

A. Supplementary material

B. Method details

In this section, we include additional information regarding our representation and learning method.

B.1. CHOIR: Anchor assignment

Table 4. Average reconstruction error for MANO meshes fitted onto ground-truth CHOIRs with the *ordered* and *random* anchor assignment schemes. Mean Per-Joint Pose Error (MPJPE) and Mean Per-Vertex Pose Error (MPVPE) are averaged across the entire ContactPose [8] dataset.

| | <i>Ordered</i> | <i>Random</i> |
|------------|----------------|---------------|
| MPJPE (mm) | 0.18 | 0.19 |
| MPVPE (mm) | 0.22 | 0.22 |

Tab. 4 shows that both the *ordered* and *random* anchor assignment schemes produce the same reconstruction error when fitting a ground-truth CHOIR from the ContactPose [8] dataset. The Mean Per-Joint Pose Error (MPJPE) and Mean Per-Vertex Pose Error (MPVPE) metrics were averaged across the entire dataset. Note that with ground-truth hand-object meshes, the obtained CHOIR allows fitting a MANO mesh with less than 1mm error.

B.2. Test-Time Optimization: Fitting loss

The Python code for the stage 1 of the TTO loss fits in a few lines of code:

```
1 anchor_dist = torch.cdist(  
2     bps, anchors  
3 ) # Anchors predicted in TTO  
4 distances = torch.gather(  
5     anchor_dist, 2, anchor_ids  
6 )  
7 choir_loss = F.mse_loss(  
8     distances, choir[... , -1]  
9 ) # Agreement of anchors and CHOIR
```

Source Code 1. Minimal Python code for the stage 1 TTO loss.

B.3. Keypoint baseline

To evaluate the expressiveness and efficacy of each component of CHOIR, we design a diffusion model backbone that allows us to fit a simpler alternative to CHOIR. This simpler representation only encodes the hand pose and shape as 21 MANO joints $\mathbf{j}_H \in \mathbb{R}^{21 \times 3}$ and 32 MANO anchors $\mathbf{a}_H \in \mathbb{R}^{32 \times 3}$. The object is encoded as a vector of K randomly sampled surface points $\mathbf{p}_O \in \mathbb{R}^{K \times 3}$ where we set $K = 4096$ to match CHOIR which uses a grid of

$16 \times 16 \times 16$ basis points. The final keypoint representation is defined as

$$\mathbf{r}_{\text{kp}} = [\mathbf{p}_O \in \mathbb{R}^{K \times 3}, \mathbf{j}_H \in \mathbb{R}^{21 \times 3}, \mathbf{a}_H \in \mathbb{R}^{32 \times 3}]. \quad (13)$$

However, as in *JointDiffusion*, this model learns to predict the hand part only, defined as

$$\mathbf{r}_{\text{kp}}^H = [\mathbf{j}_H \in \mathbb{R}^{21 \times 3}, \mathbf{a}_H \in \mathbb{R}^{32 \times 3}] \quad (14)$$

The backbone of this diffusion model is composed only of residual blocks made of multi-layer perceptrons (MLPs). We use 4 residual blocks with a hidden dimensionality of 512.

In effect, in this baseline, we only replace the 3D U-Net component of *JointDiffusion* with a residual MLP and remove the contact prediction branch, while keeping cross-attention and the same timestep conditioning scheme. The context encoder is also replaced with a residual MLP of hidden dimensionality 2048. We experimented with a PointNet++-based encoder but observed a degradation in performance.

B.4. Runtime costs

To evaluate the computational costs of CHOIR, we timed its computation and that of TOCH [53] for 50 grasps on an RTX 2080Ti and Intel i9-7900X. On average, TOCH takes $\sim 8.89\text{s}$ (± 3.99) while CHOIR takes $\sim 0.13\text{s}$ (± 0.015), a $68\times$ reduction. When looking at the total inference time, including the model representation computation, forward pass and TTO, ours converges in $\sim 49\text{s}$ (± 16) and TOCH in $\sim 23\text{s}$ (± 4.3). Our diffusion model accounts for $\sim 13\text{s}$ of the total (27%), hence is a major runtime bottleneck. Diffusion Models are inherently slow, but they are becoming faster, and new alternatives with similar properties can be easily integrated since our representation is agnostic to the learning method.

