

ClickDiff: Click to Induce Semantic Contact Map for Controllable Grasp Generation with Diffusion Models

Supplementary Materials

Anonymous Authors

A OVERVIEW

In the supplementary paper, we present human evaluation (Sec. B), our model architecture (Sec. C), the algorithm to generate SCM (Sec. D), and the additional results of our method on GRAB and ARCTIC datasets (Sec. E), respectively.

B HUMAN EVALUATION

In this study, we further validate the perceptual quality and stability of the grasps generated by our method through a comprehensive user study following [2–4, 6]. We select a total of six objects from the GRAB dataset [5] for evaluation. For each object, our assessment incorporate three randomly selected ground truth grasps from the dataset, alongside three grasps generated by benchmarked methods and three by our approach. Participants are asked to rate the quality of each grasp based on its naturalness and the stability of holding the object using a five-point scale ranging from strongly disagree (1) to strongly agree (5). Fig. 1 shows the score distribution and Tab. 1 summarizes the average ratings. Notably, our method demonstrates superior performance in both naturalness and stability compared to other evaluated methods, but still lags behind ground truth grasps.

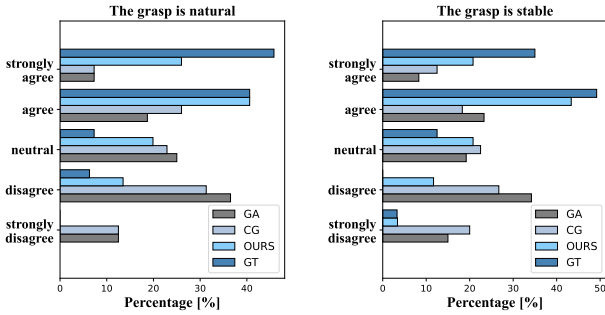


Figure 1: Grasp human studies score distribution. The distribution of scores shows that our method achieves comparable performance to the GT in both naturalness and stability.

Table 1: Grasp human study statistics. While the gap between ours and the GT exists, our method performs better than GraspTTA [2] and ContactGen [4] in terms of naturalness and stability.

GRAB	GraspTTA [2]	ContactGen [4]	Ours (SCM)	GT
Natural	2.76	2.77	3.67	4.13
Stability	2.72	2.84	3.79	4.26

C MODEL ARCHITECTURE

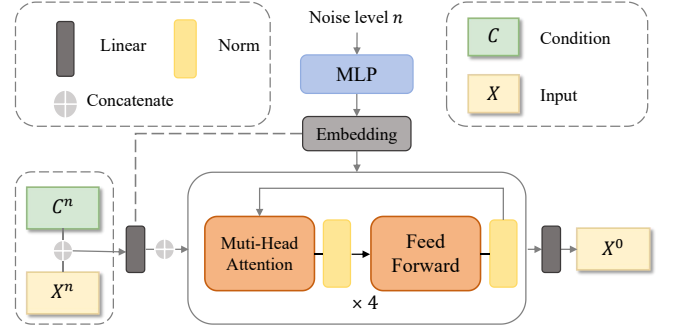


Figure 2: Model architecture of the denoising network in a single step of the reverse diffusion process.

D ALGORITHM

Algorithm 1 Semantic Contact Map (SCM) Generation

Input: Hand point cloud \mathcal{H} with $N_{\mathcal{H}}$ points, object point cloud \mathcal{O} with $N_{\mathcal{O}}$ points, finger range definitions $\mathcal{R} = \{R_{\text{thumb}}, R_{\text{index}}, R_{\text{middle}}, R_{\text{ring}}, R_{\text{little}}\}$, distance thresholds $\tau_{\text{threshold}}$.

Output: Semantic Contact Map matrix of size $N_{\mathcal{O}} \times 5$.

// Contact Map Generation

for $o = 1$ to $N_{\mathcal{O}}$ do

for $h = 1$ to $N_{\mathcal{H}}$ do

$d_{oh} \leftarrow$ Compute Euclidean distance from \mathcal{O}_o to \mathcal{H}_h ;

end for

$D_o \leftarrow \min_{j \in \{1, 2, 3, \dots, N_{\mathcal{H}}\}} (d_{oj})$;

$\text{Index}_o \leftarrow \text{argmin}_{j \in \{1, 2, 3, \dots, N_{\mathcal{H}}\}} (d_{oj})$;

end for

Normalize D to get the contact map C ;

