Supplementary Materials: Task-Oriented Human Grasp Synthesis via Context- and Task-Aware Diffusers

An-Lun Liu[1] Yu-Wei Chao[2] Yi-Ting Chen[1*]

[1]National Yang Ming Chiao Tung University [2]NVIDIA

1. Implementation Details

1.1. Hyperparameters

All models use AdamW [5] as the optimizer, set the learning rate as 5×10^{-3}, and apply CosineAnnealingLR [4] as the learning rate scheduler. We set batch sizes as 84 and 96 for ContactDiffuser and GraspDiffuser respectively.

1.2. Model Architecture

Our models adopt the Transformer [7] encoder layer. We design individual layers for them. Contactdiffuser is composed of 8 layers, and GraspDiffuser is composed of 4 layers. The architecture can be found in Figure 1.

2. More Experimental Results

This section performs more detailed experiments on the proposed method and baselines.

2.1. Analysis on Different Penetration Volume for Grasp Prediction Quality

To further evaluate the quality of generated human grasps, we set different thresholds for QR. The penetration from is ranged from $1 \times 10^{-6}\,\mathrm{cm}^3$ to $4 \times 10^{-6}\,\mathrm{cm}^3$ and the simulation displacement is ranged from 1 cm to 3 cm. The comparsions are shown in Figs. 2 to 4. In most combinations, our method outperforms all other baselines. However, in the strictest setting, where PV is smaller than $1 \times 10^{-6}\,\mathrm{cm}^3$, SceneDiffuser shows better performance.

2.2. Performance Analysis on the Impact of Object Size for Stacking

We conduct a comprehensive study on stacking. There are a total of six bricks in the testing set. We divide them into 2 categories, small and large. The bricks F, I, and K belong to the small object set. The bricks N, R, and V belong to the large object set. As discussed in the main paper, the two sets have a performance gap. In Table 1, we show the quantitative results for each of them. All methods struggle to generate task-oriented grasping poses for small bricks, leading to large **Init OPP**. We show the failure case for small bricks in the main paper. This suggests the future direction of generating realistic contact maps for tiny objects.

3. Additional Qualitative Results

This section displays more qualitative results of the proposed method and baseline algorithms. Our method outperforms all the baselines quantitatively. ContactDiffuser generates a more realistic contact map, and GraspDiffuser can synthesize more natural and stable human grasps.

3.1. Visualization of Predicted Human Grasps

In Figs. 5 to 9 we show the synthesized grasping poses of our method and baselines. We also show the failure cases from our method in Figs. 10 to 12. In the Figs. 13 to 28 we show the synthesized grasping poses of our method for all objects in the testing set.

3.2. Contact Maps

In Figs. 29 to 55, we show the contact maps synthesized by ContactDiffuser and ContactCVAE [3] and corresponding human grasps synthesize by GraspDiffuser.

3.3. Limitation

Figs. 30 and 39 show the predicted contact map and grasping poses for **bowl** and **tape**. Our method struggles to predict an appropriate contact map for them, which leads to severe penetration or unstable grasp. ContactCVAE [3] can also not predict realistic contact maps for them. As mentioned in the main paper, ContactCVAE tends to predict over-smooth results. Noticeably, GraspDiffuser can generate more natural and stable grasping poses with the output of ContactCVAE.

*Corresponding Author

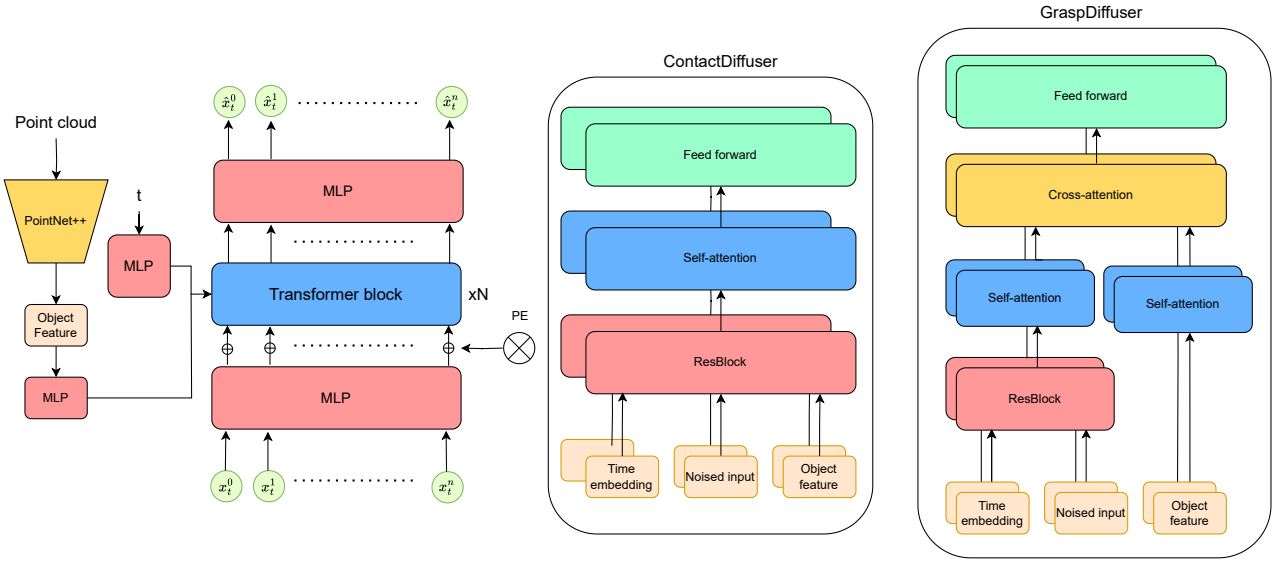

Figure 1. The architectural designs of the proposed context- and task-aware diffusers.

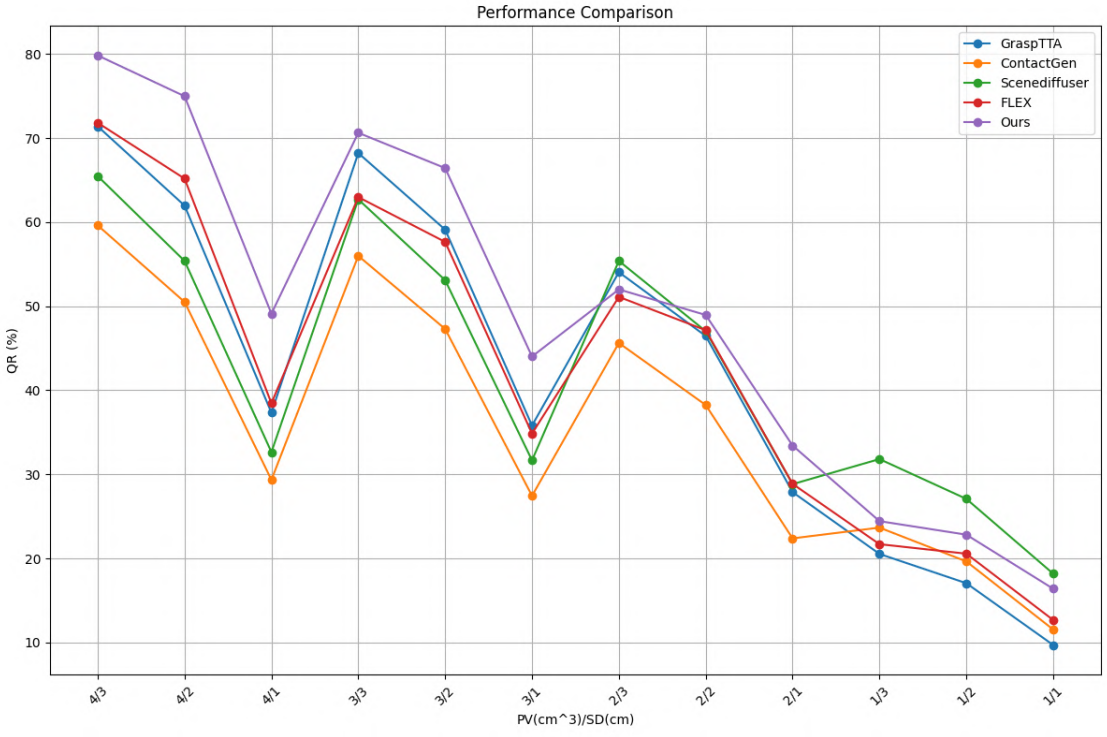

Figure 2. Different thresholds of QR for **Placing**.

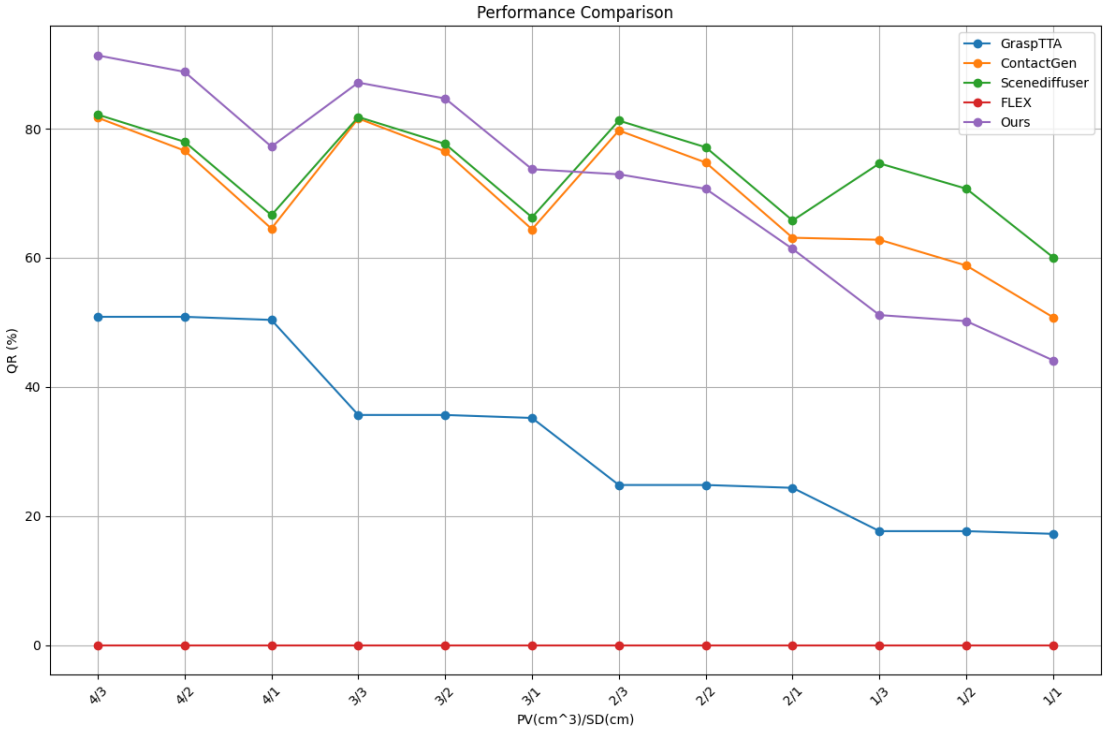

Figure 3. Different thresholds of QR for **Stacking**.

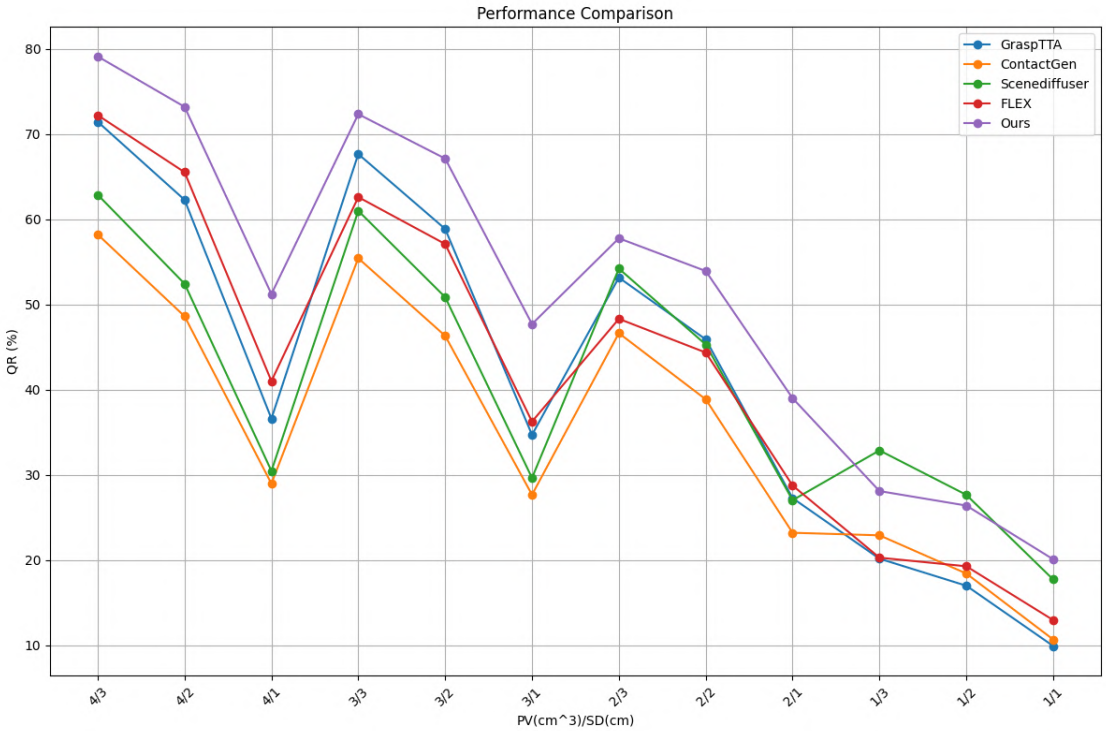

Figure 4. Different thresholds of QR for **Shelving**.

Table 1. Ablation study of our method. We denote GraspDiffuser as **GD**, ContactDiffuser as **CD**, and ContactCVAE as **CC**. The bricks F, I, and K belong to the small object set. The bricks N, R, and V belong to the large object set.

Brick	Method	PV↓	SD↓	CR(%)↑	QR(%)↑	DS ↑	Init OPP(%)↓	Goal OPP(%)↓	TS ↑
F	Simple-GD	0.31	2.55	81.62	63.00	**57.25**	28.03	15.36	0.383
	CC [3] + GD	0.32	2.34	**84.75**	69.50	49.58	**24.94**	**7.34**	**0.483**
	CD + GD	**0.25**	**2.03**	83.50	**71.25**	46.41	26.80	7.35	**0.483**
I	Simple-GD	0.34	0.70	98.62	92.75	**57.25**	30.68	11.11	0.571
	CC [3] + GD	0.47	**0.52**	**99.62**	**96.37**	48.70	**24.62**	**3.18**	**0.703**
	CD + GD	**0.30**	0.75	95.87	94.25	53.44	27.06	6.29	0.644
K	Simple-GD	0.60	1.06	**96.62**	87.00	**57.32**	30.64	13.72	0.520
	CC [3] + GD	0.57	**0.87**	96.50	**90.50**	50.03	**26.63**	**5.45**	**0.627**
	CD + GD	**0.38**	1.06	90.50	86.50	50.60	29.44	6.22	0.572
N	Simple-GD	**0.67**	1.88	95.50	76.00	**67.79**	22.17	5.40	0.559
	CC [3] + GD	1.69	1.58	99.62	**77.75**	57.76	9.54	**2.00**	0.689
	CD + GD	2.25	**0.98**	**100.00**	76.12	48.99	**3.12**	3.86	**0.709**
R	Simple-GD	**0.59**	1.87	97.85	73.37	**69.70**	21.83	7.05	0.533
	CC [3] + GD	1.33	0.87	99.87	**90.37**	54.49	8.27	**1.16**	0.819
	CD + GD	1.63	**0.70**	**100.00**	88.00	45.93	**1.59**	2.04	**0.848**
V	Simple-GD	**0.68**	1.81	99.50	74.25	**68.41**	18.46	6.25	0.567
	CC [3] + GD	1.52	0.75	**100.00**	**91.00**	54.11	7.33	**0.74**	0.837
	CD + GD	1.74	**0.64**	**100.00**	89.75	46.15	**1.64**	1.28	**0.871**

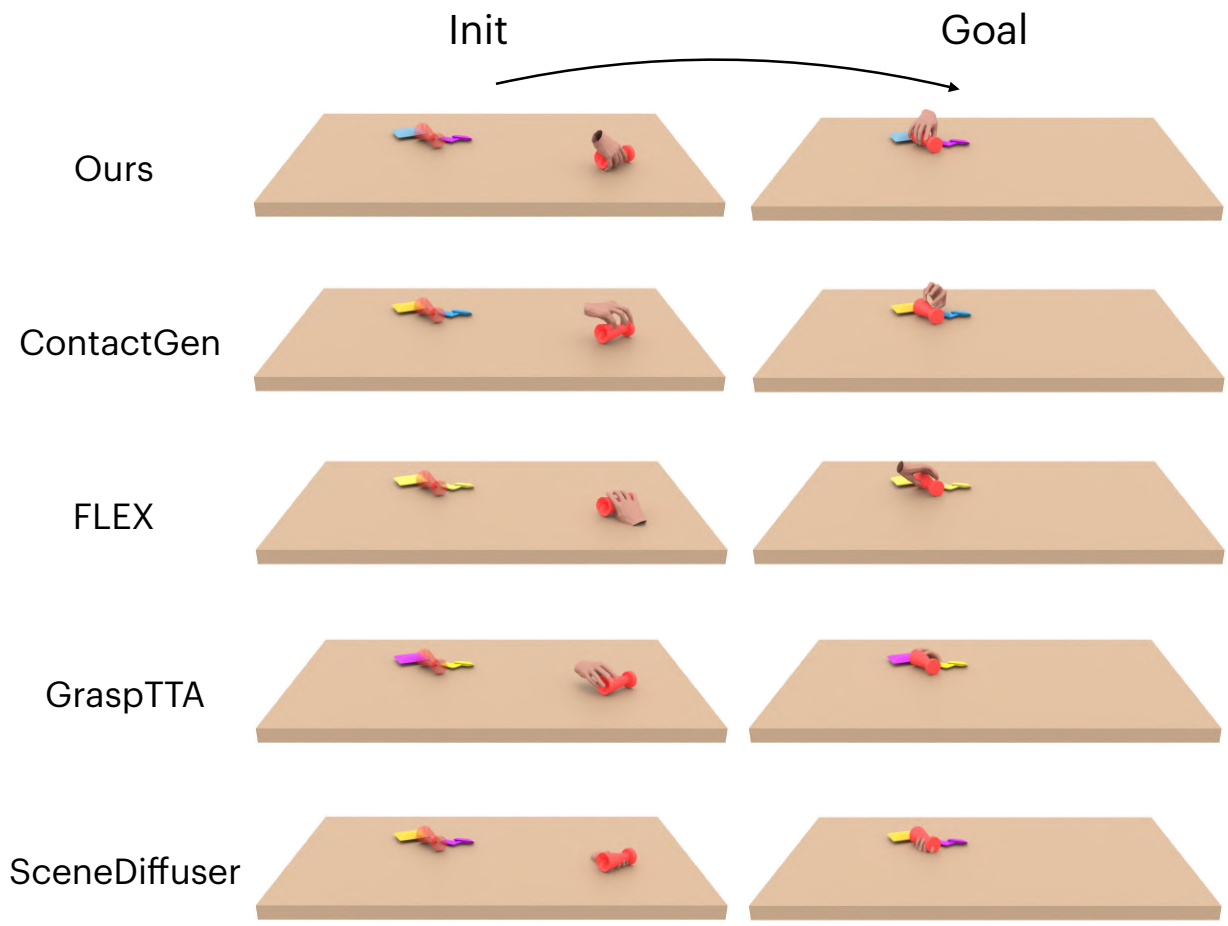

Figure 5. Visualization of predicted human grasps for **trophy** from Ours and ContactGen [3], FLEX [6], GraspTTA [2], SceneDiffuser [1].