C. Additional experiments and results

C.1. Evaluation metrics

In our experiments, we use the following metrics to evaluate the fitted hand mesh to the predicted CHOIR:

- **Mean Per-Joint Pose Error (MPJPE)/(R-MPJPE) (mm)**: L2 norm between ground-truth (GT) and predicted hand joints. We compute both absolute (MPJPE) and root-aligned (R-MPJPE) metrics. The former tells us about the position of the hand around the object, and the latter tells us about the hand grasp error regardless of the spatial pose.
- **Intersection Volume (IV) (cm^3)**: A measure of hand-object mesh penetration. It is computed by voxelizing the hand and object meshes (1mm voxels) and computing the volume of the intersecting voxels.

- **Hand contact F1/precision/recall (%)**: The precision and recall scores are measured on binary hand contact maps obtained by upsampling the MANO mesh and computing the Chamfer distance to the object point cloud. Hand vertices within 2mm of their nearest object point are considered in contact, to emulate soft tissue deformation as in [16]. A high precision means a low false positives count, while a high recall means a low false negatives count. The F1 score is the harmonic mean of both and is a measure of predictive performance.
- **Simulation Displacement (SD) (cm)**: The distance of displacement of the object in world space when applying inward forces to the hand grasp in a physics simulation. This tells how stable the grasp is, since more hand-object contact patches result in higher friction and therefore lower displacement.

C.2. Perturbed ContactPose

We show a qualitative comparison of our method vs. ContactOpt [16] on several objects. Fig. 7 shows failure cases in some challenging cases. While ContactOpt [16] fails to produce a plausible grasp for each object and noisy input, our method delivers satisfying results that still closely match the contacts of the ground-truth hand pose. Further qualitative samples are shown in Fig. 8, Fig. 9, and Fig. 10, where our method demonstrates fidelity in the reconstructed finger contacts, as opposed to ContactOpt [16].



















| Method | Ground truth | Observation | Prediction |
|-----------------------|---|---|---|
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| ContactOpt | | | |

Figure 7. Failure cases on a comparison of *JointDiffusion* and ContactOpt for the Perturbed ContactPose benchmark. While ContactOpt consistently fails at producing a plausible mesh after multiple restarts, our method results in minimal penetration and respected finger contacts with only one sample.







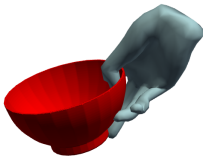
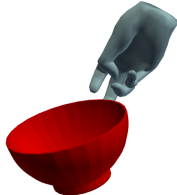
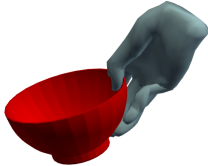
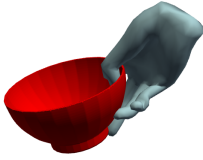
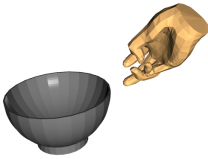




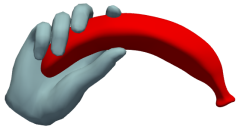


| Method | Ground truth | Observation | Prediction |
|-----------------------|---|---|---|
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |

Figure 8. Qualitative comparison of *JointDiffusion* vs. ContactOpt on the Perturbed ContactPose benchmark. Our method, *JointDiffusion*, produces plausible grasps and maintains the fidelity of finger contacts while ContactOpt fails in challenging cases even with several random restarts. Our method only draws one sample and performs TTO without random restarts.


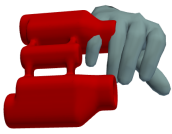


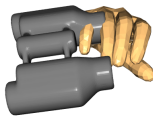


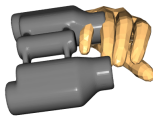


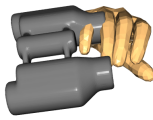

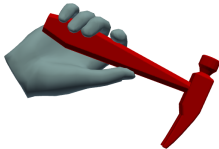


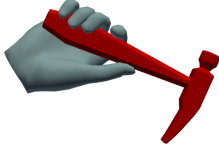


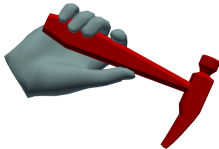


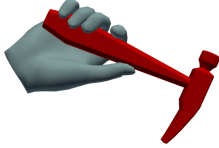





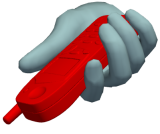





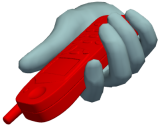


| Method | Ground truth | Observation | Prediction |
|-----------------------|---|---|---|
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>ContactOpt</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>ContactOpt</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>ContactOpt</i> |  |  |  |
| |  |  |  |