// Binary Contact Map Generation

for $o = 1$ to $N_{\mathcal{O}}$ do

$B_o \leftarrow$ threshold contact map C_o with $\tau_{\text{threshold}}$;

end for

// Semantic Contact Map Matrix Construction

for $o = 1$ to $N_{\mathcal{O}}$ do

for each finger f in \mathcal{R} do

if Index_o in f then

$\text{SCM}_{of} \leftarrow B_o$;

end if

end for

end for

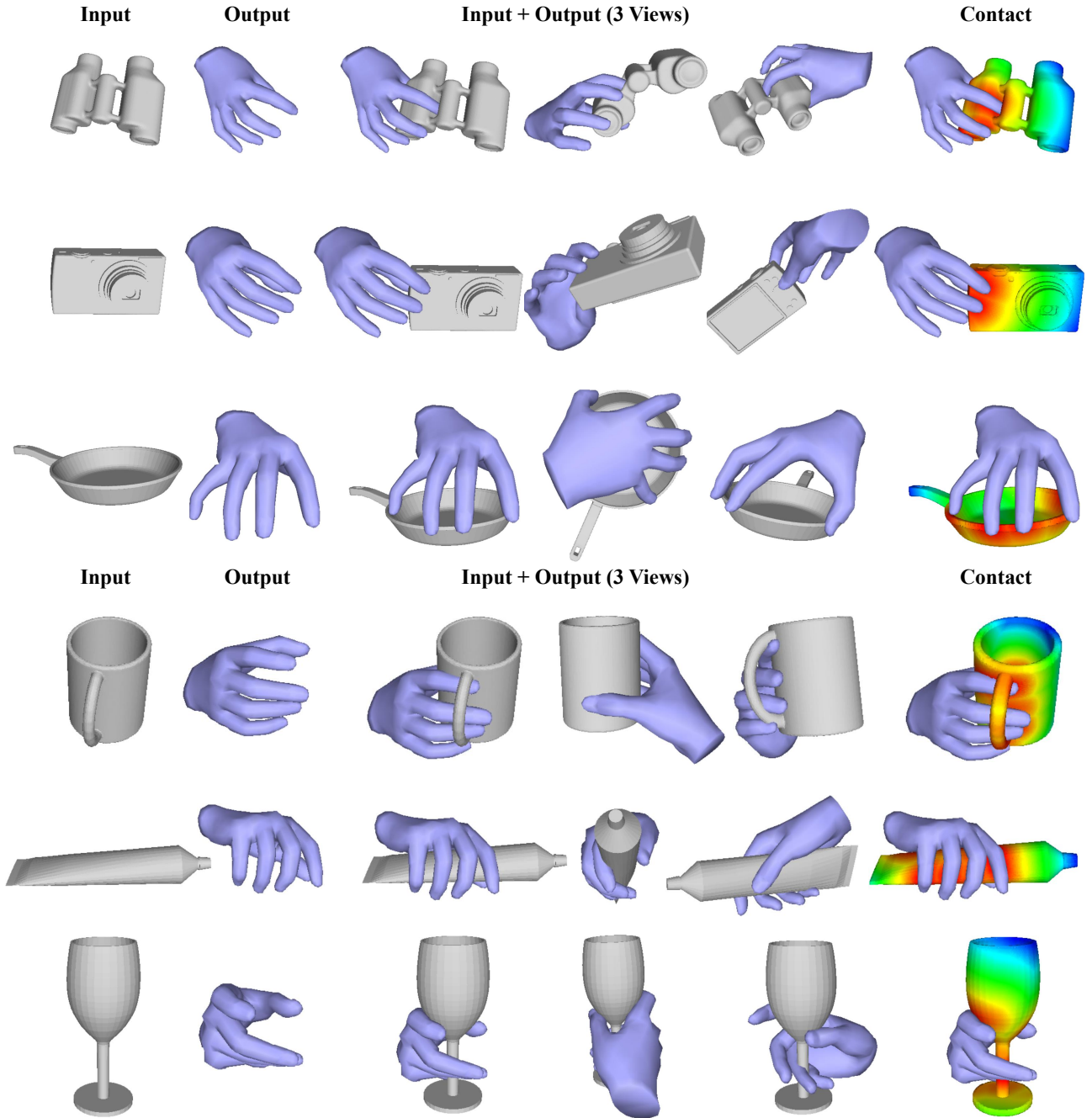


Figure 3: Additional results of generated grasps given objects. Every result is shown in a row with input object, output hand mesh, both input and output in 3 views and in contact.