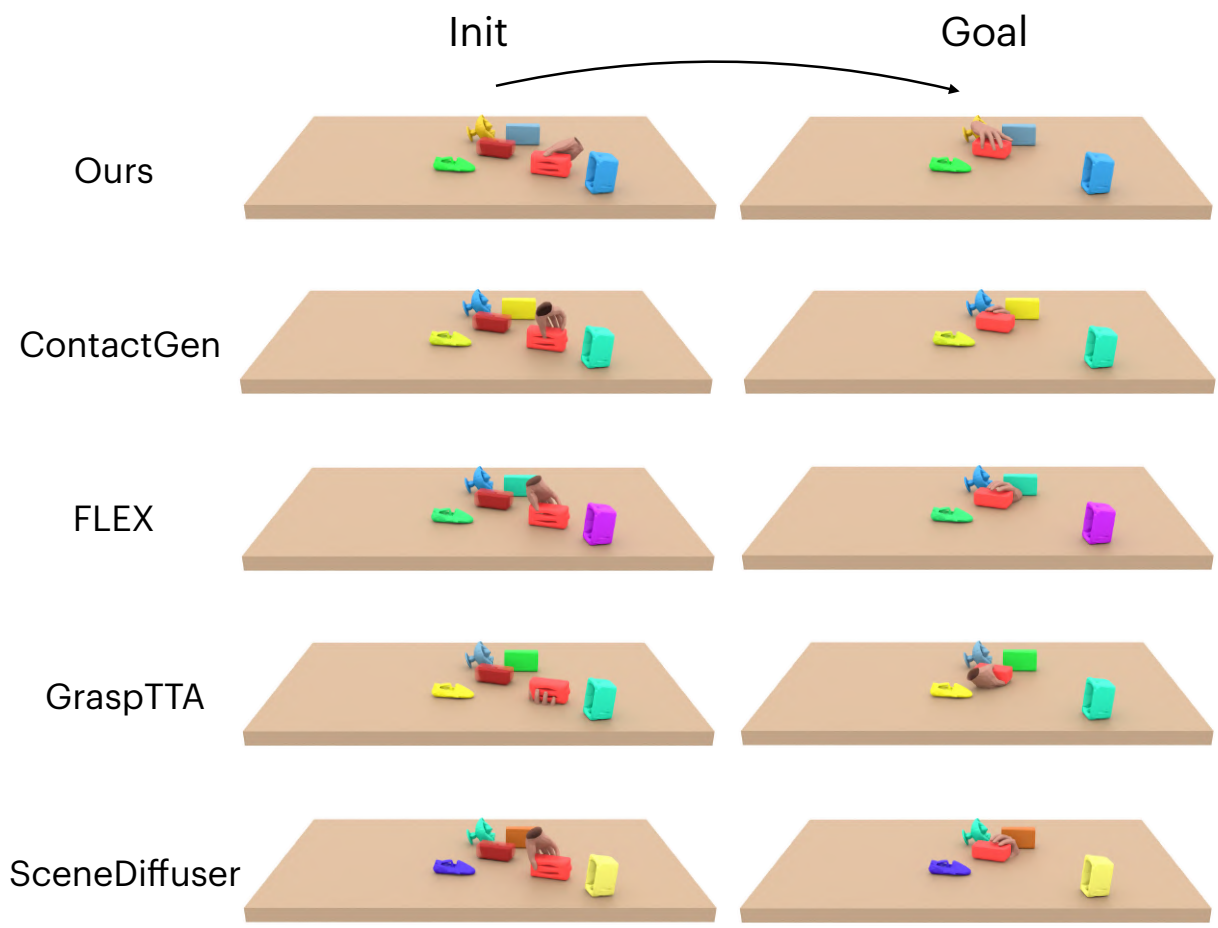

Figure 6. Visualization of predicted human grasps for **toaster** from Ours and ContactGen [3], FLEX [6], GraspTTA [2], SceneDiffuser [1].

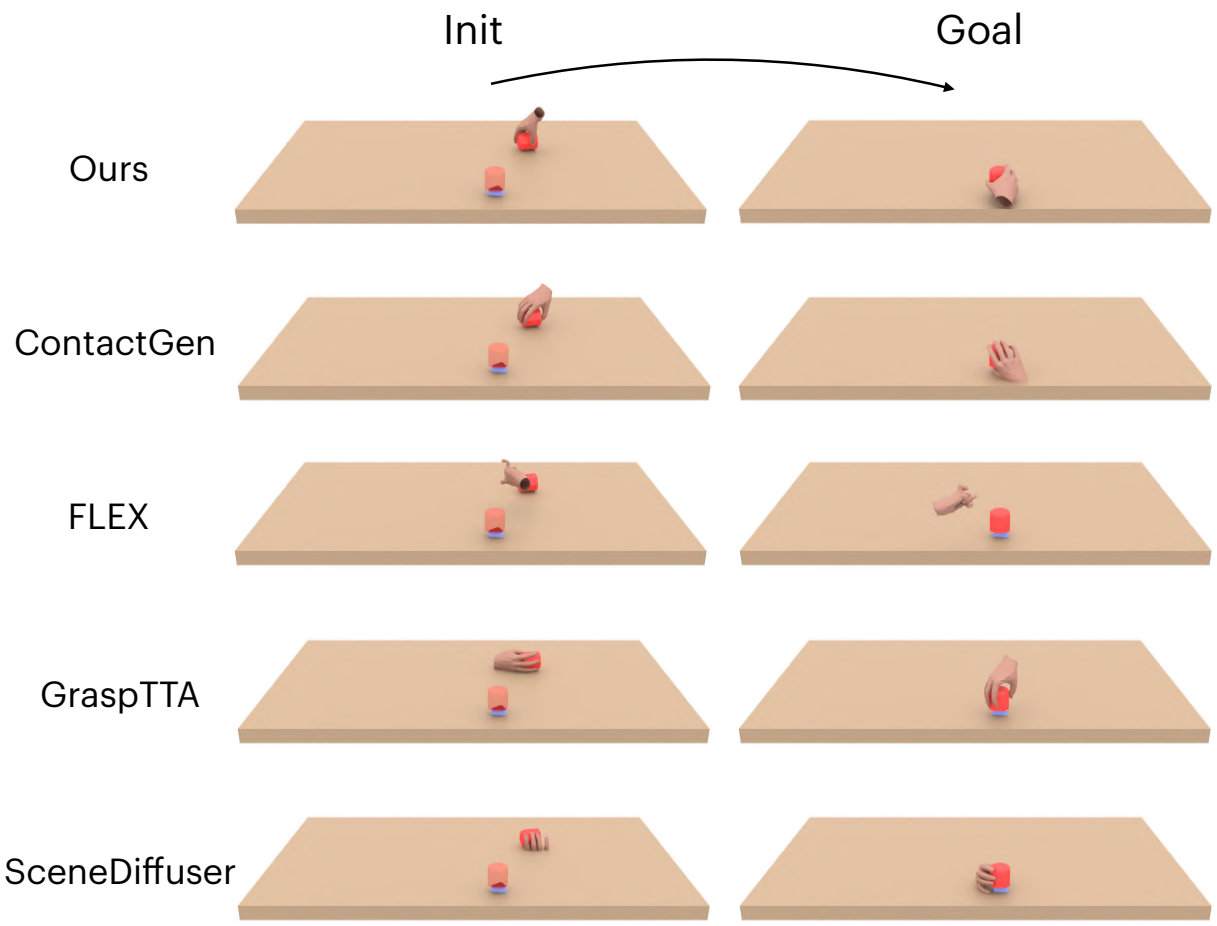

Figure 7. Visualization of predicted human grasps for **brick V** from Ours and ContactGen [3], FLEX [6], GraspTTA [2], SceneDiffuser [1].

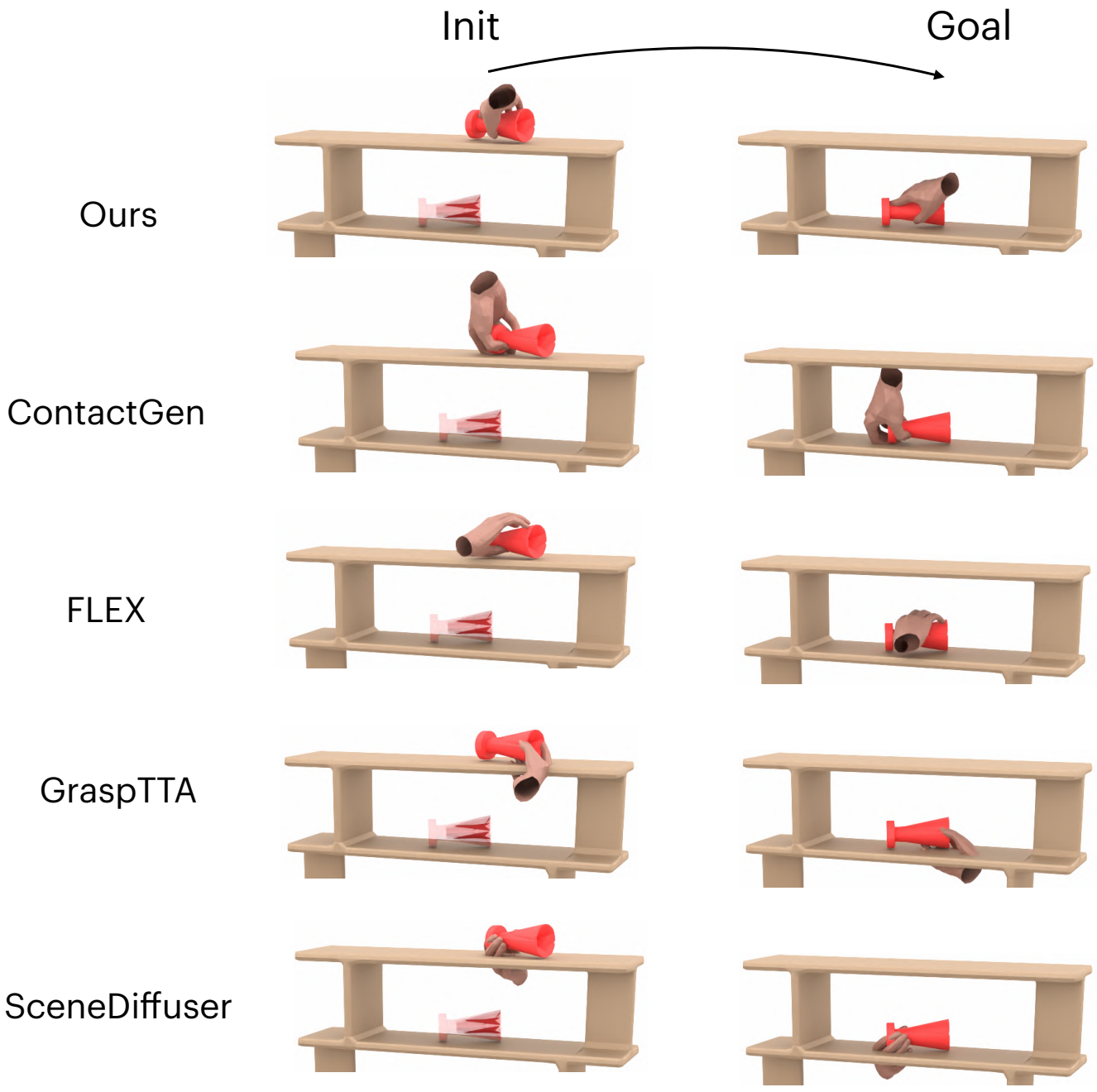

Figure 8. Visualization of predicted human grasps for **trophy** from Ours and ContactGen [3], FLEX [6], GraspTTA [2], SceneDiffuser [1].

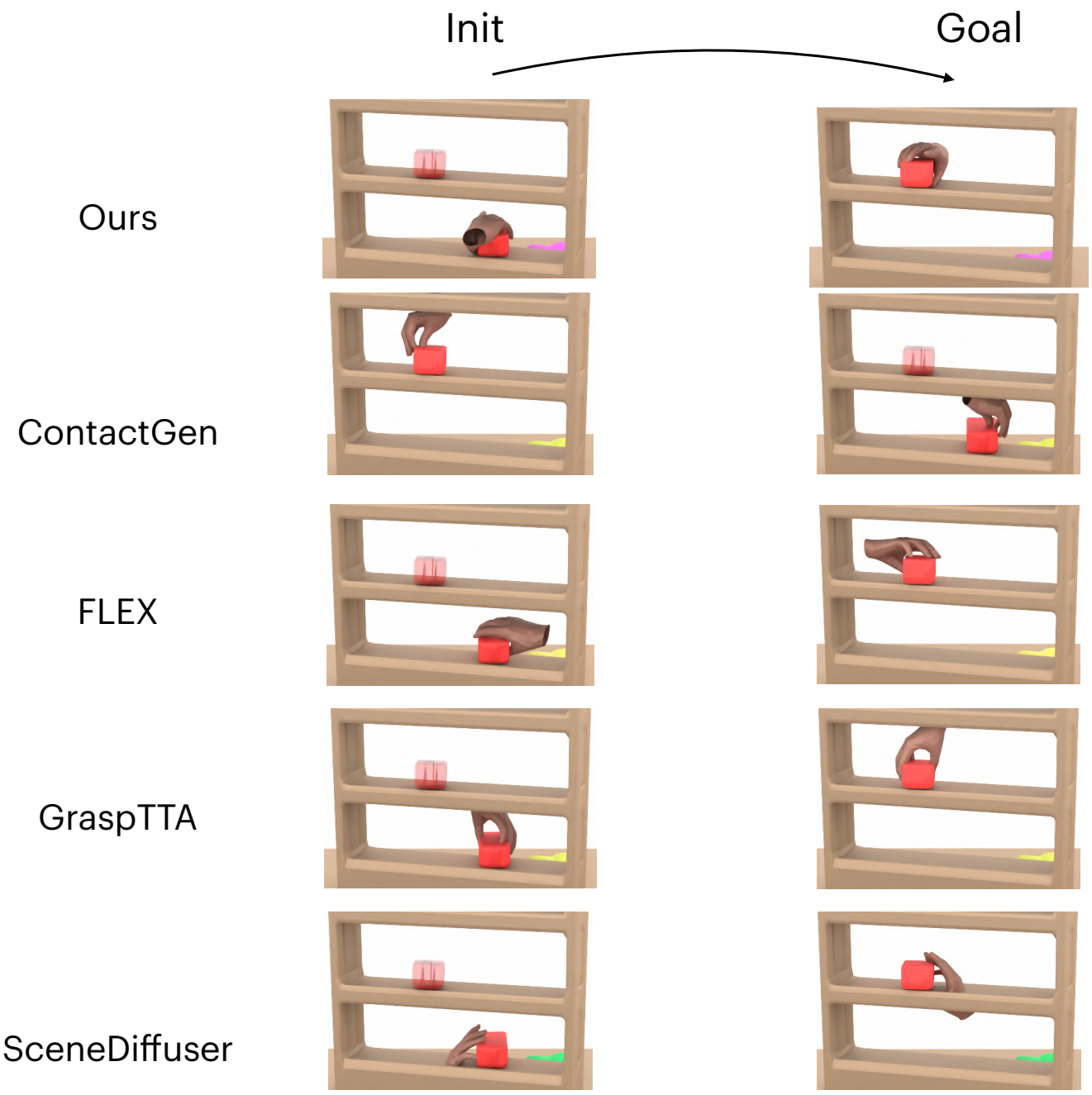

Figure 9. Visualization of predicted human grasps for **toaster** from Ours and ContactGen [3], FLEX [6], GraspTTA [2], SceneDiffuser [1].

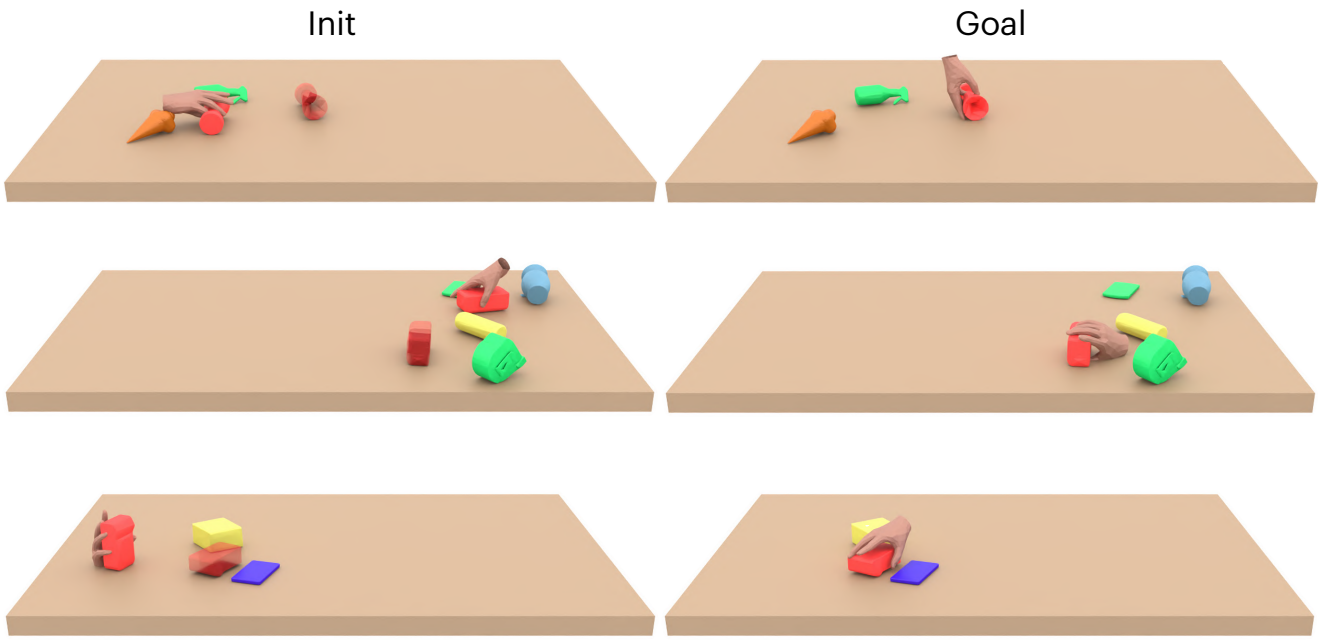

Figure 10. Visualization of **failure** predicted human grasp for **trophy** and **camera** from Ours

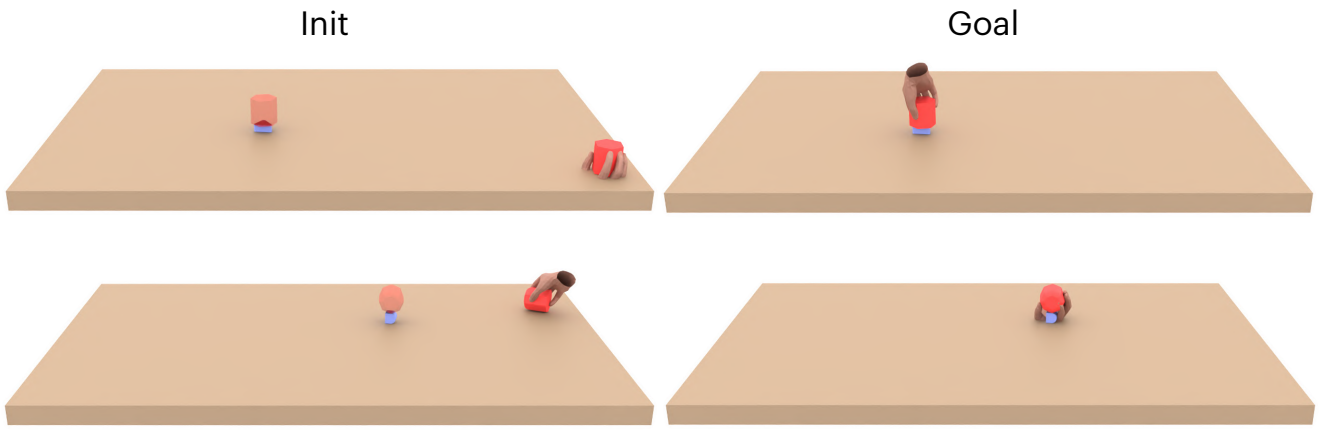

Figure 11. Visualization of **failure** predicted human grasp for **brick R** and **brick V** from Ours

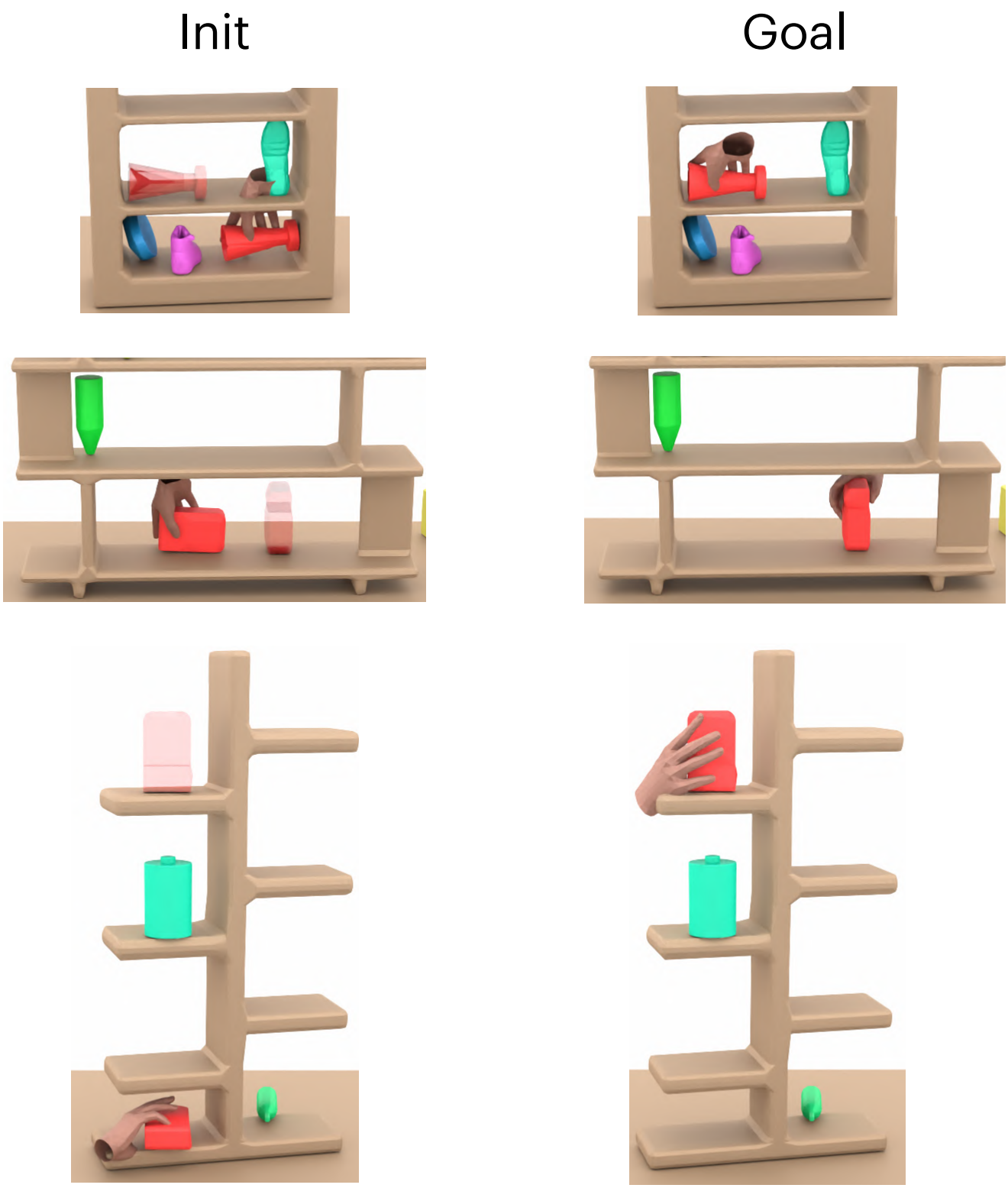

Figure 12. Visualization of **failure** predicted human grasp for **trophy** and **camera** from Ours

