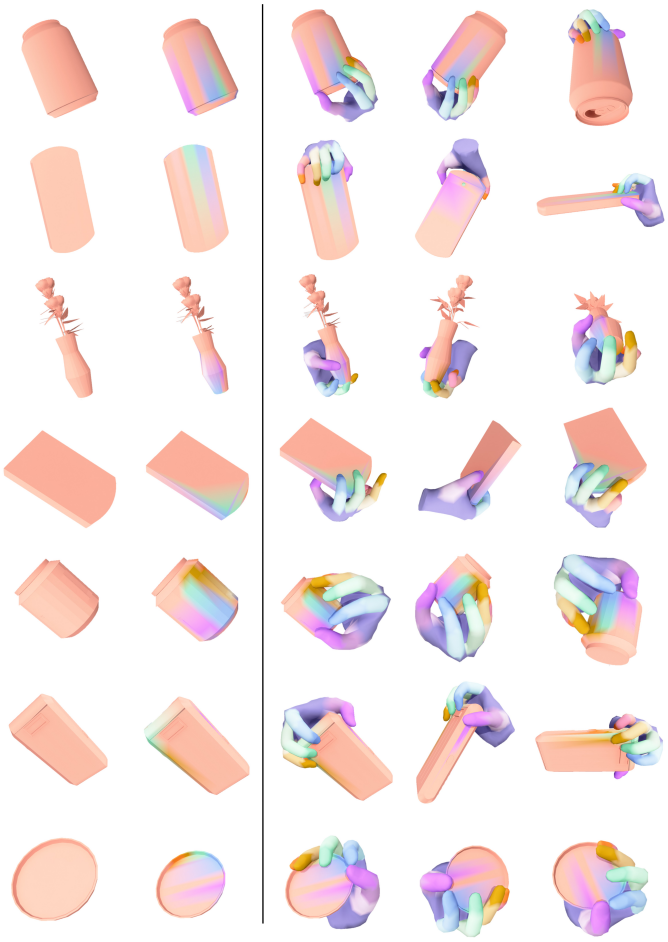

Fig. 3. Visualizations of the generation that are jointly guided by object point clouds and part maps. We display three views for each generated hand grasp.

map region, for those non-contact object points, we set the corresponding vector to 0. **iii) 2D images**: Benefitting from the open-source dataset, we directly obtain 2D images from the dataset.

B.2 Training Loss

MSE Loss. MSE loss is adopted to supervise the generated hand grasps $\hat{H}$. The following equation describes the implementation.

$$L_{MSE} = \left\| H - \hat{H} \right\|_2^2. \tag{1}$$

Chamfer Distance Loss. In practice, we have found that relying solely on MSE loss makes it difficult to ensure global consistency between the generated grasping posture and the ground truth, including global rotations and global translations. Therefore, we introduce additional geometric constraints to achieve more appropriate generations.

$$\begin{aligned} L_{CD}\left(\mathbf{V}_h, \hat{\mathbf{V}}_h\right) = & \sum_{x_i \in \mathbf{V}_h} \min_{x_j \in \hat{\mathbf{V}}_h} \|x_i - x_j\|_2^2 \\ & + \sum_{x_j \in \hat{\mathbf{V}}_h} \min_{x_i \in \mathbf{V}_h} \|x_i - x_j\|_2^2, \end{aligned} \tag{2}$$

where $\mathbf{V}_h$ and $\hat{\mathbf{V}}_h$ denote the hand vertices from ground truth and generated grasp respectively. The first term finds the nearest neighbor from the ground-truth vertices to the generated hand vertices. While the second item is the opposite. Consistent with the conclusion summarized in [4], although chamfer distance loss is implemented simply, it produces reasonable and high-quality generation in practice.

Penetration Loss. Similar to the above purpose, the penetration loss is also considered to supervise the training process to improve the plausibility of generation. It is formulated as:

$$L_{Pene.} = \frac{1}{|\mathbf{P}_{\text{in}}^o|} \sum_{p \in \mathbf{P}_{\text{in}}^o} \min_i \left\| p - \hat{\mathbf{V}}_h^i \right\|_2^2, \tag{3}$$

where $\mathbf{P}_{\text{in}}^o$ denotes the object points in penetration state, and penetration loss is devised to minimize their distances to their closest hand vertices $\hat{\mathbf{V}}_h$. Ideally, $L_{Pene.}$ should tend towards 0, which means no penetration existed.

C More Results

Besides the visualizations demonstrated in the main paper, we report additional results here.

Fig. 2, Fig. 3 and Fig. 4 further report the generated grasps with multimodal conditions as guidance. Visualizations in Fig. 2 are jointly conditioned by object point clouds

Fig. 4. Visualizations of the generation that are jointly guided by object point clouds and 2D images. We display two views for each generated hand grasp.

and contact maps. Visualizations in Fig. 3 are jointly conditioned by object point clouds and part maps. Visualizations in Fig. 4 are jointly conditioned by object point clouds and 2D images, which could also be approximated as a monocular image reconstruction task.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[3] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *CVPR*, 2021, pp. 1471–1481.

[4] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *CVPR*, 2017, pp. 605–613.