Figure 9. Qualitative comparison of *JointDiffusion* vs. *ContactOpt* on the Perturbed ContactPose benchmark. Our method, *JointDiffusion*, produces plausible grasps and maintains the fidelity of finger contacts while *ContactOpt* fails in challenging cases even with several random restarts. Our method only draws one sample and performs TTO without random restarts.











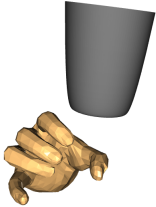






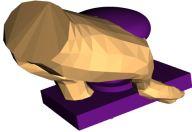
| Method | Ground truth | Observation | Prediction |
|-----------------------|---|---|---|
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |
| <i>JointDiffusion</i> |  |  |  |
| |  |  |  |

Figure 10. Qualitative comparison of *JointDiffusion* vs. ContactOpt on the Perturbed ContactPose benchmark. Our method, *JointDiffusion*, produces plausible grasps and maintains the fidelity of finger contacts while ContactOpt fails in challenging cases even with several random restarts. Our method only draws one sample and performs TTO without random restarts.

C.3. Object splits experiment

To evaluate the generalizability of our method in the grasp refinement setting, we retrain all methods on the Perturbed ContactPose benchmark [16] with object splits instead of subject splits. We hold 2 objects out of the validation split, and reserve 5 objects for the test split, namely: *doorknob*, *eyeglasses*, *apple*, *bowl*, *toothbrush*. This increases the difficulty of the benchmark, as all test objects were unseen during training. For a method to perform well in this setting, it must learn generalizable hand-object interaction in latent space. Tab. 5 shows that our method outperforms ContactOpt [16] on most contact-based metrics, and TOCH [53] on all metrics. ContactOpt [16] retains an edge on the recall score since it maximizes the hand-object contact ratio and therefore minimizes false negatives, but at the cost of less contact fidelity since its precision score is significantly lower than *JointDiffusion*. However, TOCH [53] fails to generalize to these objects, which can be explained by the lack of object representation in the TOCH field. We consider this task to be a main challenge in hand-object interaction understanding and will focus on object generalization in future work.

C.4. Grasp synthesis

Fig. 11 and Fig. 12 show samples of our generative model given an object mesh as input. The model is trained on the improved Perturbed ContactPose benchmark [16], *i.e.* all objects are seen during training. *JointDiffusion* generates visually plausible grasps with consistent finger contacts and minimal mesh penetration. In addition, to enhance visibility, we provide non-cherry-picked supplementary videos of generated hand grasps.

Table 5. Quantitative evaluation of our approach on static grasp refinement against ContactOpt [16] on the Perturbed ContactPose benchmark with object splits. * means reported figures. *JointDiffusion* is evaluated with one non-cherry-picked generated grasp per sample. *JointDiffusion* shows greater contact accuracy and outperforms ContactOpt [16] on most contact metrics, although ContactOpt [16] retains a greater recall score due to its objective which maximizes the hand-object contact ratio, hence reducing false negatives. Best results are in bold, second best are underlined.

| Method | MPJPE (mm) ↓ | R-MPJPE (mm) ↓ | IV (cm ³) ↓ | F1 (%) ↑ | Precision (%) ↑ | Recall (%) ↑ |
|-----------------------|--------------|----------------|-------------------------|--------------|-----------------|--------------|
| Perturbed data | 83.02 | 21.55 | 6.99 | 1.55 | 1.88 | 2.74 |
| ContactOpt [16] | 35.05 | 29.13 | <u>12.83*</u> | <u>15.39</u> | <u>12.04</u> | 30.36 |
| TOCH [53] | 48.27 | 51.13 | 17.63 | 11.18 | 10.74 | 13.54 |
| <i>JointDiffusion</i> | 42.54 | 29.55 | 2.90 | 21.40 | 21.94 | <u>23.05</u> |