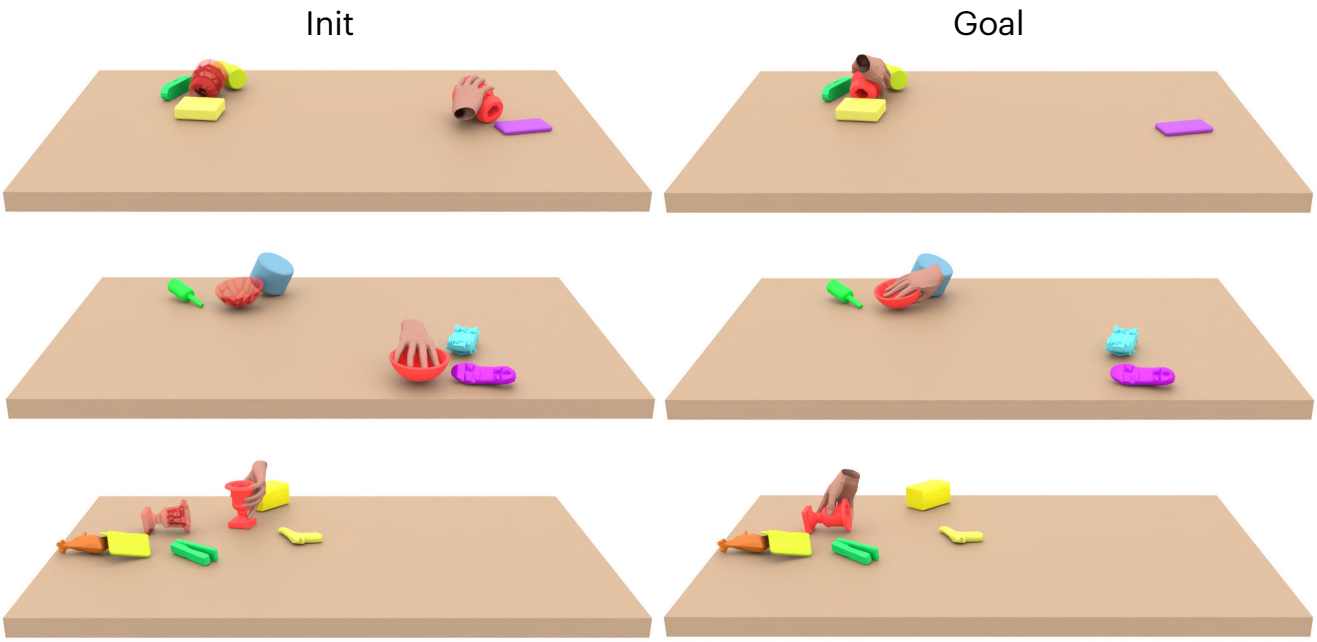

Figure 13. Visualization of predicted human grasps on bottle, bowl, and jar.

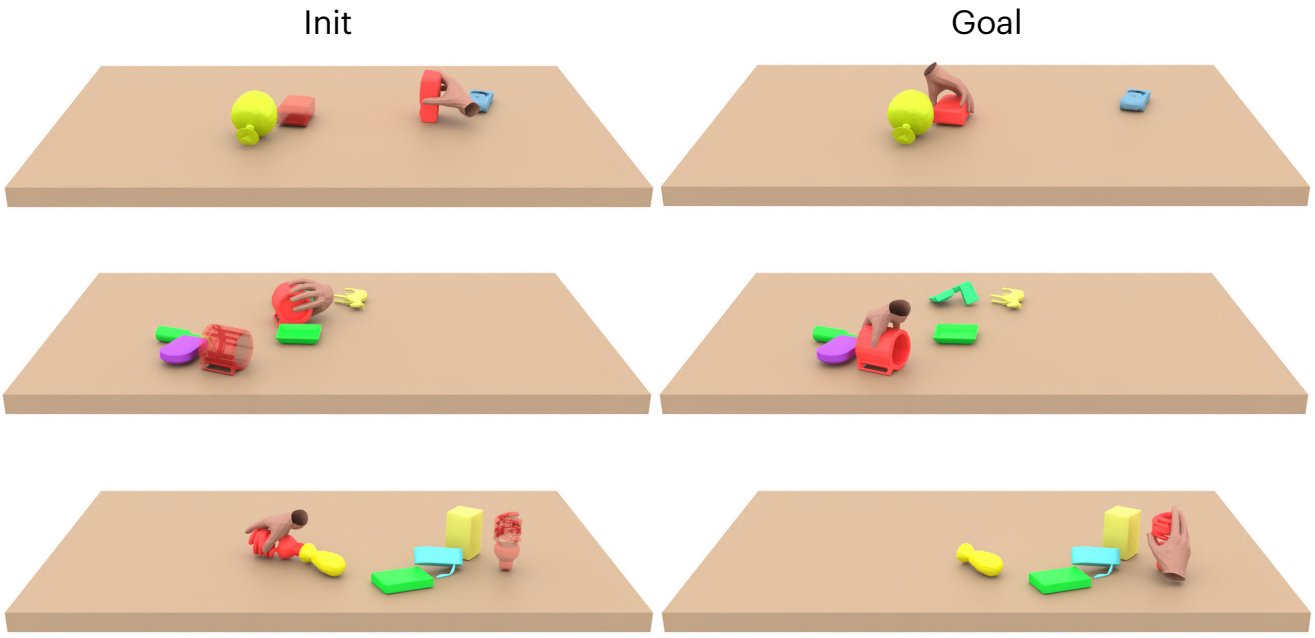

Figure 14. Visualization of predicted human grasps on camera, mug, and lightbuld.

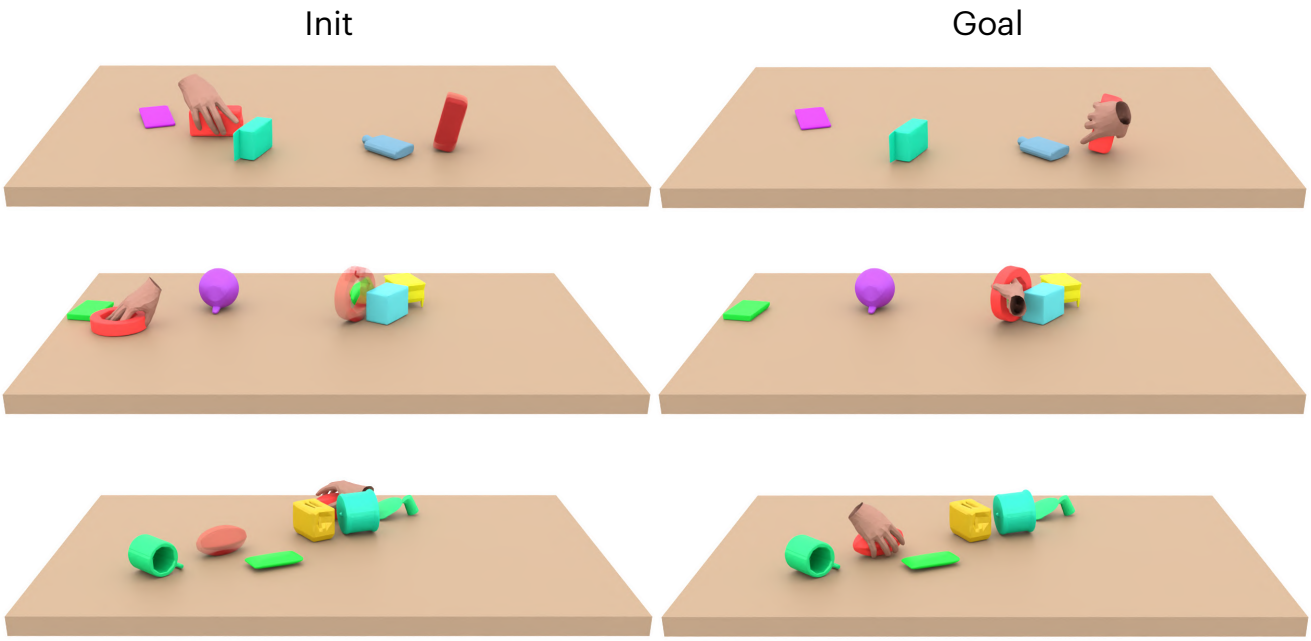

Figure 15. Visualization of predicted human grasps on box, tape, and star fruit

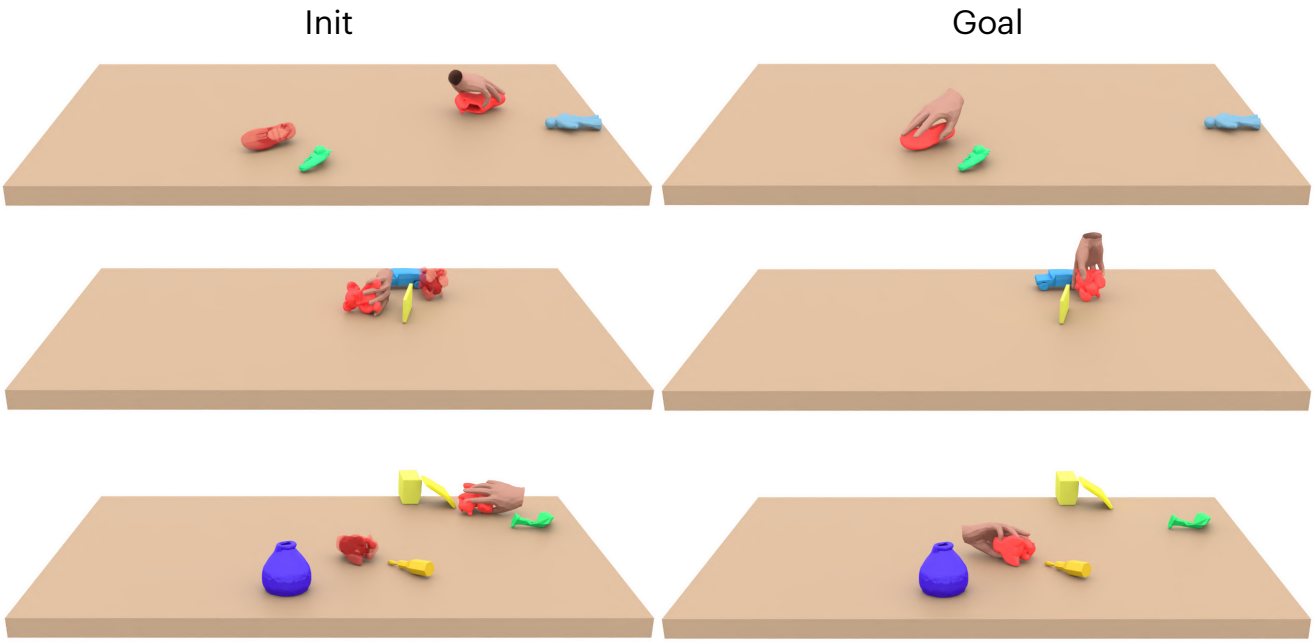

Figure 16. Visualization of predicted human grasps on shoe, elephant doll, and bear doll

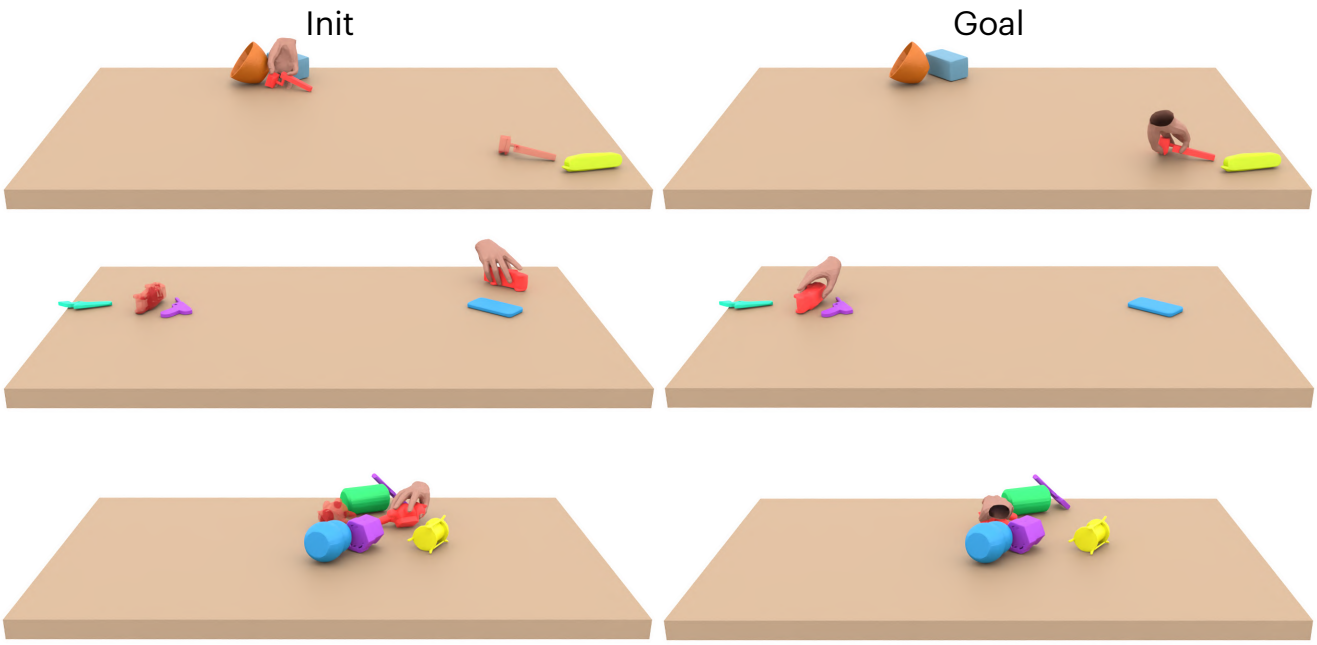

Figure 17. Visualization of predicted human grasps on camera, toy car, and doll

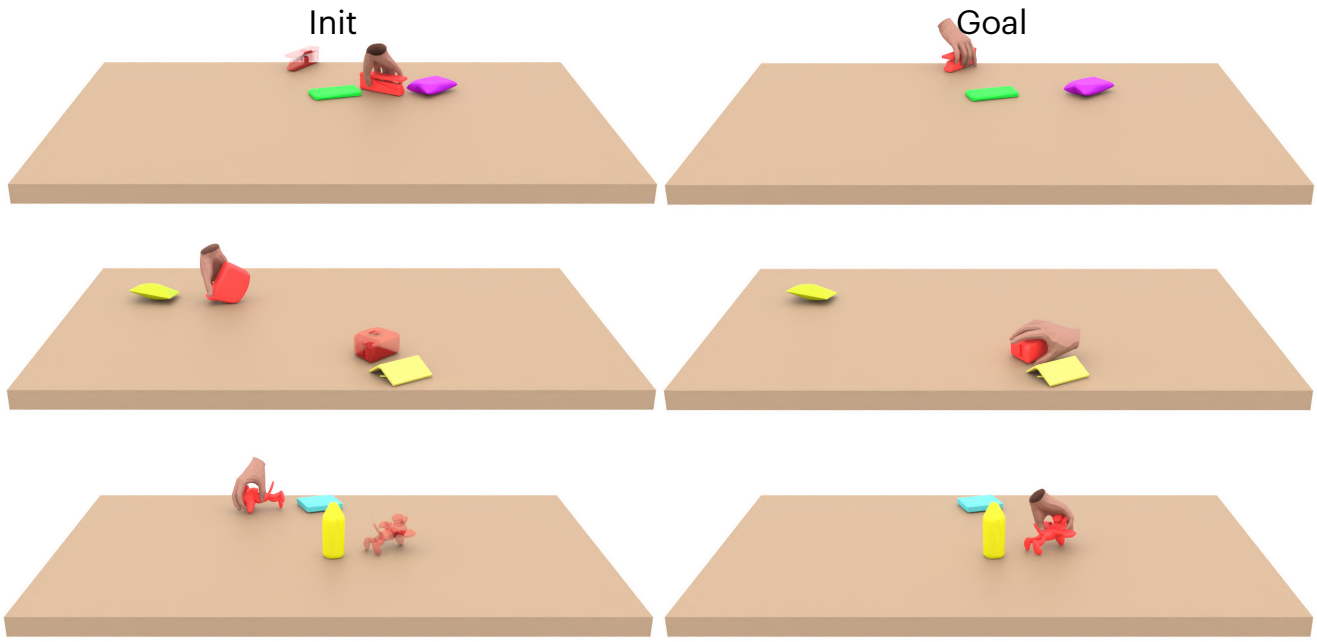

Figure 18. Visualization of predicted human grasps on stapler, tapemeasure, and toy

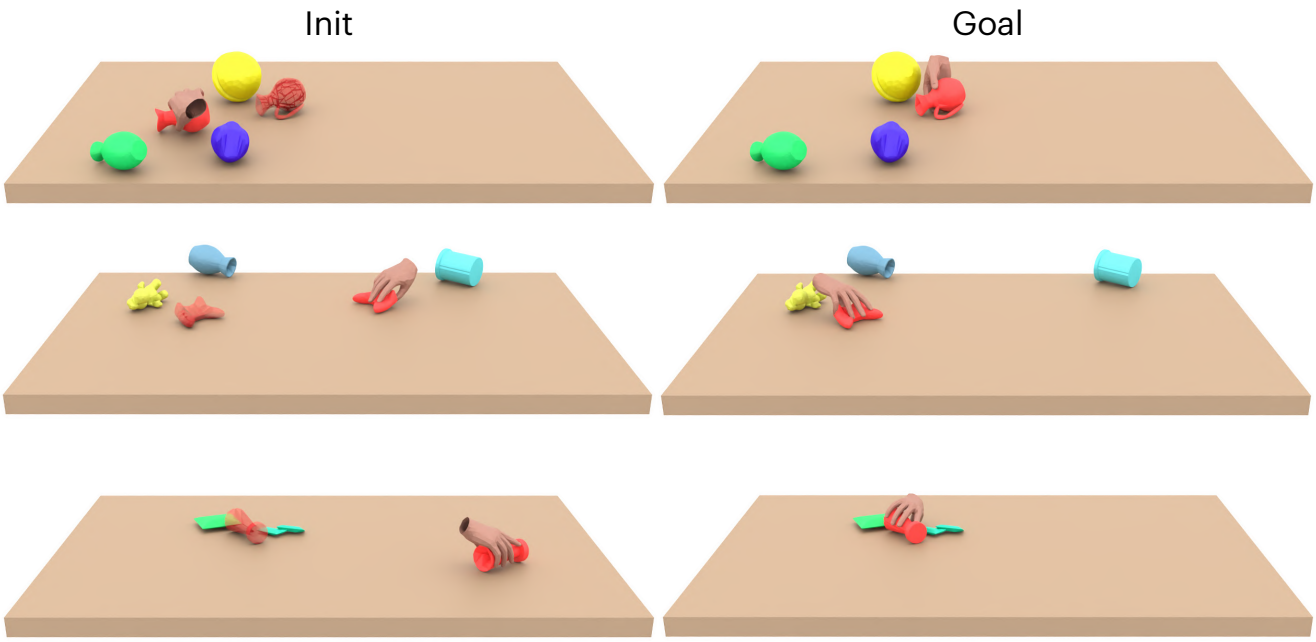

Figure 19. Visualization of predicted human grasps on vase, video game controller, and trophy

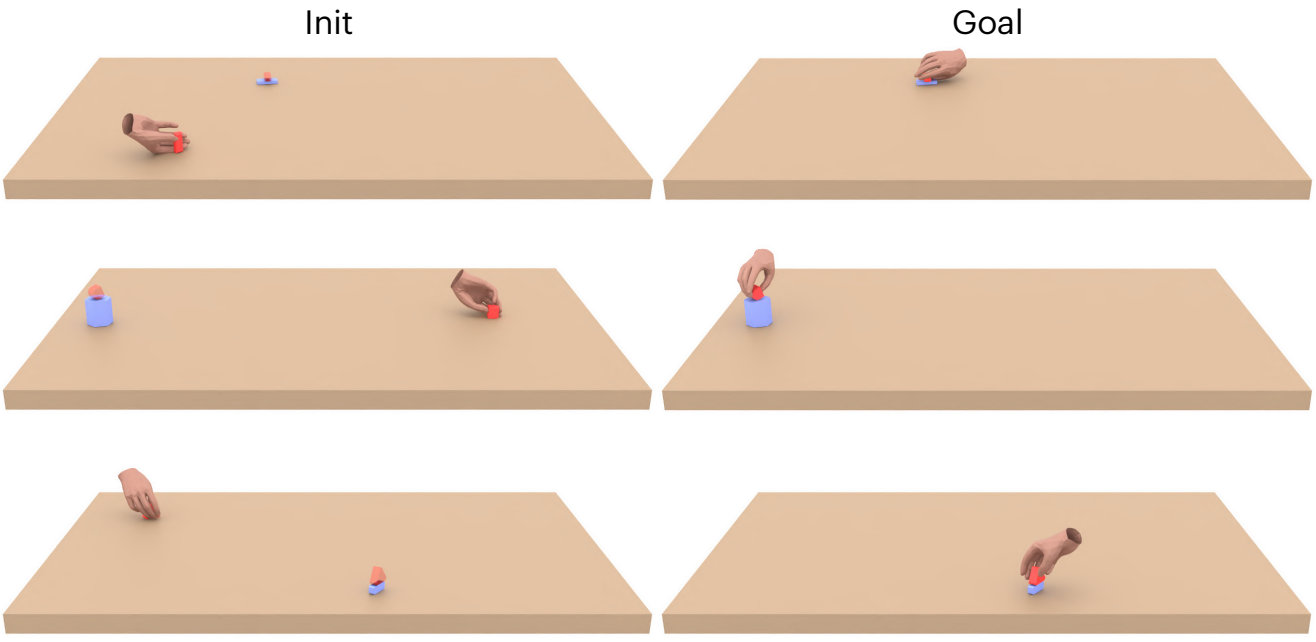

Figure 20. Visualization of predicted human grasps on brick F, brick I, and brick K