E ADDITIONAL RESULTS

REFERENCES

- [1] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12943–12954.
- [2] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. 2021. Hand-Object Contact Consistency Reasoning for Human Grasps Generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 11087–11096.
- [3] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. 2020. Grasping Field: Learning Implicit Representations for Human Grasps. In *2020 International Conference on 3D Vision (3DV 2020)*. 333–344.
- [4] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. 2023. ContactGen: Generative Contact Modeling for Grasp Generation. In *Proceedings*

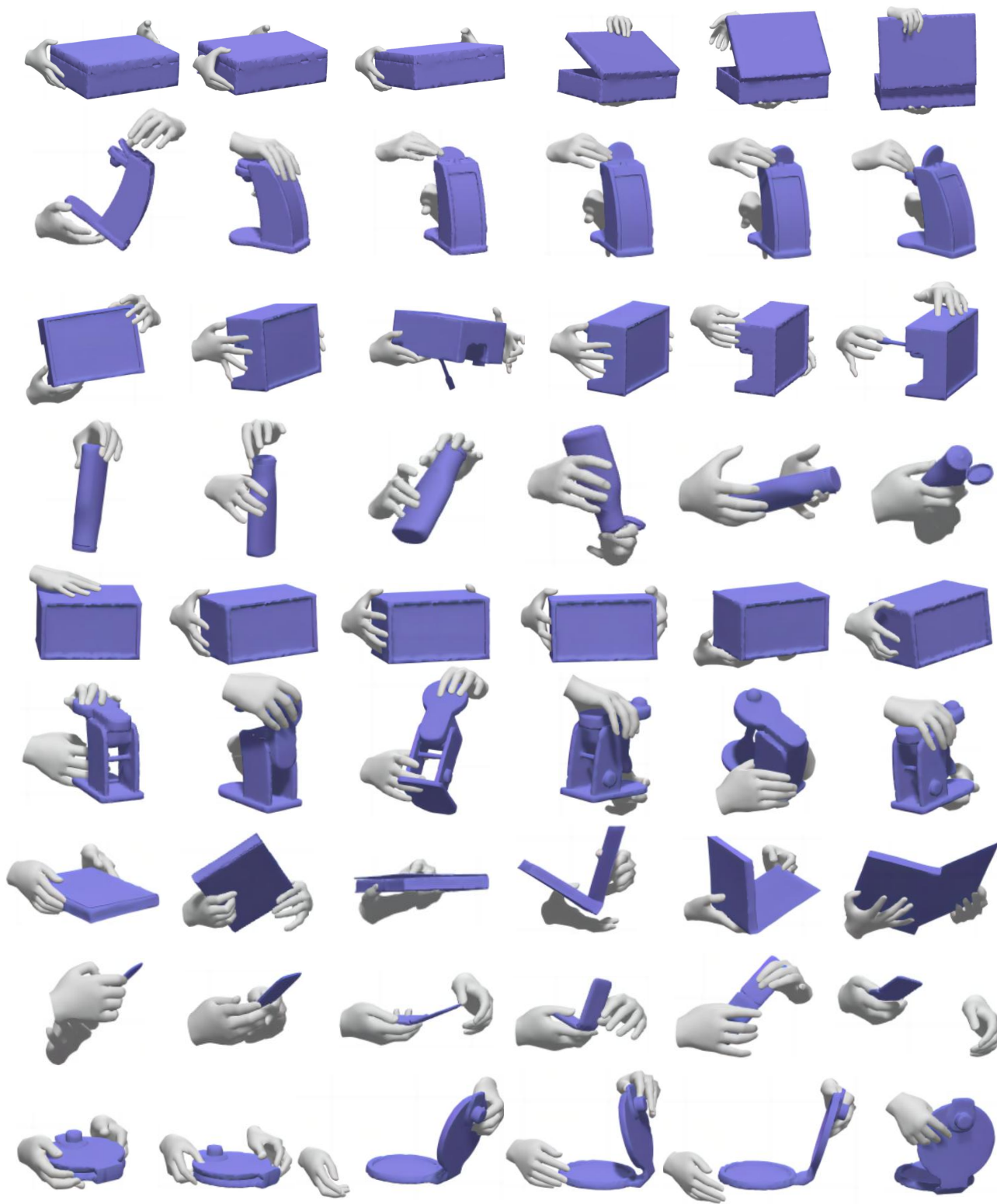


Figure 4: visualization results of generated grasps on the ARCTIC dataset[1].

of the IEEE/CVF International Conference on Computer Vision. 20609–20620.

[5] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision*.

[6] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. 2022. SAGA: Stochastic Whole-Body Grasping with Contact. In *Computer Vision – ECCV 2022*. 257–274.