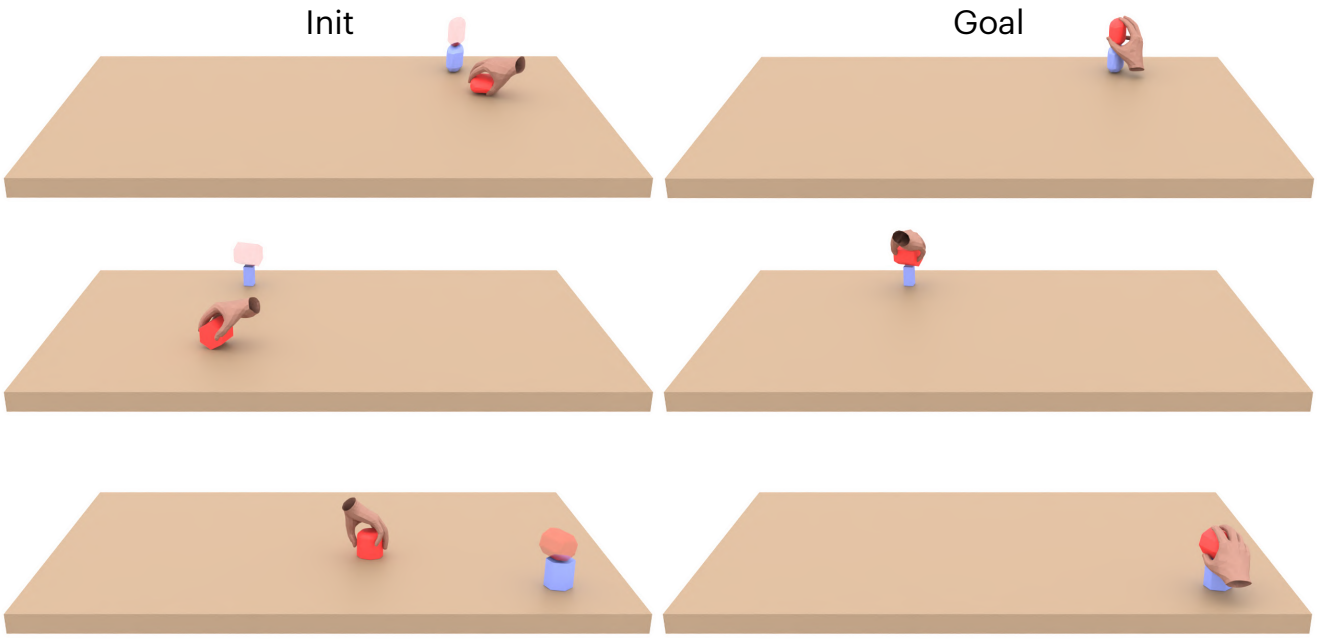

Figure 21. Visualization of predicted human grasps on brick N, brick R, and brick V

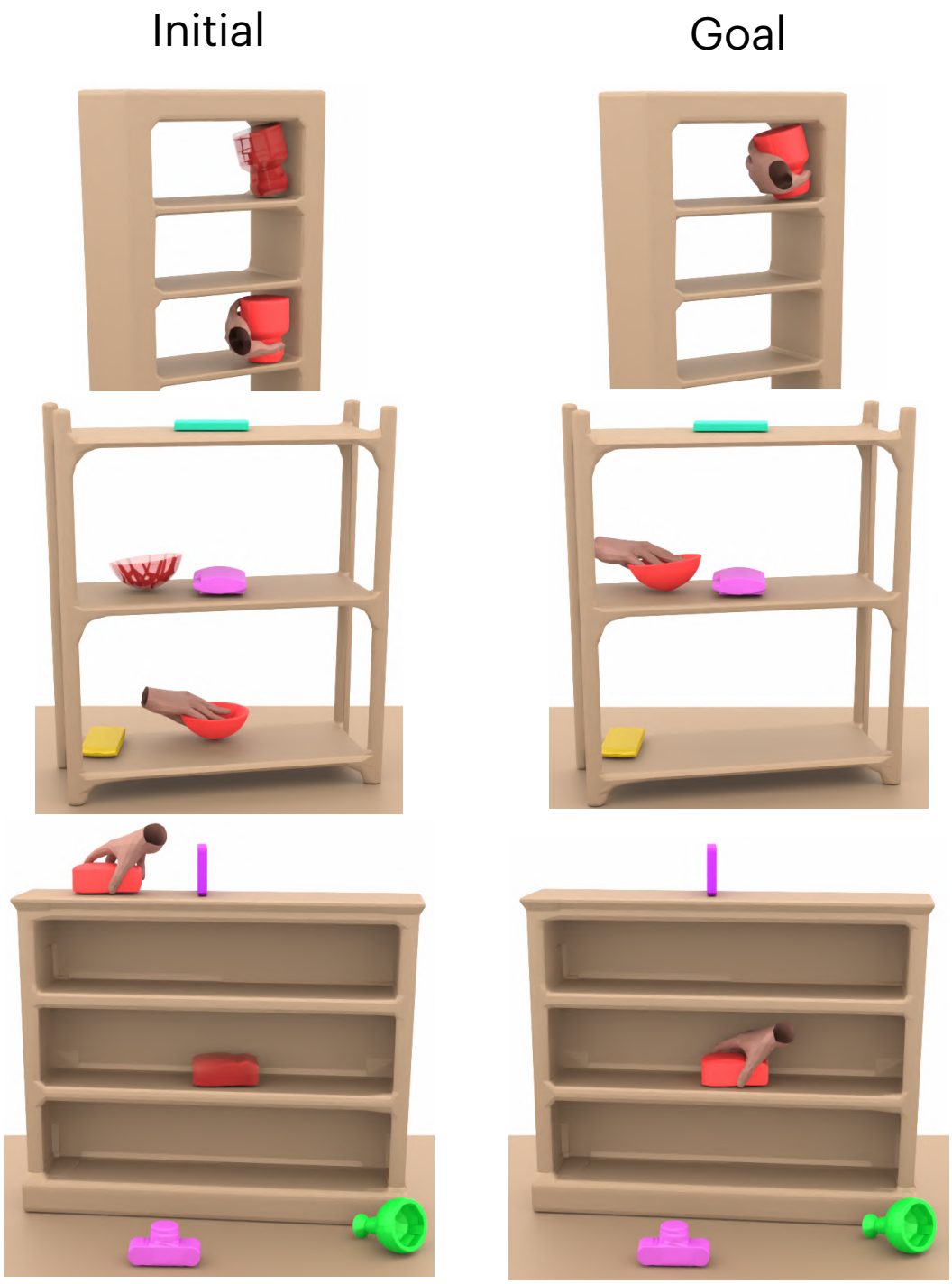

Figure 22. Visualization of predicted human grasps on bottle, bowl, and camera.

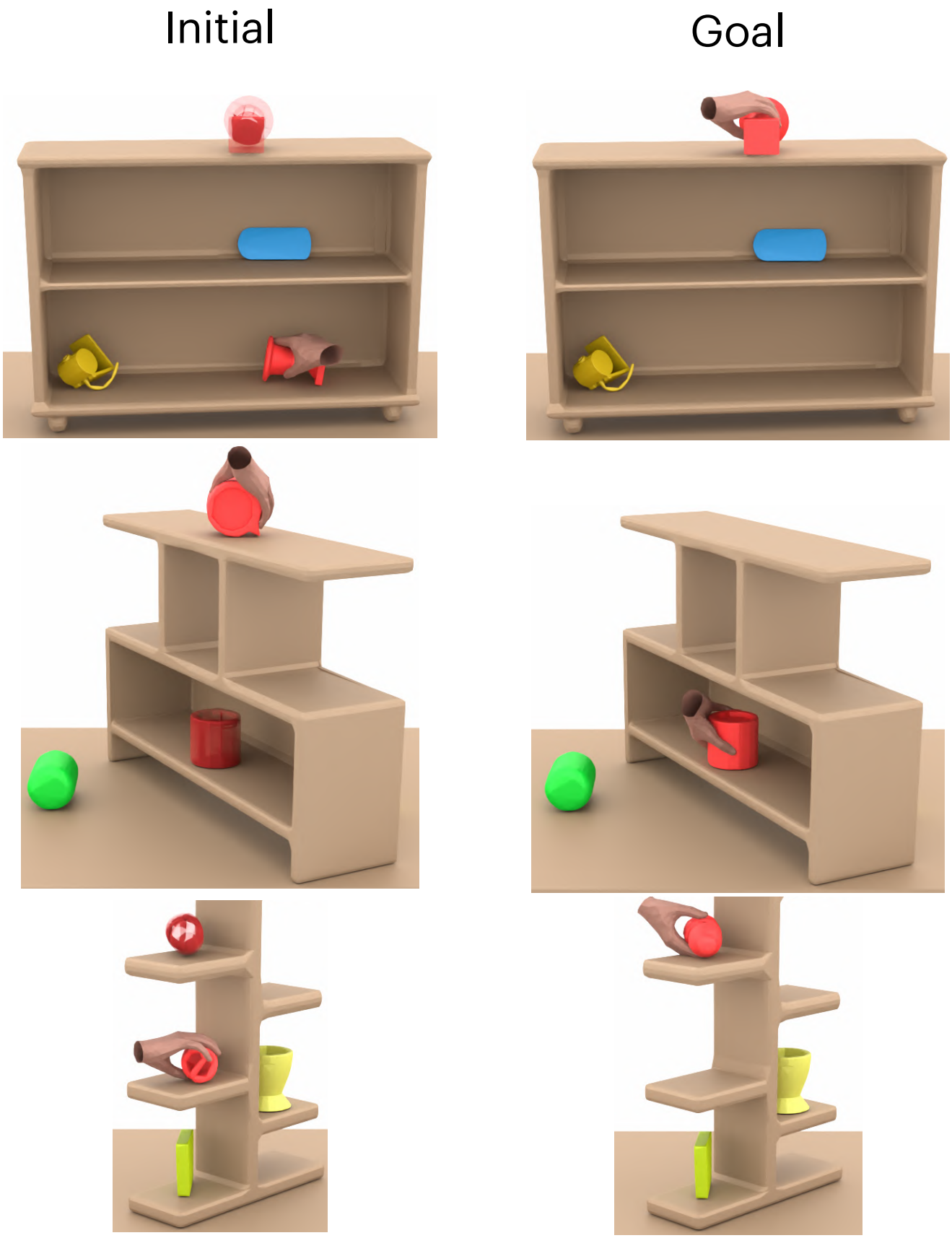

Figure 23. Visualization of predicted human grasps on jar, mug, and lightbuld.

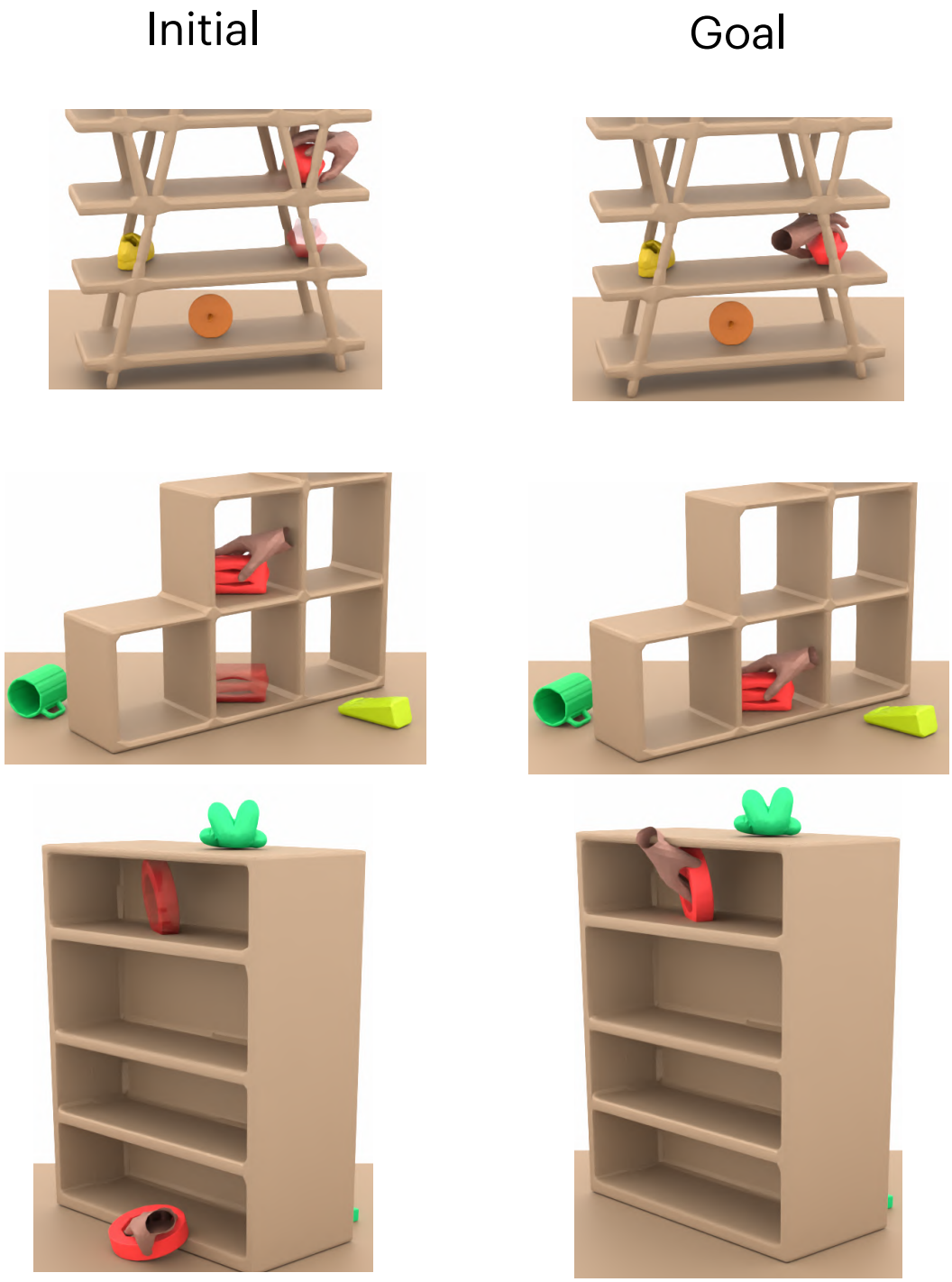

Figure 24. Visualization of predicted human grasps on start fruit, toaster, and tape

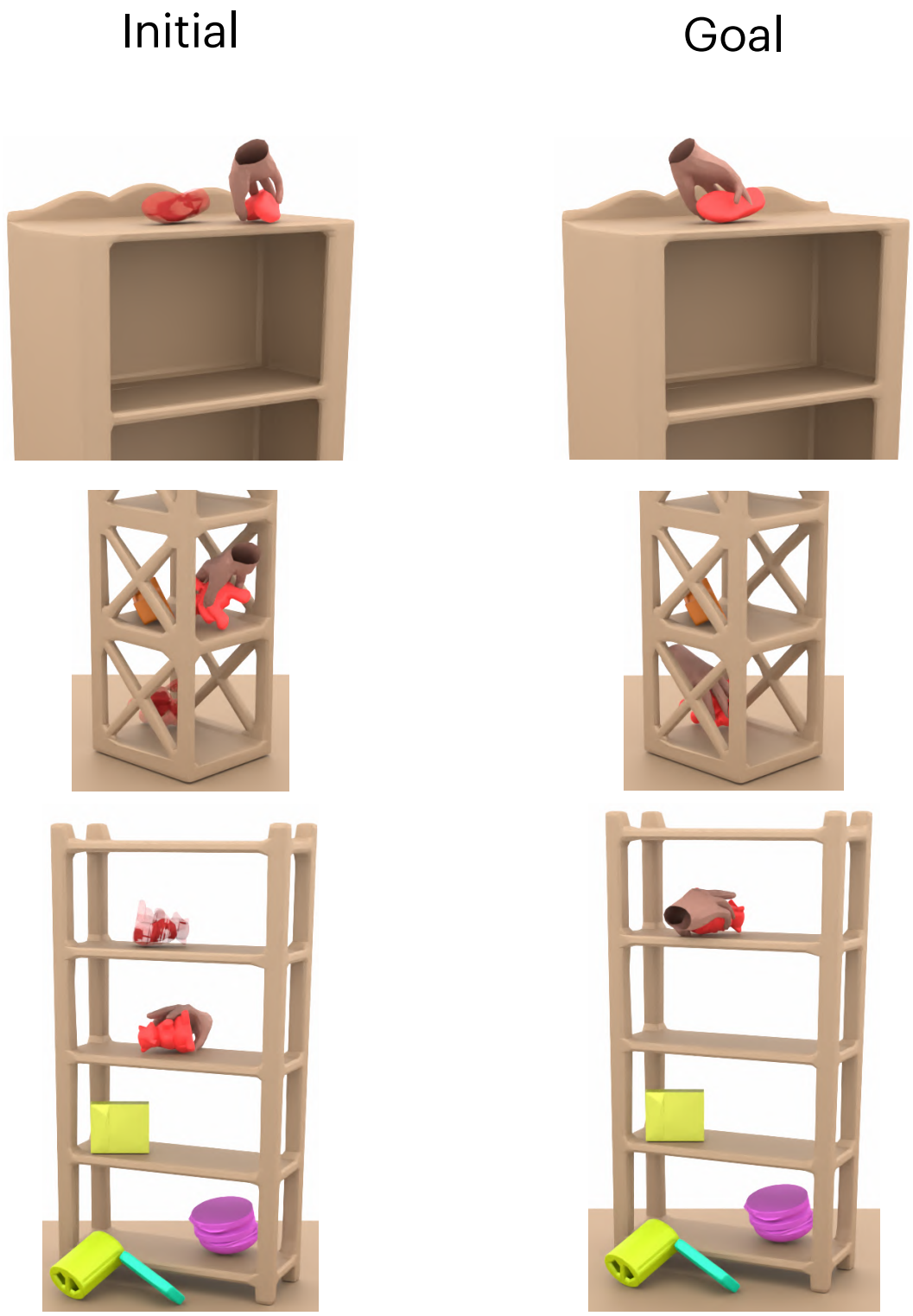

Figure 25. Visualization of predicted human grasps on shoe, elephant doll, and bear doll

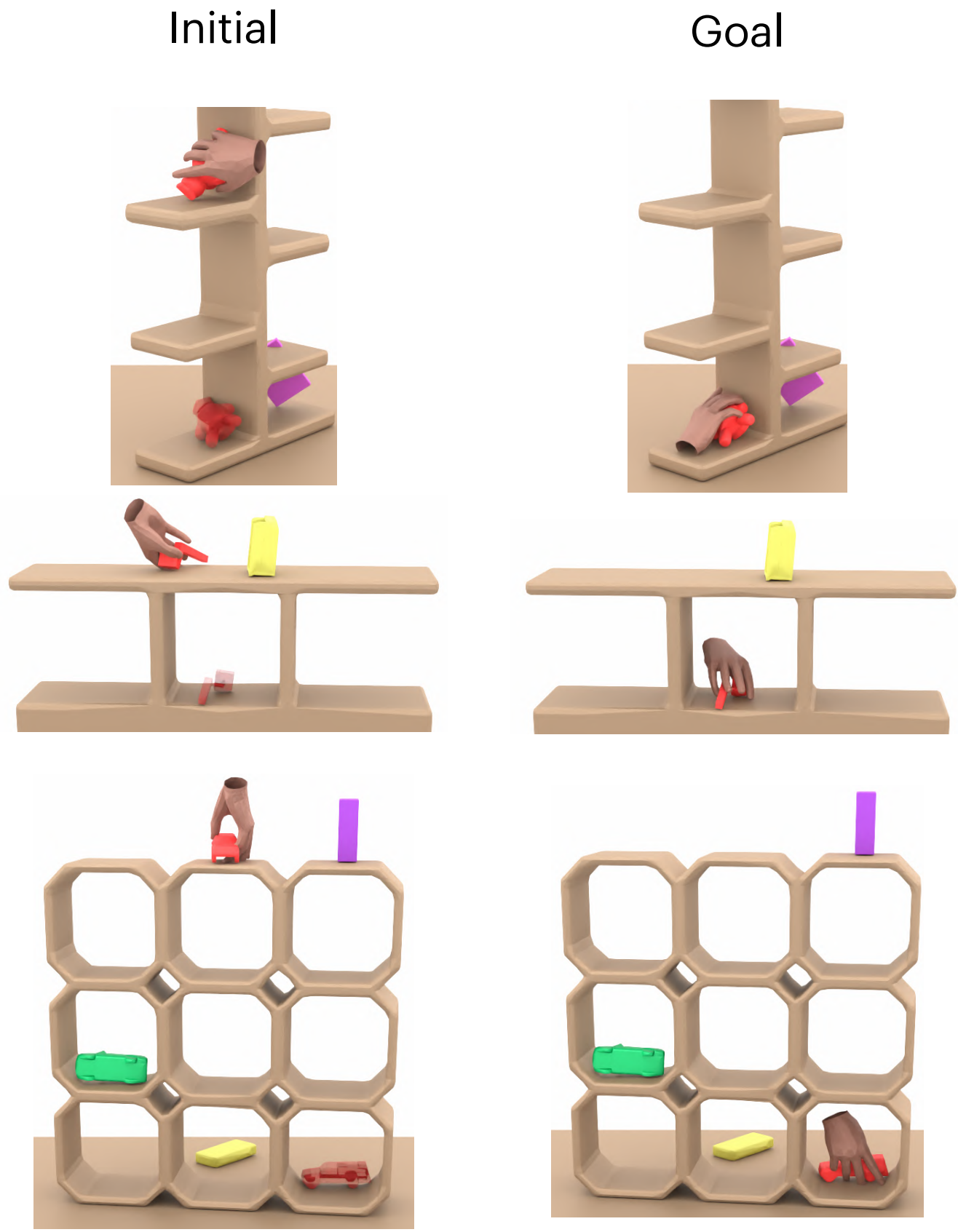

Figure 26. Visualization of predicted human grasps on doll, camera, and car

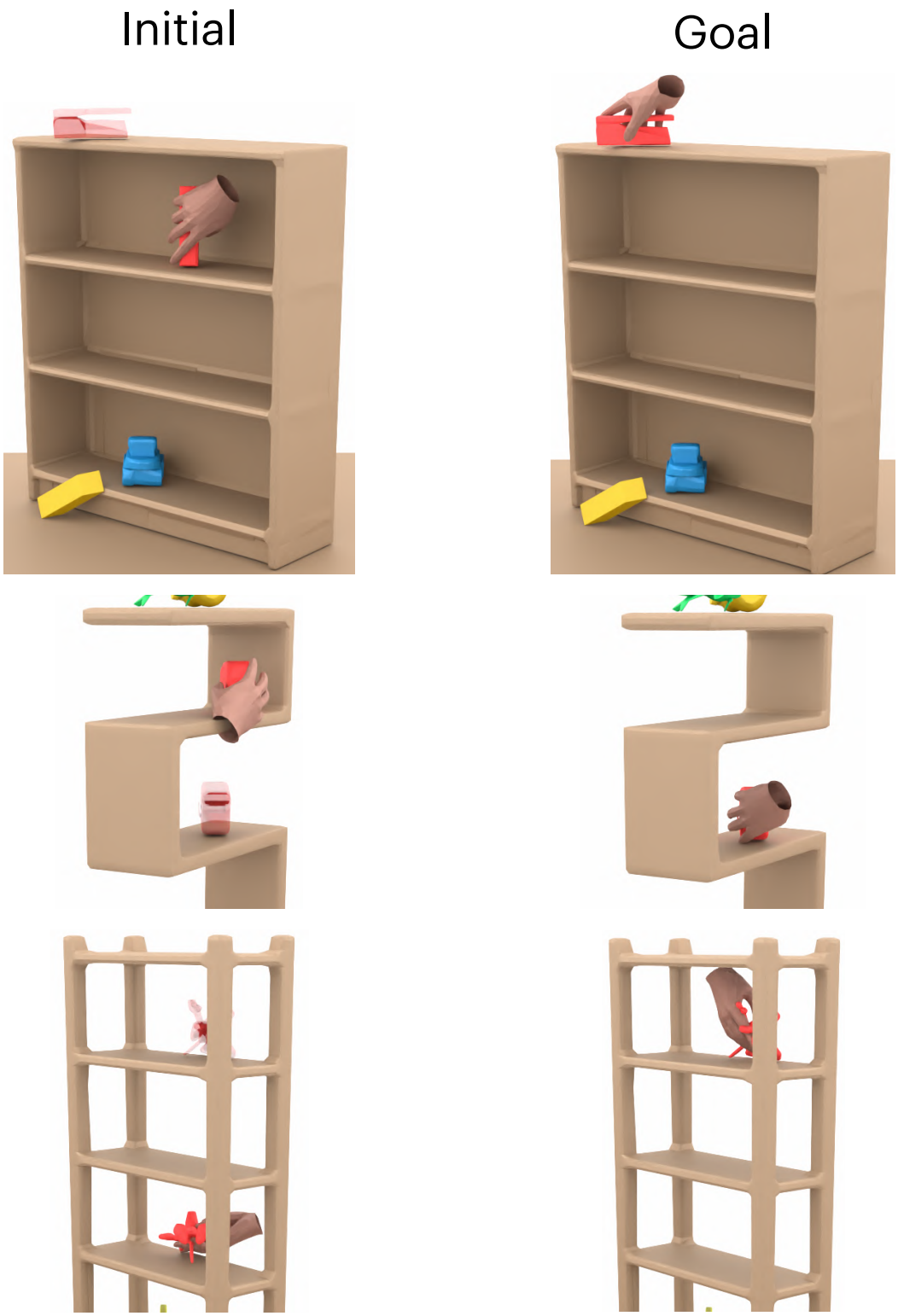

Figure 27. Visualization of predicted human grasps on stapler, tapemeasure, and toy

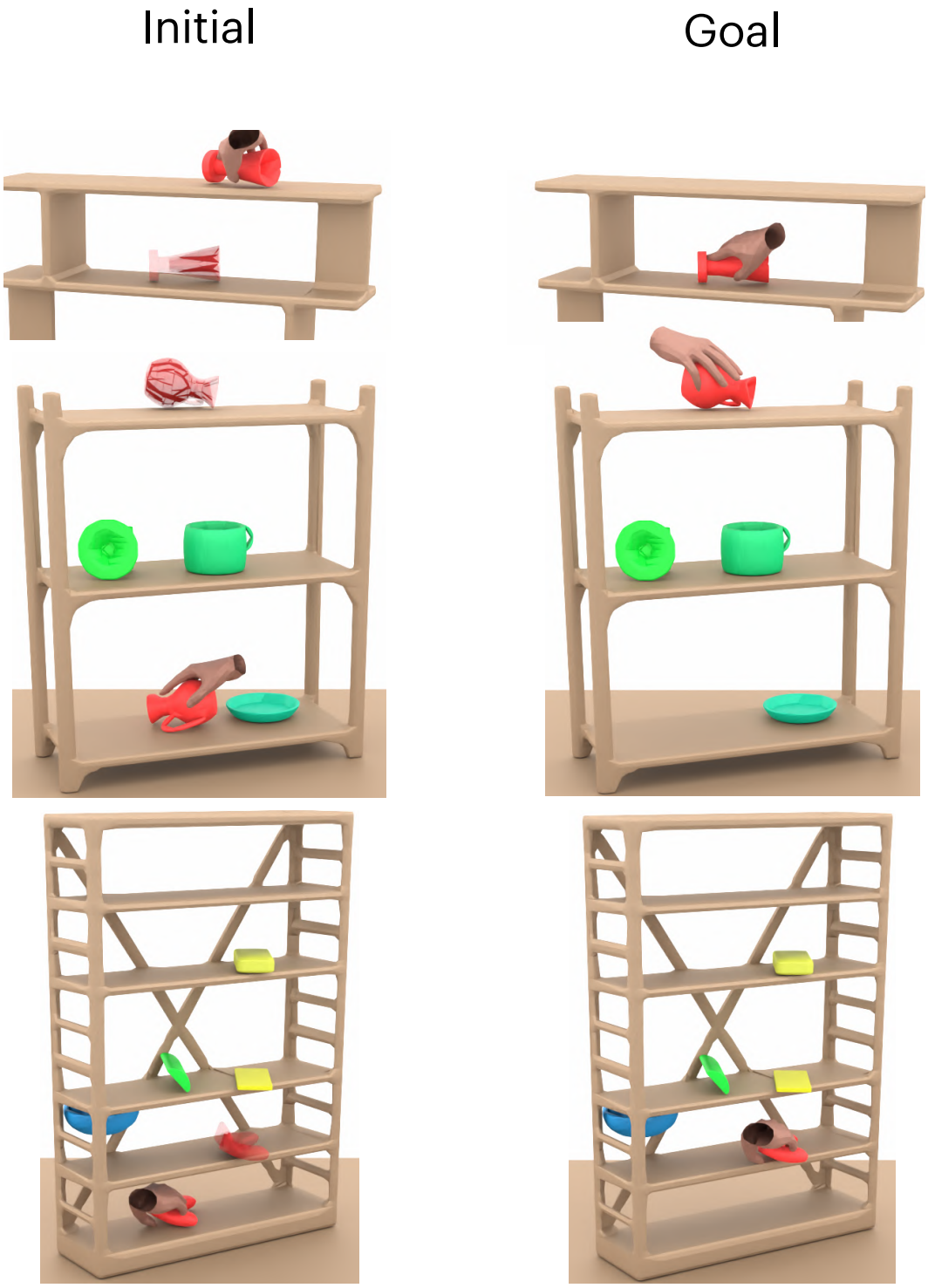

Figure 28. Visualization of predicted human grasps on trophy, vase, and video game controller

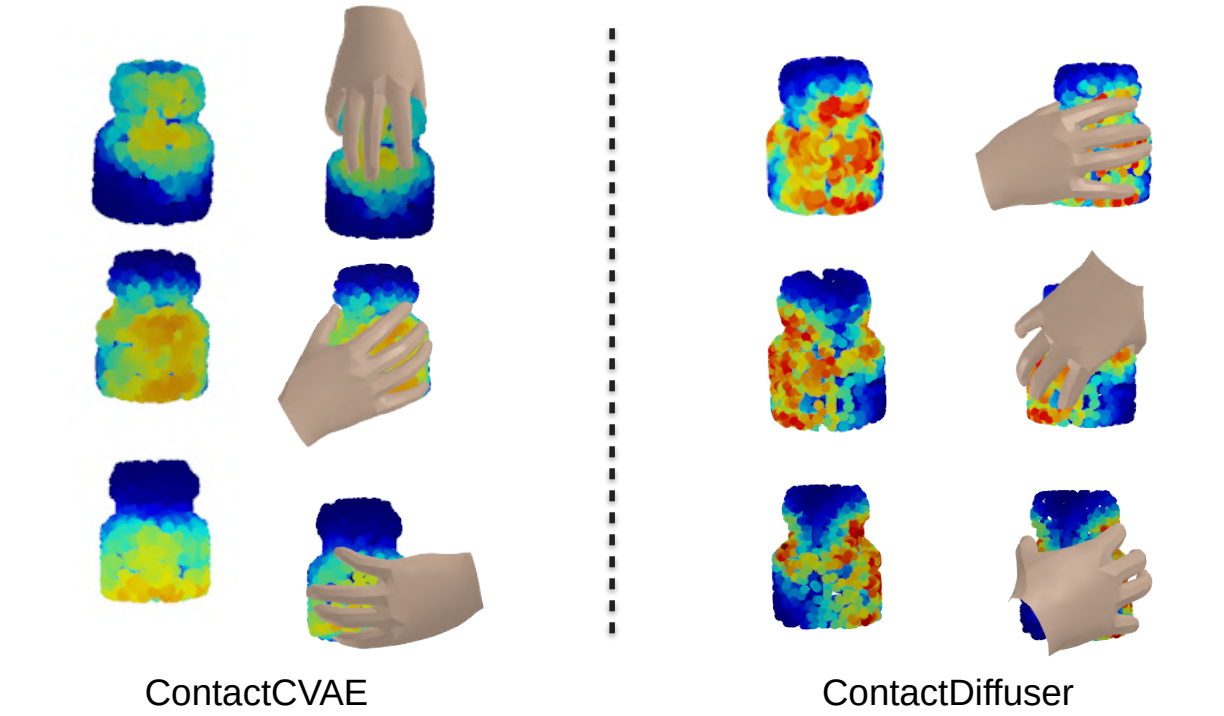

Figure 29. Visualization of predicted contact map and grasp on **bottle** from ContactCVAE [3] and Ours

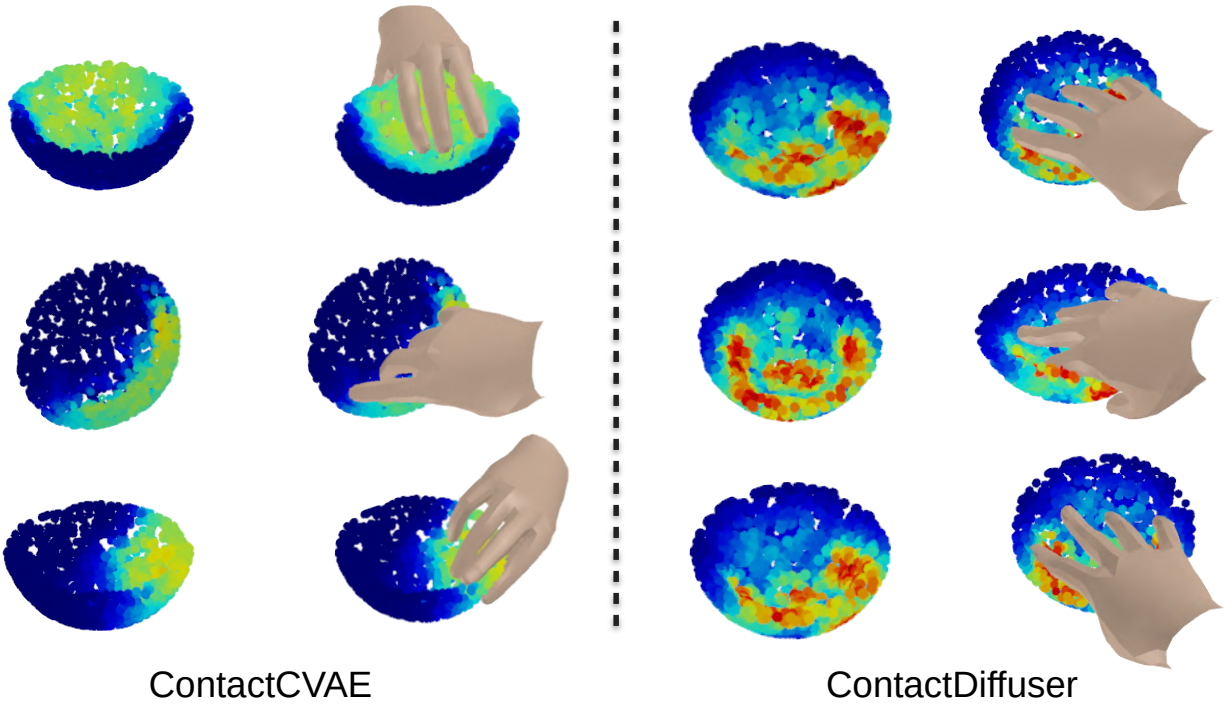

Figure 30. Visualization of predicted contact map and grasp on **bowl** from ContactCVAE [3] and Ours

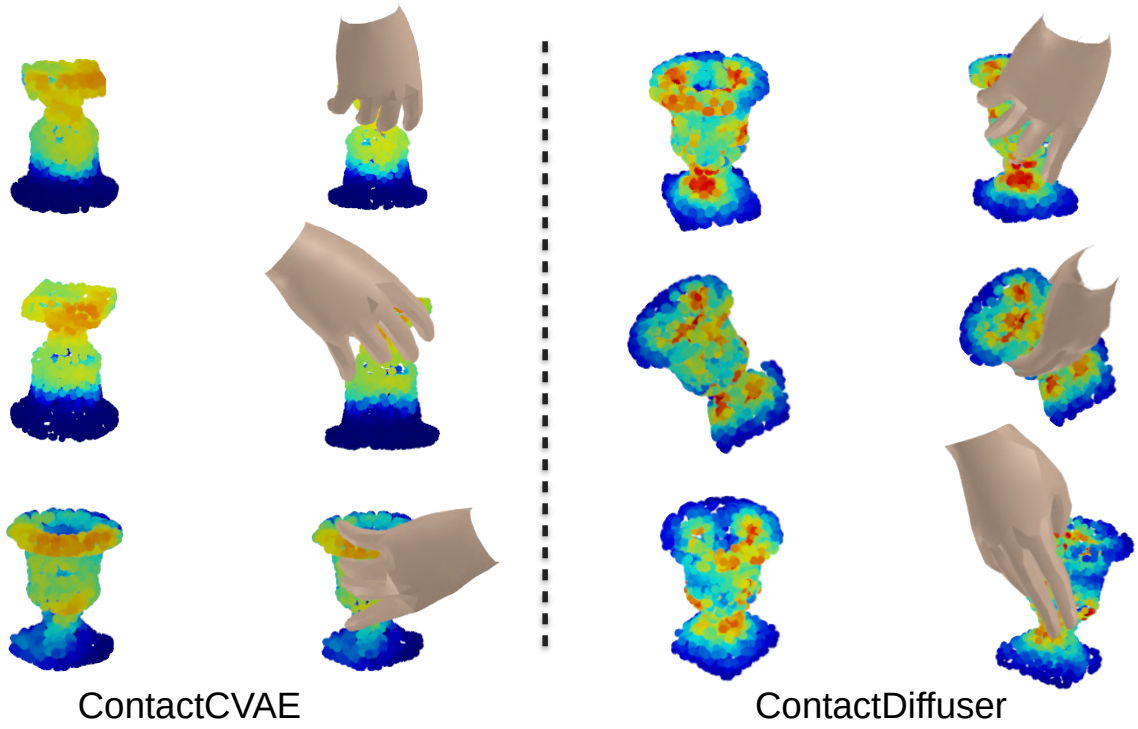

Figure 31. Visualization of predicted contact map and grasp on **jar** from ContactCVAE [3] and Ours

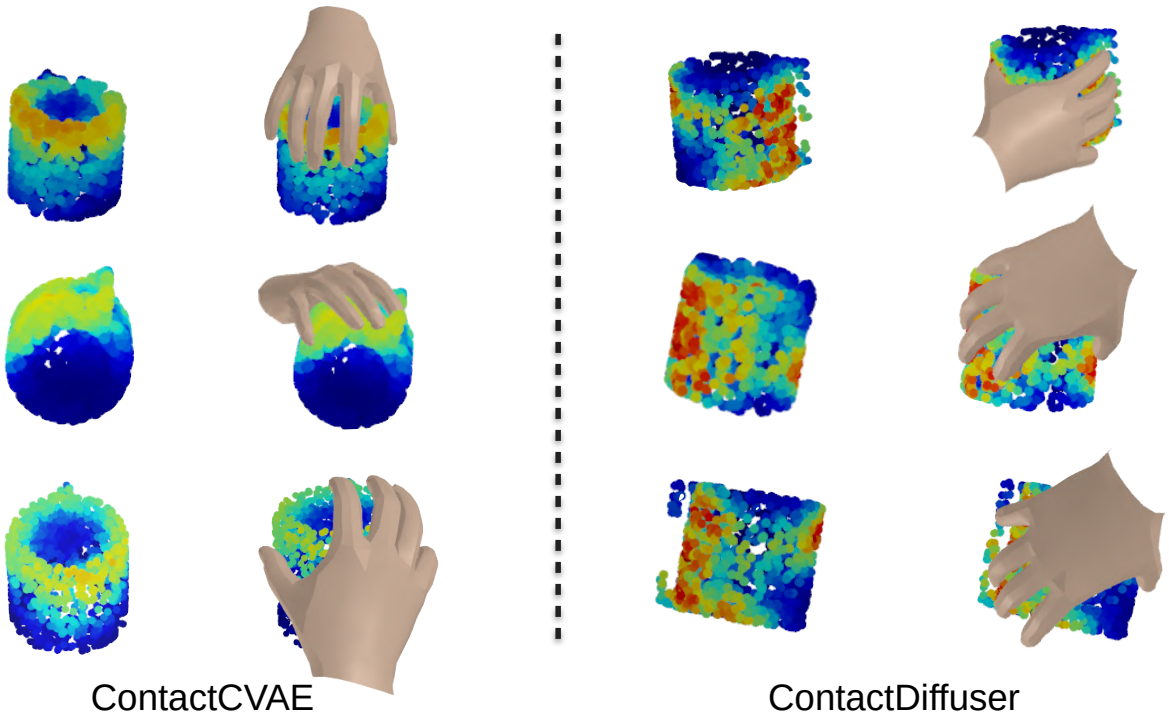

Figure 32. Visualization of predicted contact map and grasp on **mug** from ContactCVAE [3] and Ours

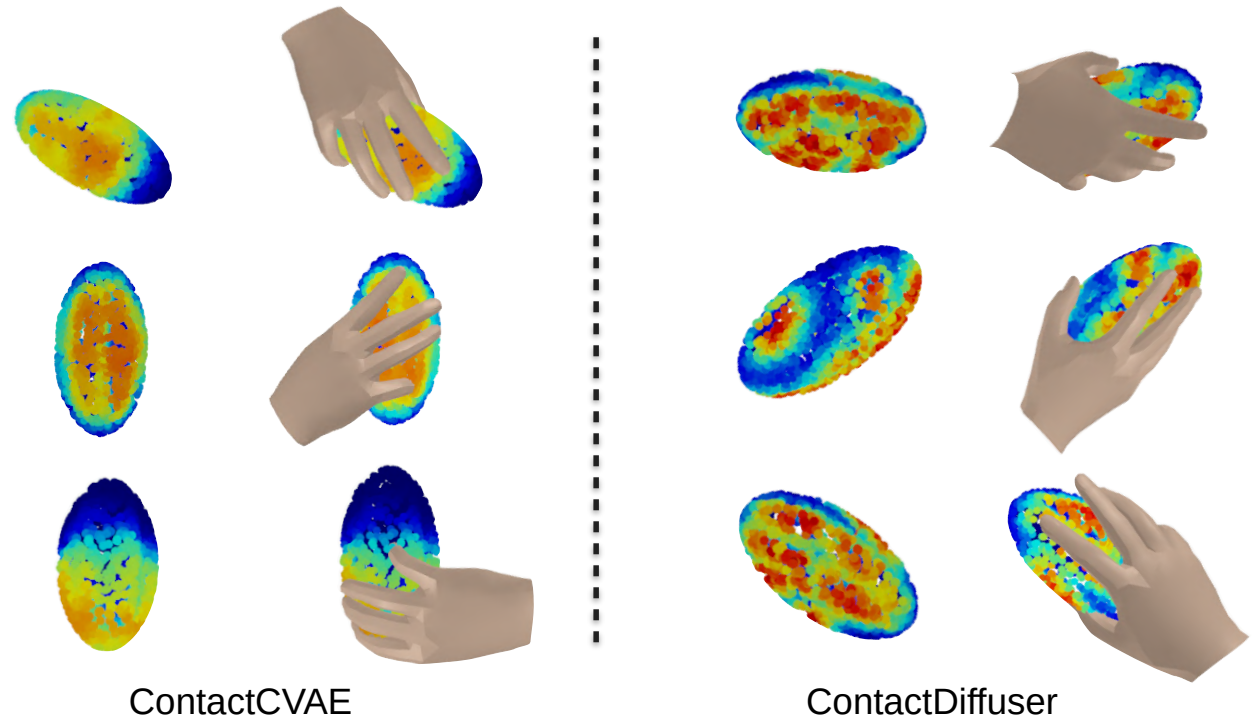

Figure 33. Visualization of predicted contact map and grasp on **starfruit** from ContactCVAE [3] and Ours

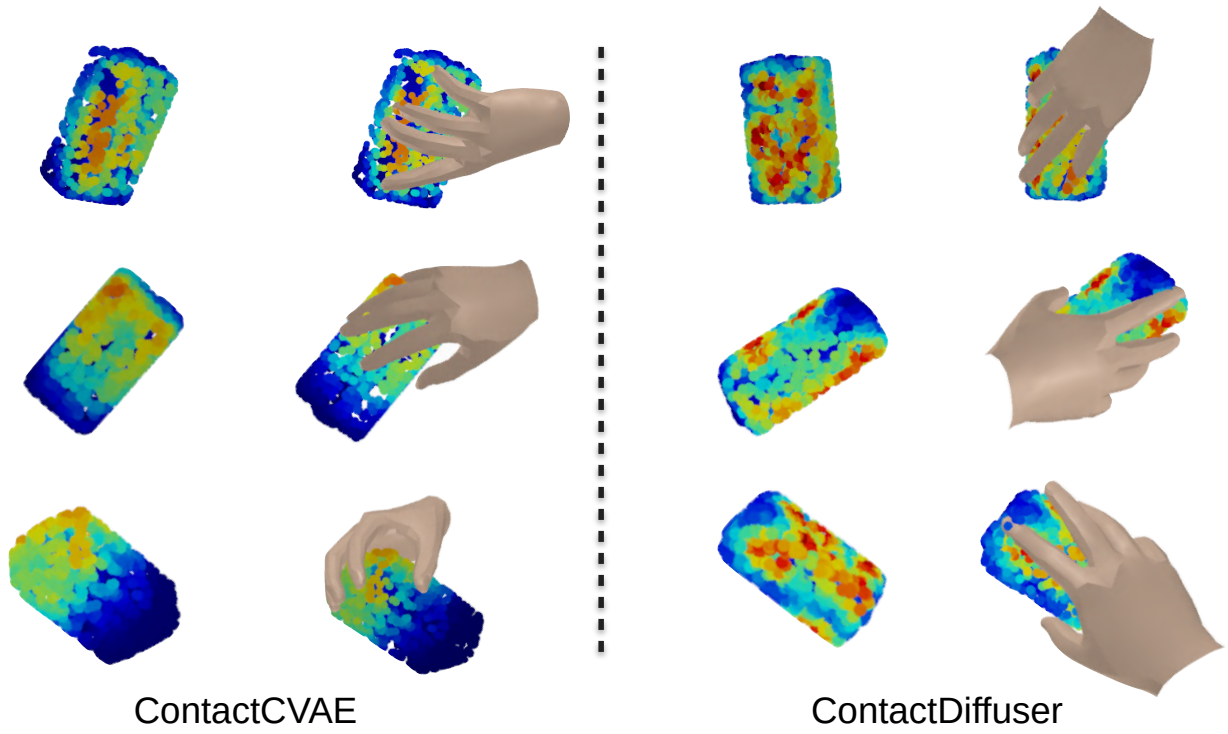

Figure 34. Visualization of predicted contact map and grasp on **toaster** from ContactCVAE [3] and Ours

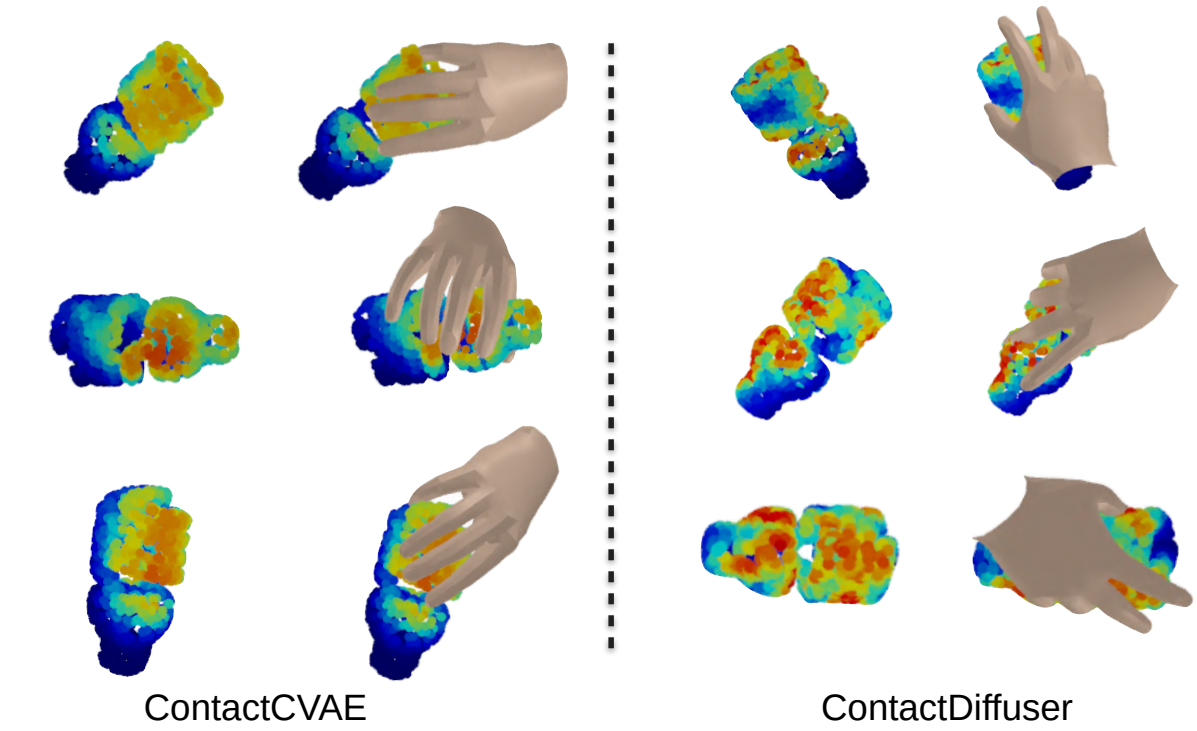

Figure 35. Visualization of predicted contact map and grasp on **lightbuld** from ContactCVAE [3] and Ours

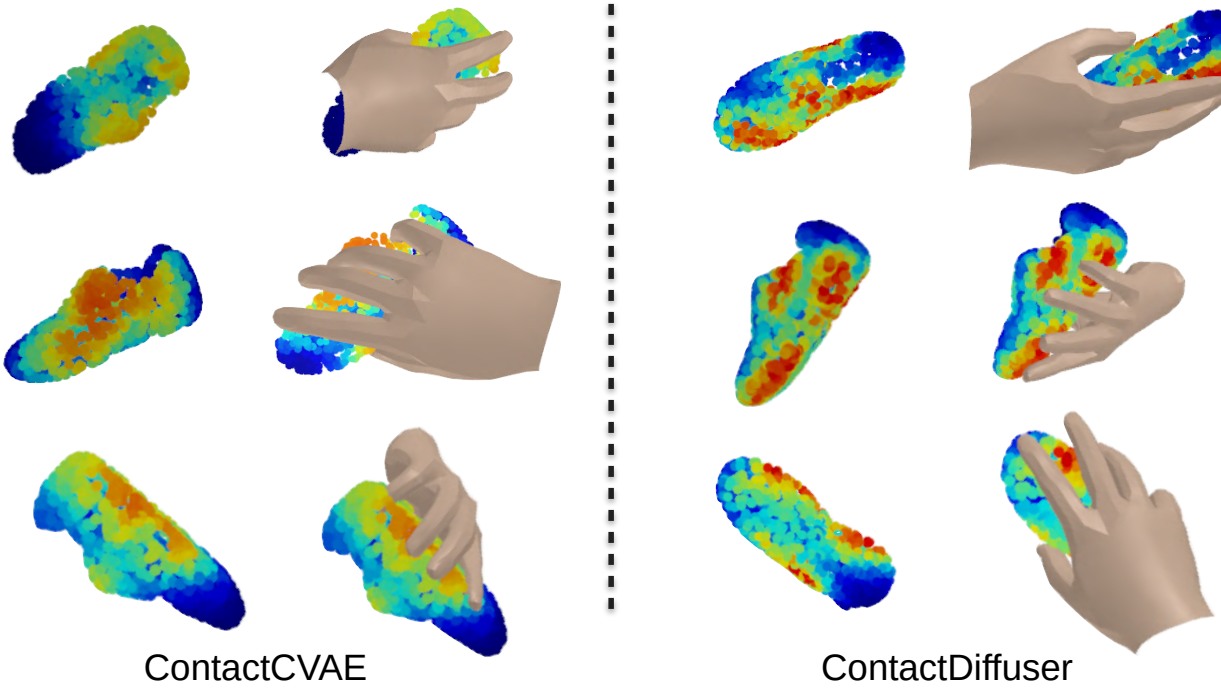

Figure 36. Visualization of predicted contact map and grasp on **shoe** from ContactCVAE [3] and Ours

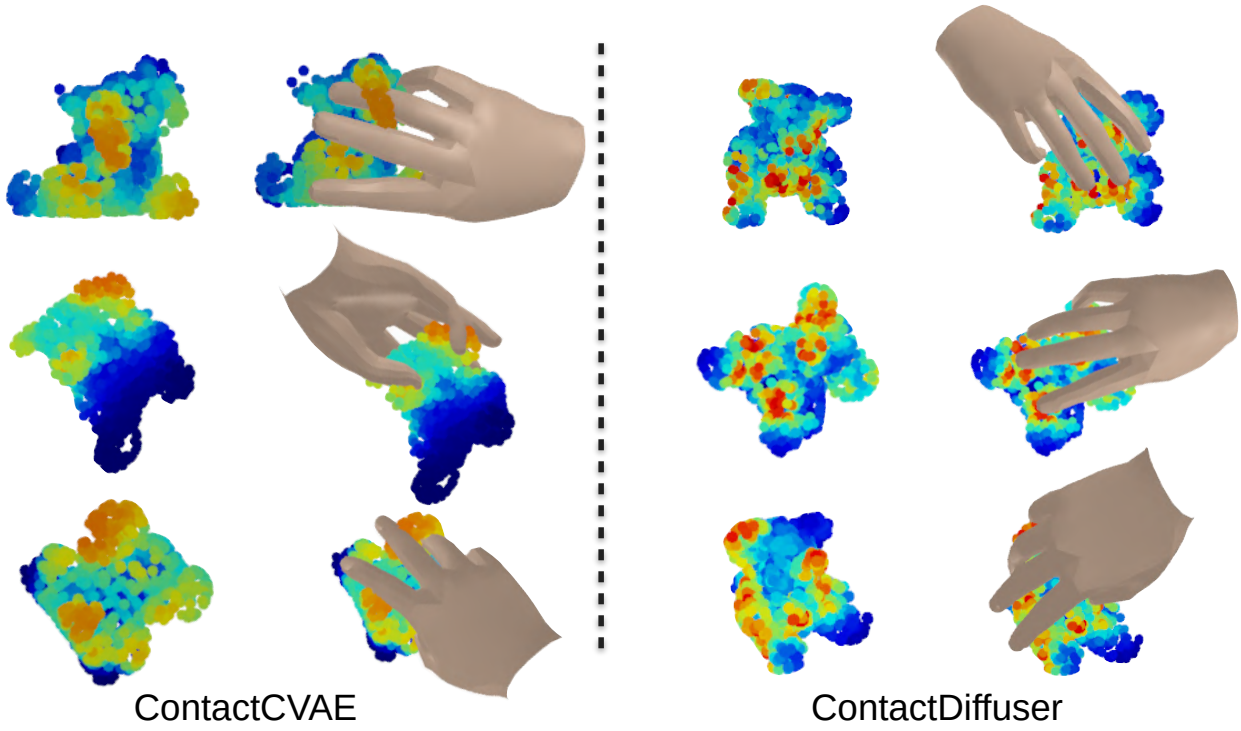

Figure 37. Visualization of predicted contact map and grasp on **elephant doll** from ContactCVAE [3] and Ours

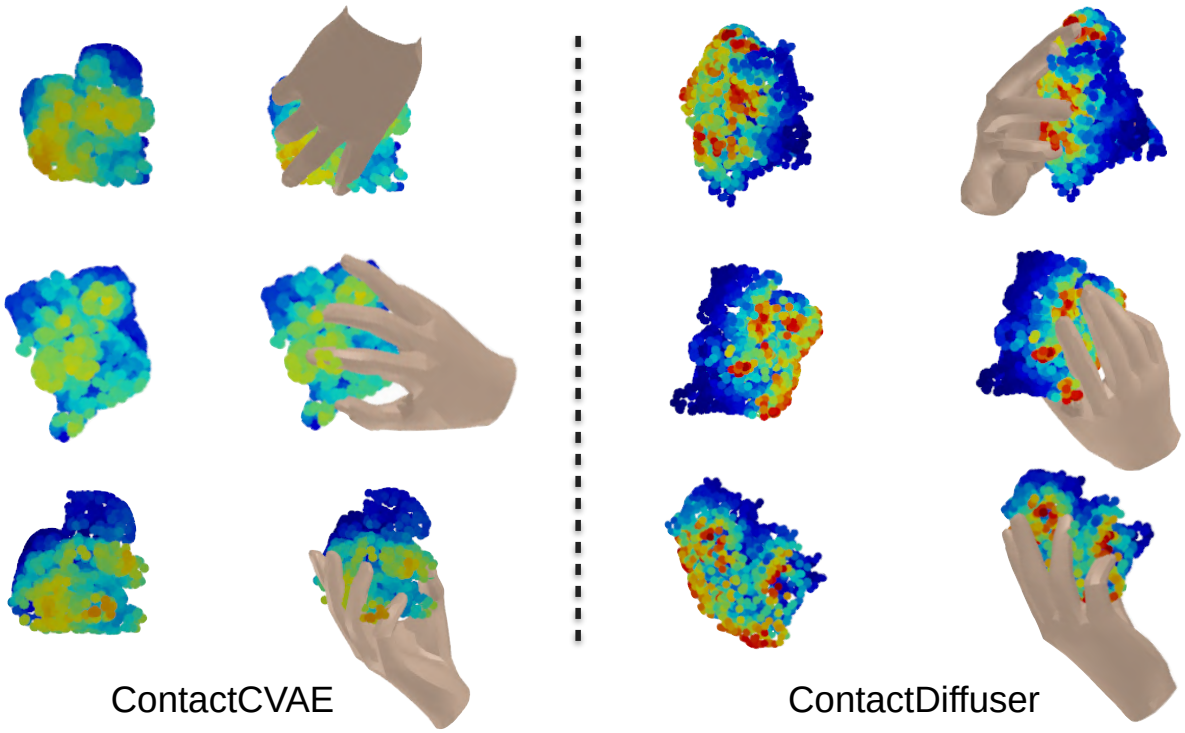

Figure 38. Visualization of predicted contact map and grasp on **doll** from ContactCVAE [3] and Ours

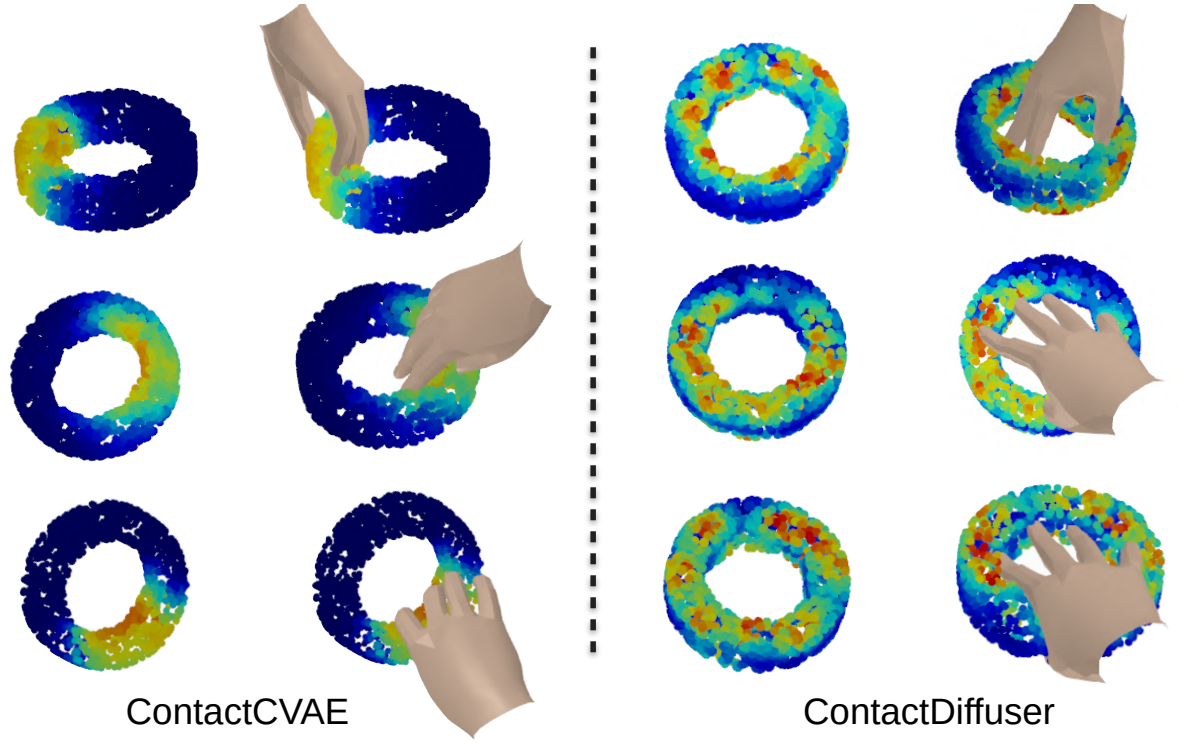

Figure 39. Visualization of predicted contact map and grasp on **tape** from ContactCVAE [3] and Ours

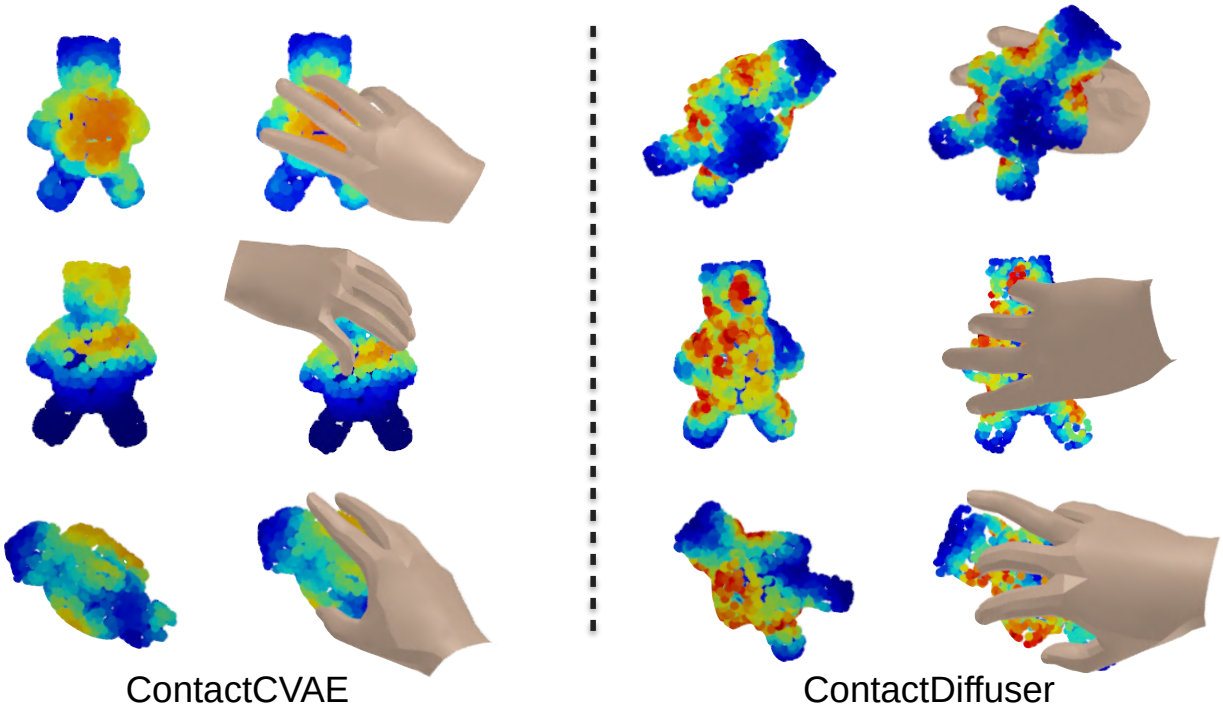

Figure 40. Visualization of predicted contact map and grasp on **bear doll** from ContactCVAE [3] and Ours

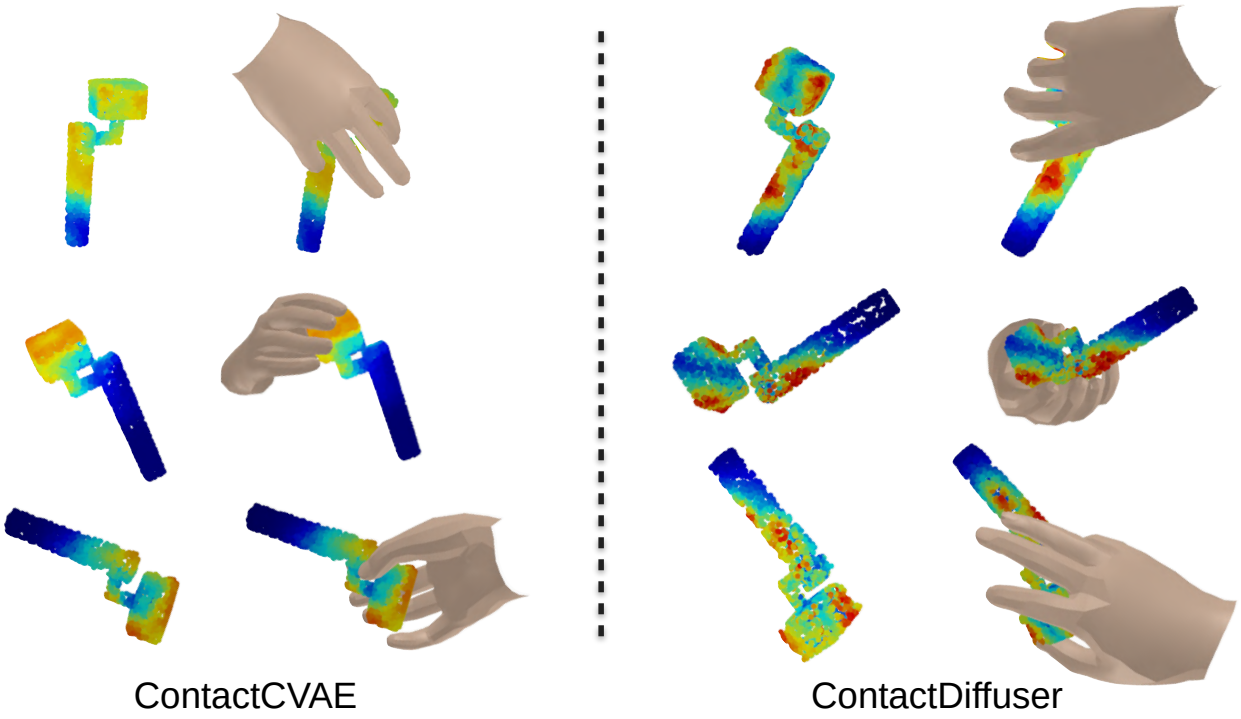

Figure 41. Visualization of predicted contact map and grasp on **camera 2** from ContactCVAE [3] and Ours

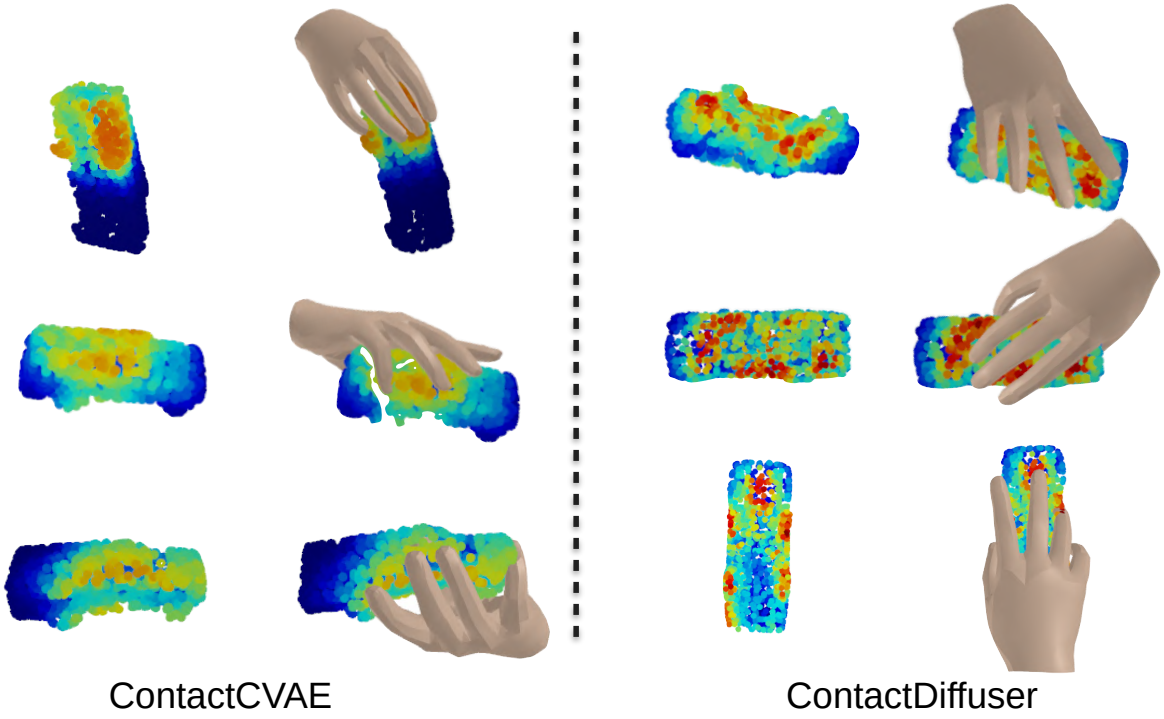

Figure 42. Visualization of predicted contact map and grasp on **toy car** from ContactCVAE [3] and Ours

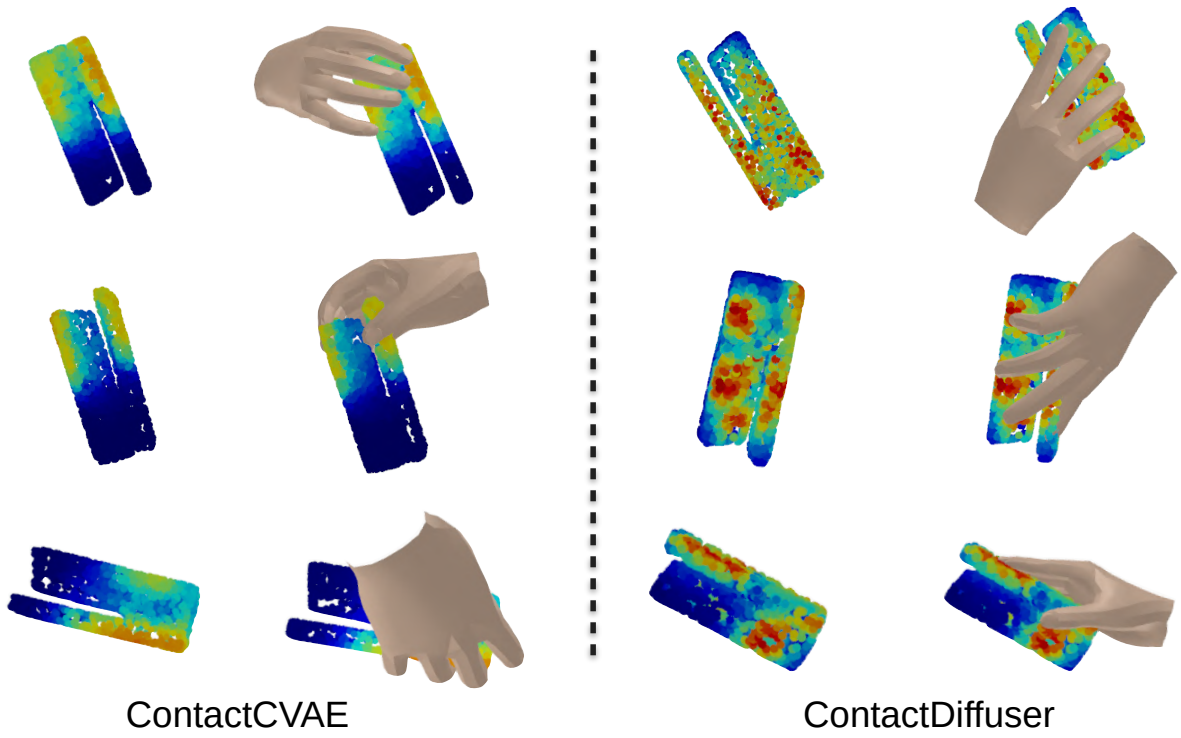

Figure 43. Visualization of predicted contact map and grasp on **stapler** from ContactCVAE [3] and Ours

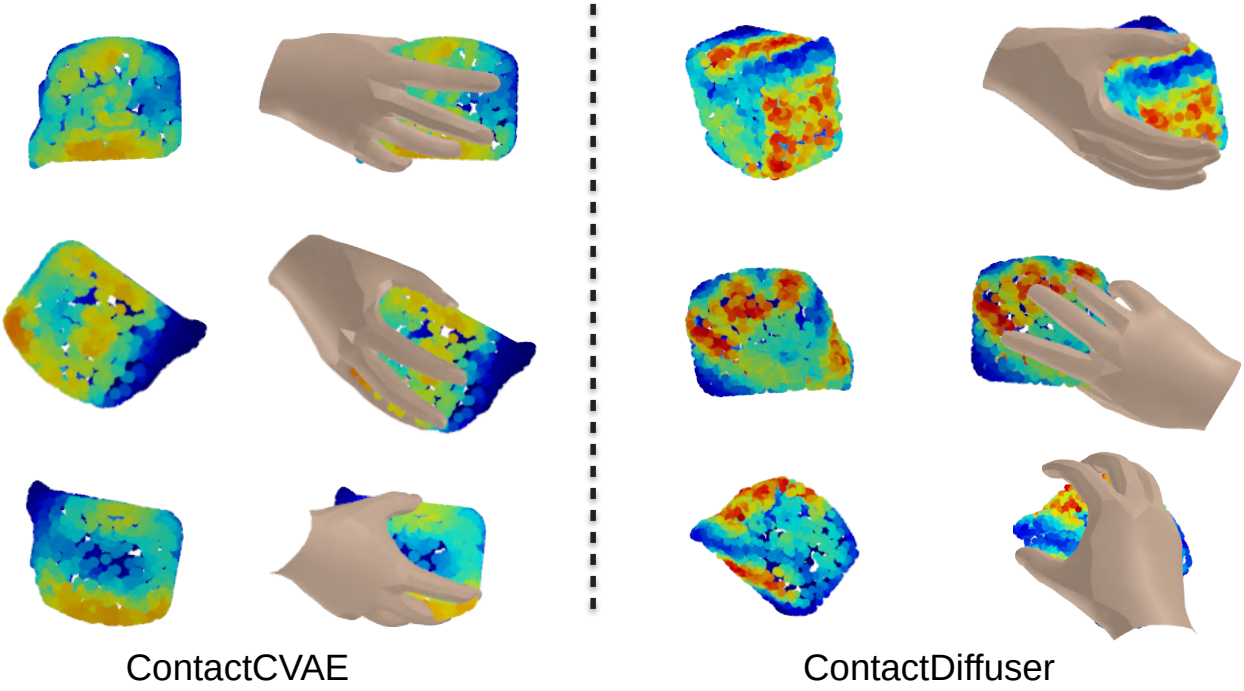

Figure 44. Visualization of predicted contact map and grasp on **tape measure** from ContactCVAE [3] and Ours

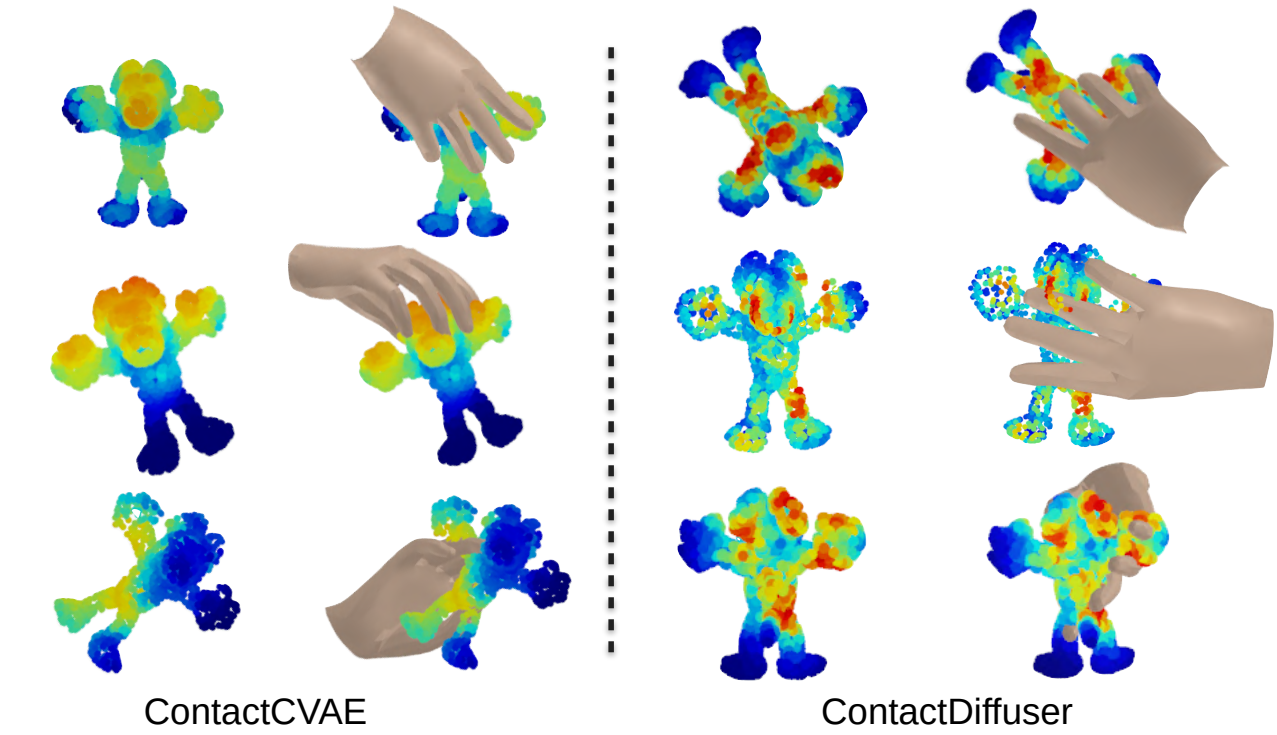

Figure 45. Visualization of predicted contact map and grasp on **toy** from ContactCVAE [3] and Ours

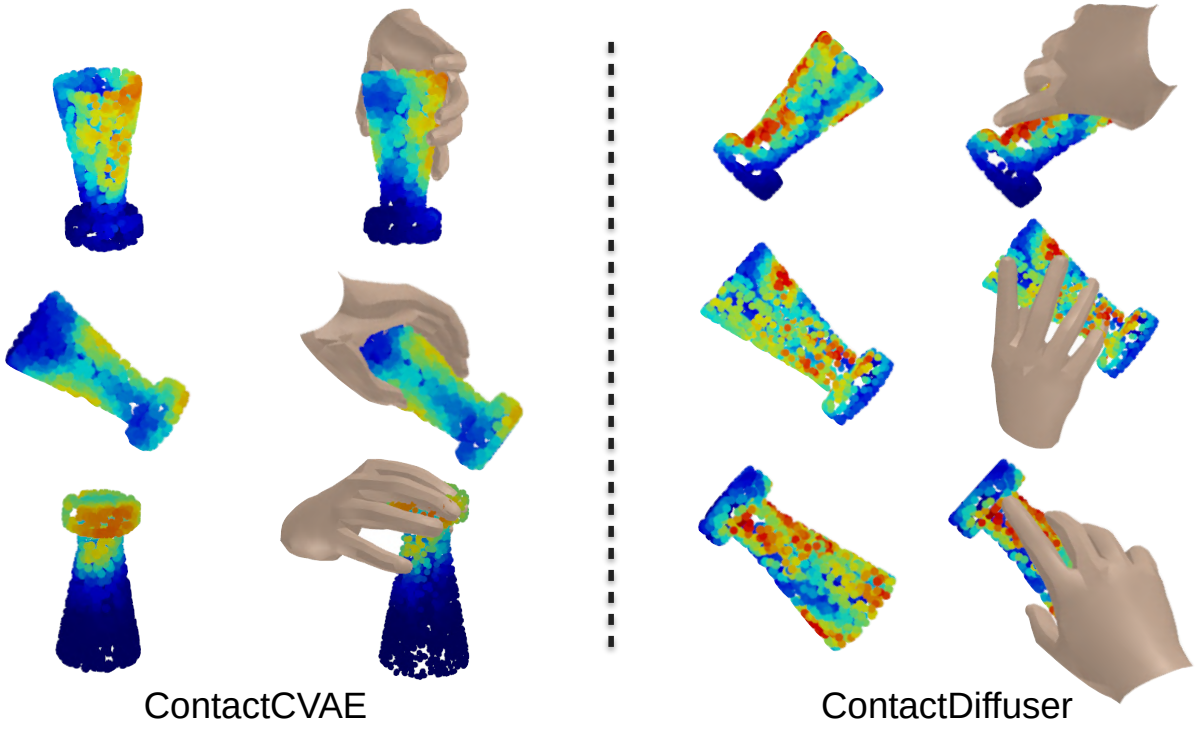

Figure 46. Visualization of predicted contact map and grasp on **trophy** from ContactCVAE [3] and Ours

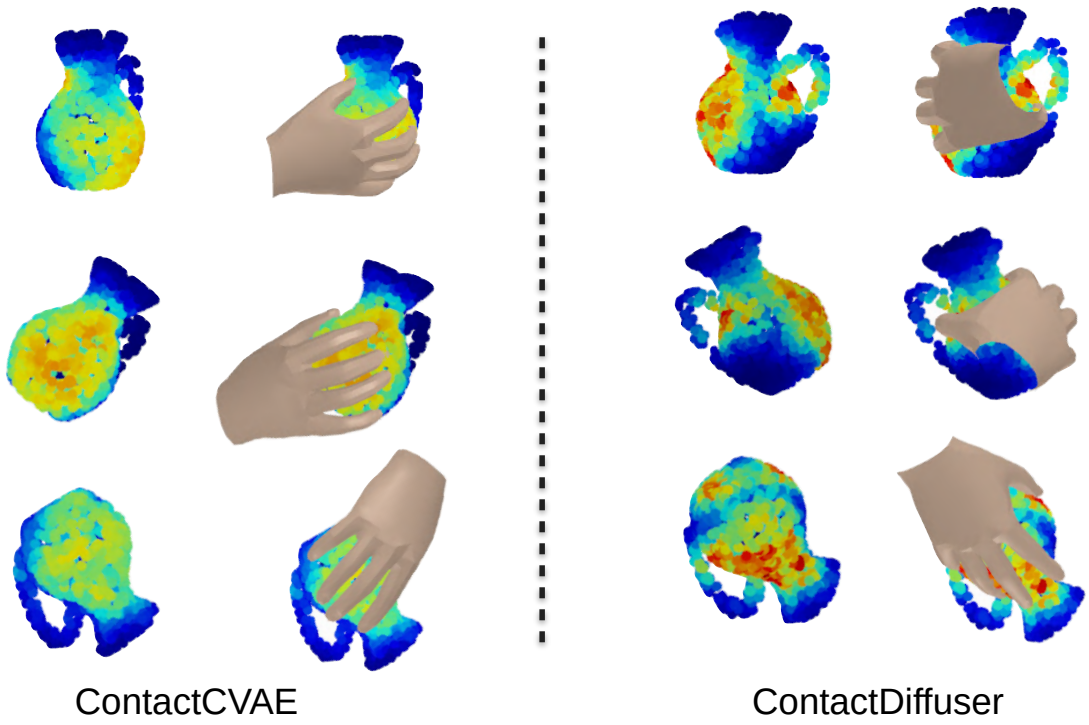

Figure 47. Visualization of predicted contact map and grasp on **vase** from ContactCVAE [3] and Ours

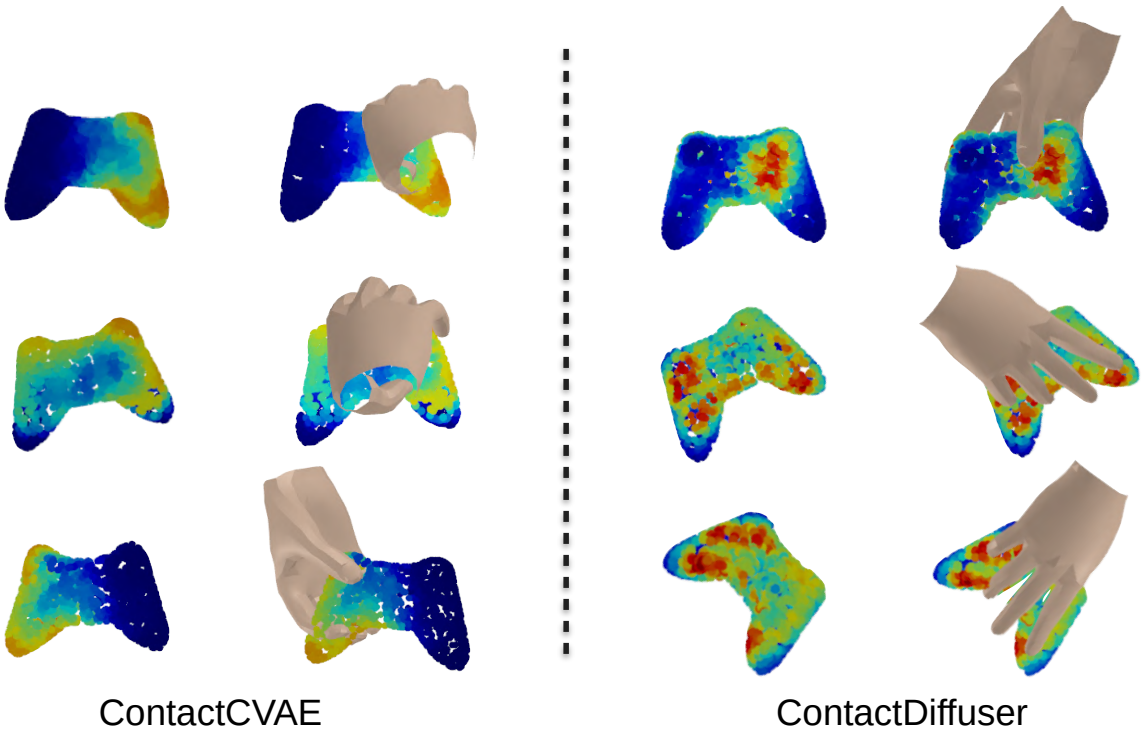

Figure 48. Visualization of predicted contact map and grasp on **video game controller** from ContactCVAE [3] and Ours

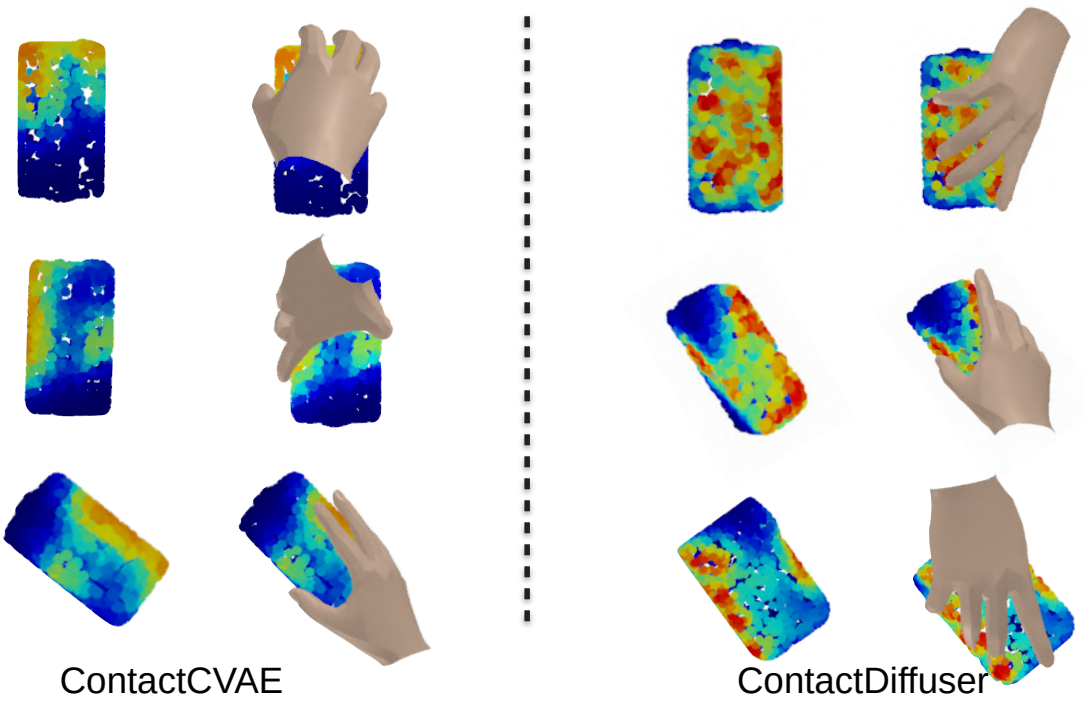

Figure 49. Visualization of predicted contact map and grasp on **camera** from ContactCVAE [3] and Ours

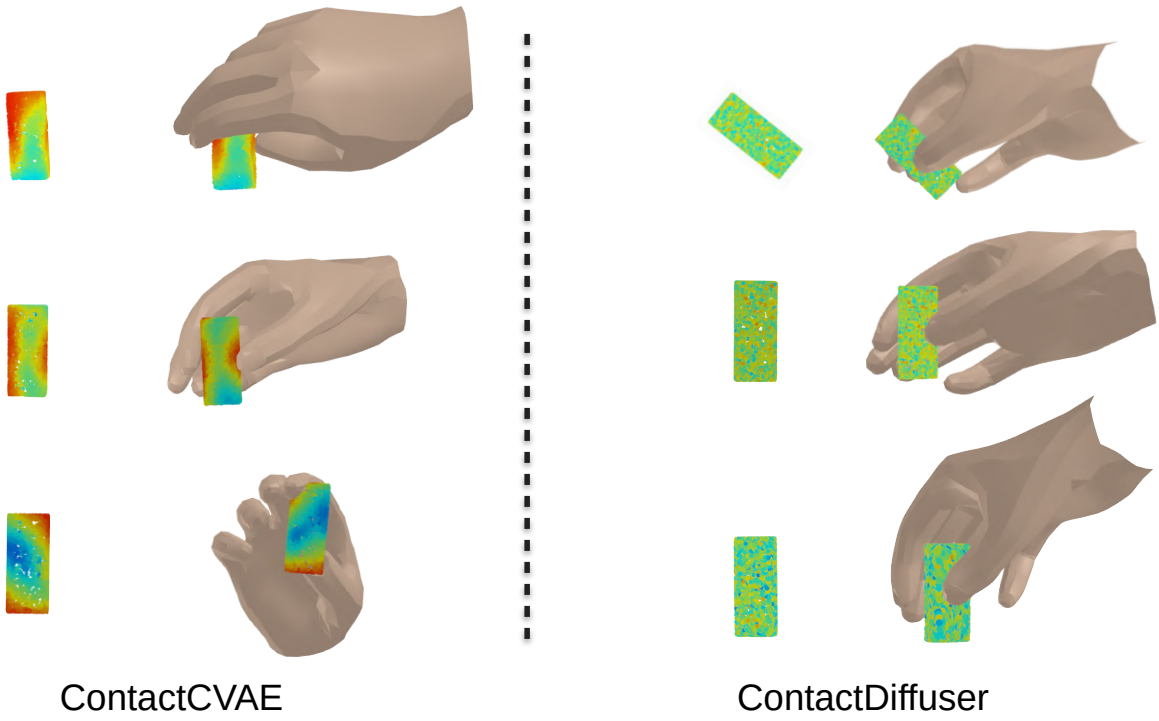

Figure 50. Visualization of predicted contact map and grasp on **brick F** from ContactCVAE [3] and Ours

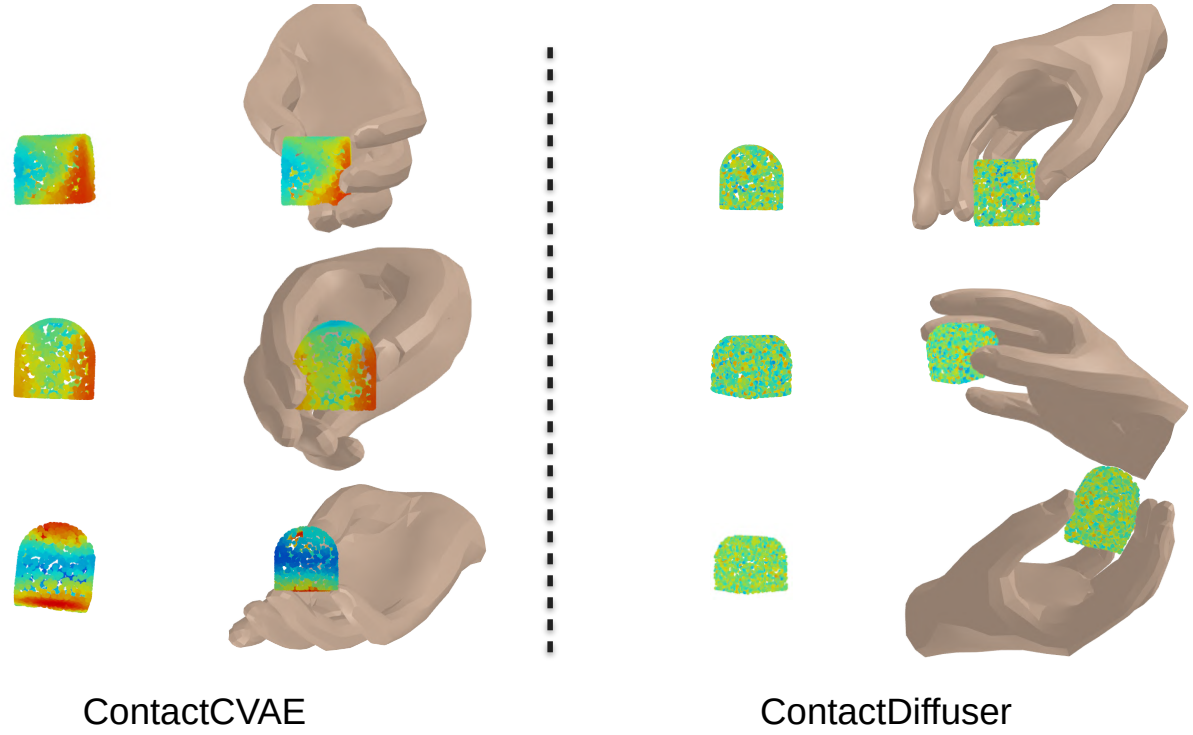

Figure 51. Visualization of predicted contact map and grasp on **brick I** from ContactCVAE [3] and Ours

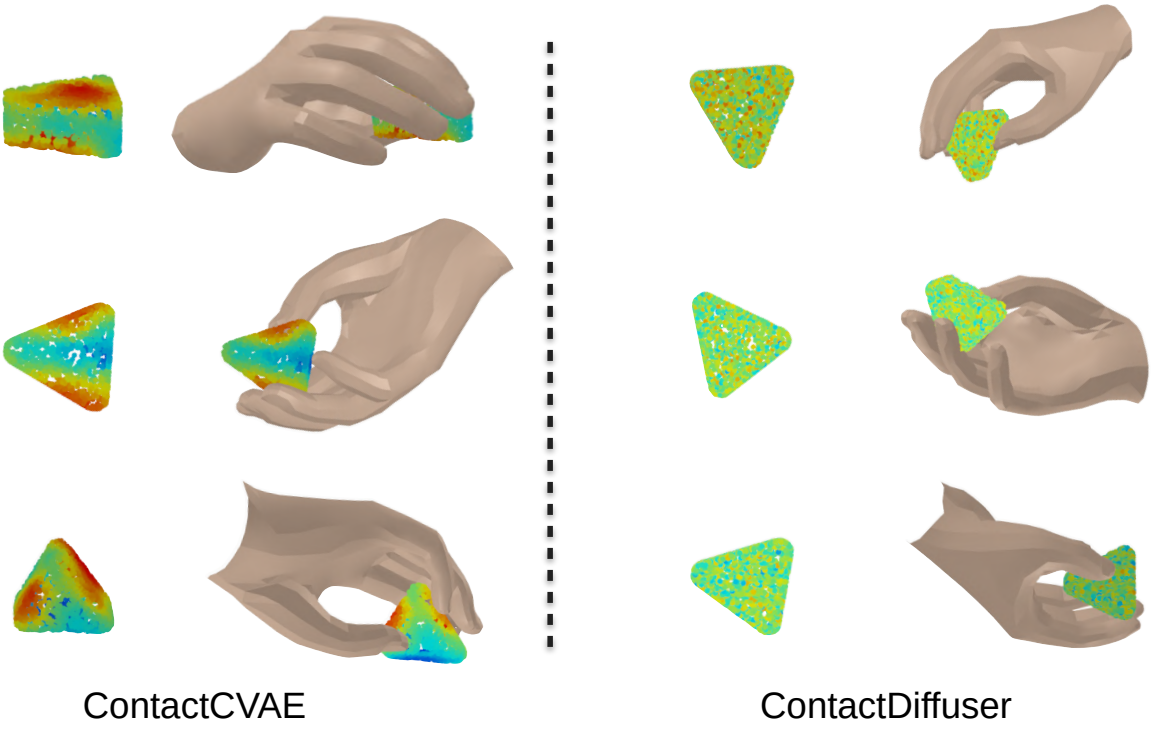

Figure 52. Visualization of predicted contact map and grasp on **brick K** from ContactCVAE [3] and Ours

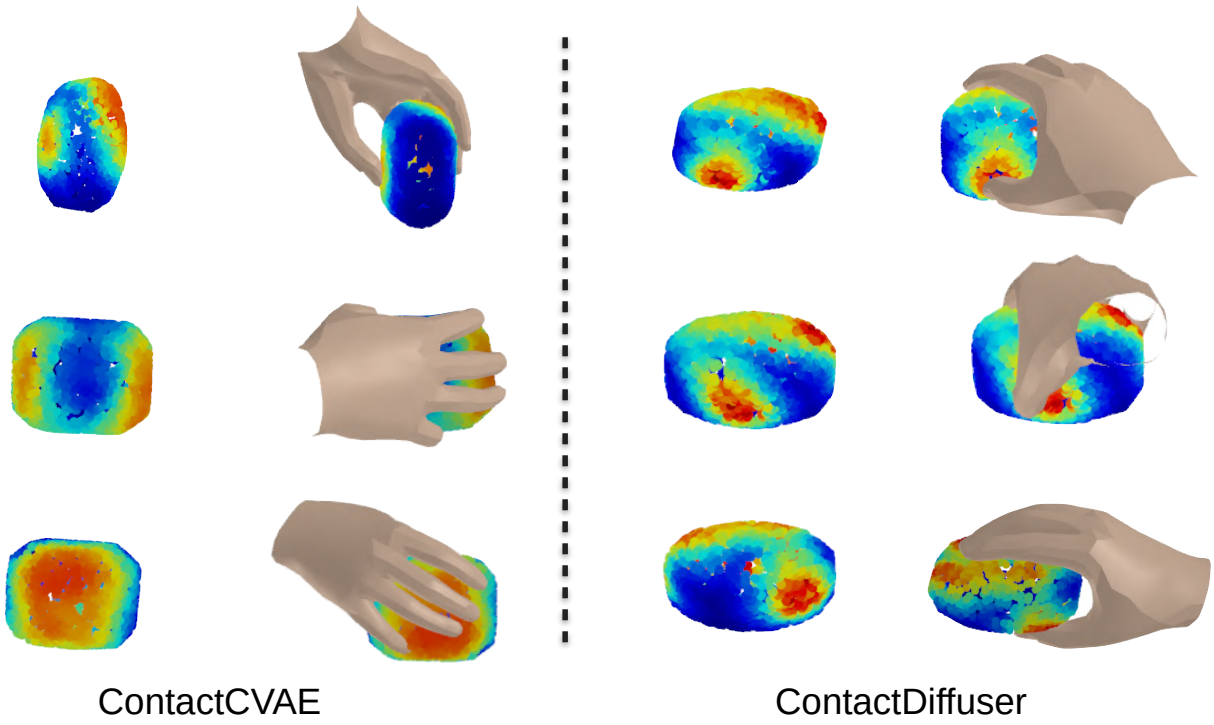

Figure 53. Visualization of predicted contact map and grasp on **brick N** from ContactCVAE [3] and Ours

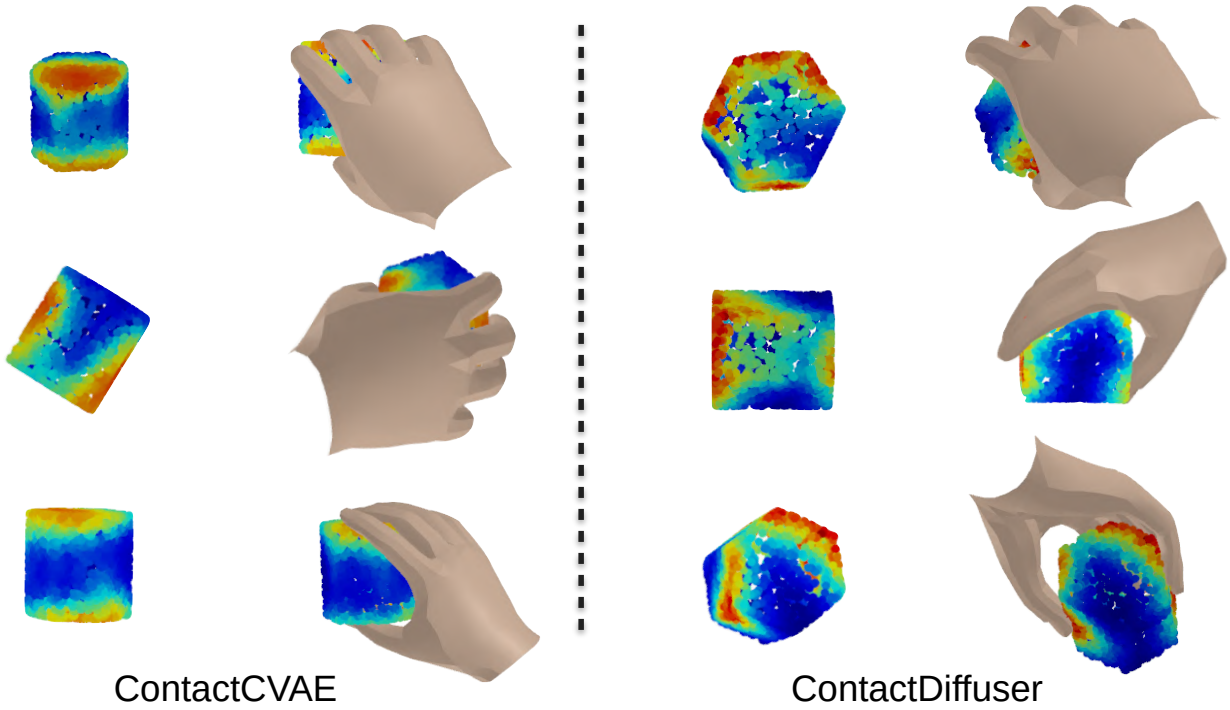

Figure 54. Visualization of predicted contact map and grasp on **brick R** from ContactCVAE [3] and Ours

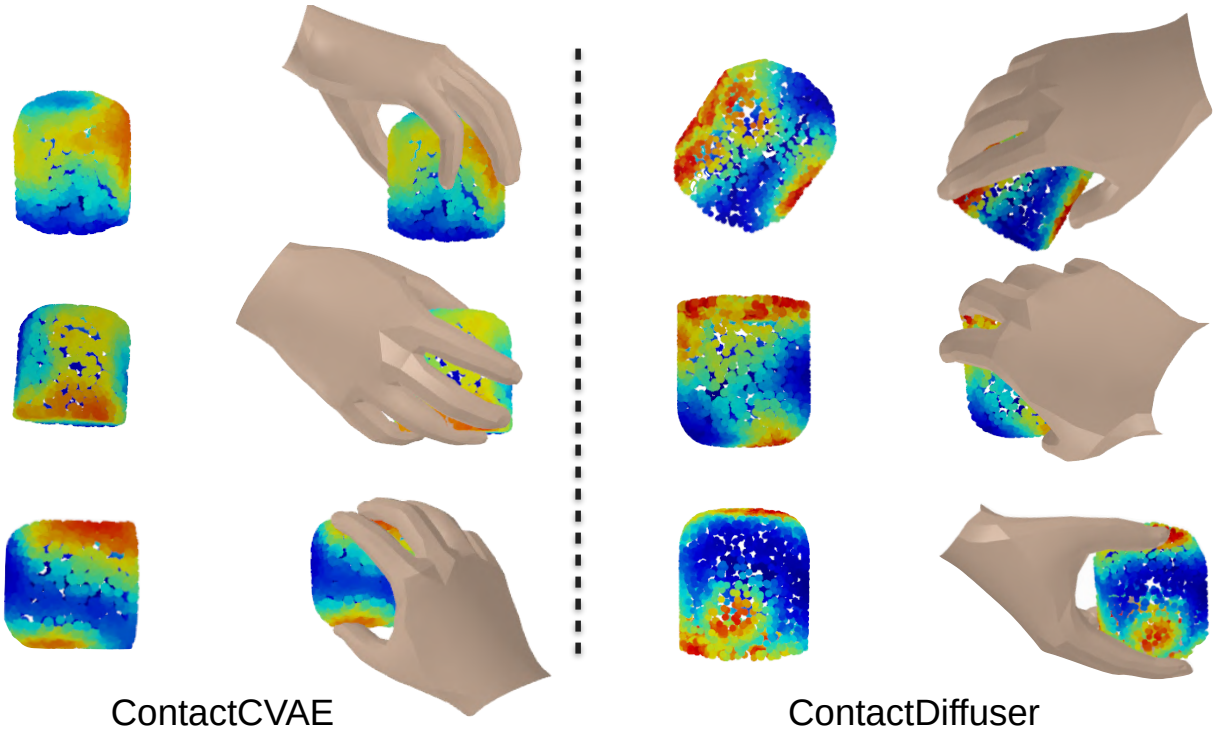

Figure 55. Visualization of predicted contact map and grasp on **brick V** from ContactCVAE [3] and Ours

References

[1] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 5, 6, 7, 8, 9

[2] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 5, 6, 7, 8, 9

[3] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20609–20620, 2023. 1, 4, 5, 6, 7, 8, 9, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[6] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 5, 6, 7, 8, 9

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1