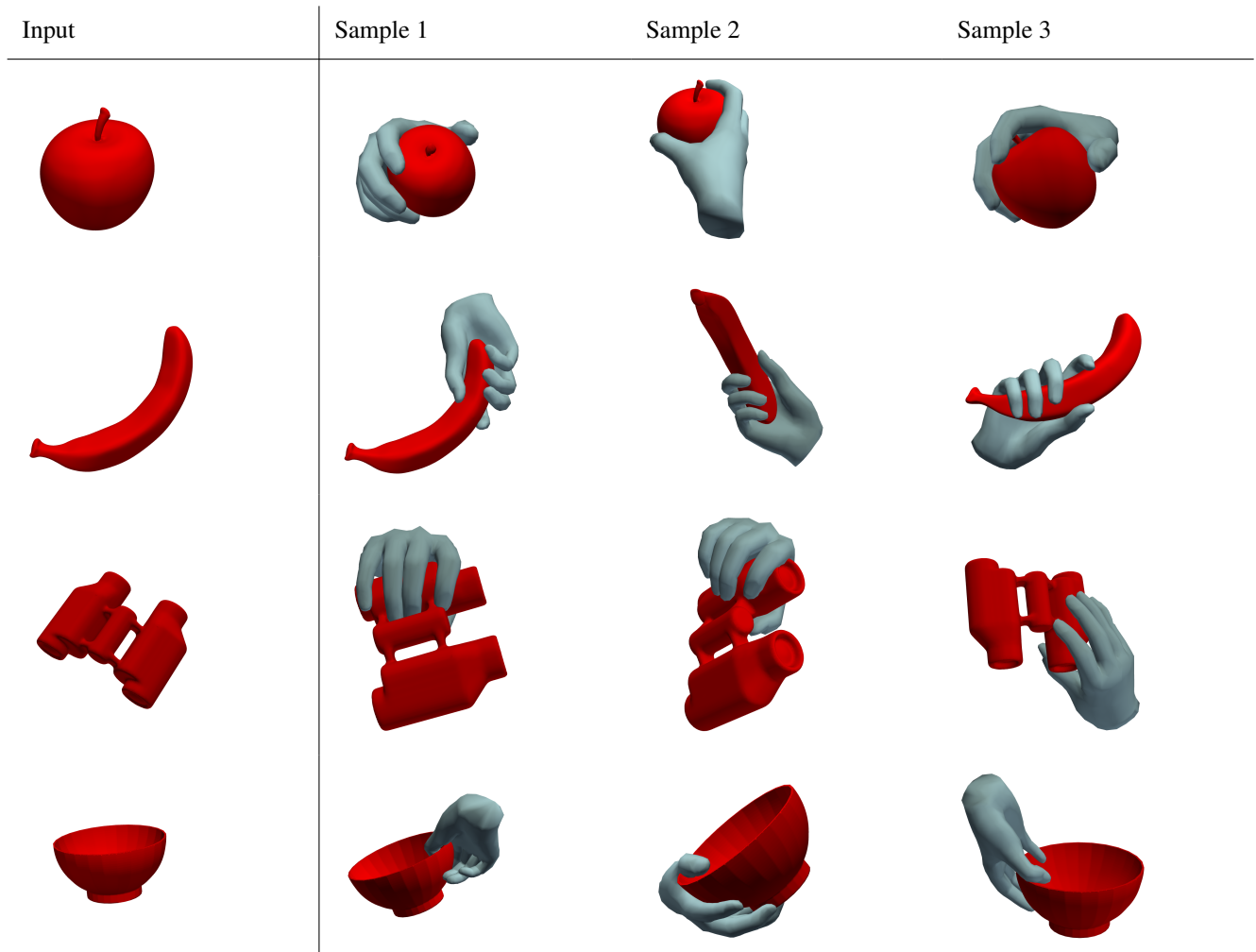


Figure 11. Qualitative evaluation of our method, *JointDiffusion*, trained on the object modality of input for grasp synthesis. Each sample is generated from the same input, the object mesh in canonical pose. *JointDiffusion* produces plausible grasps with minimal mesh penetration and consistent finger contacts.

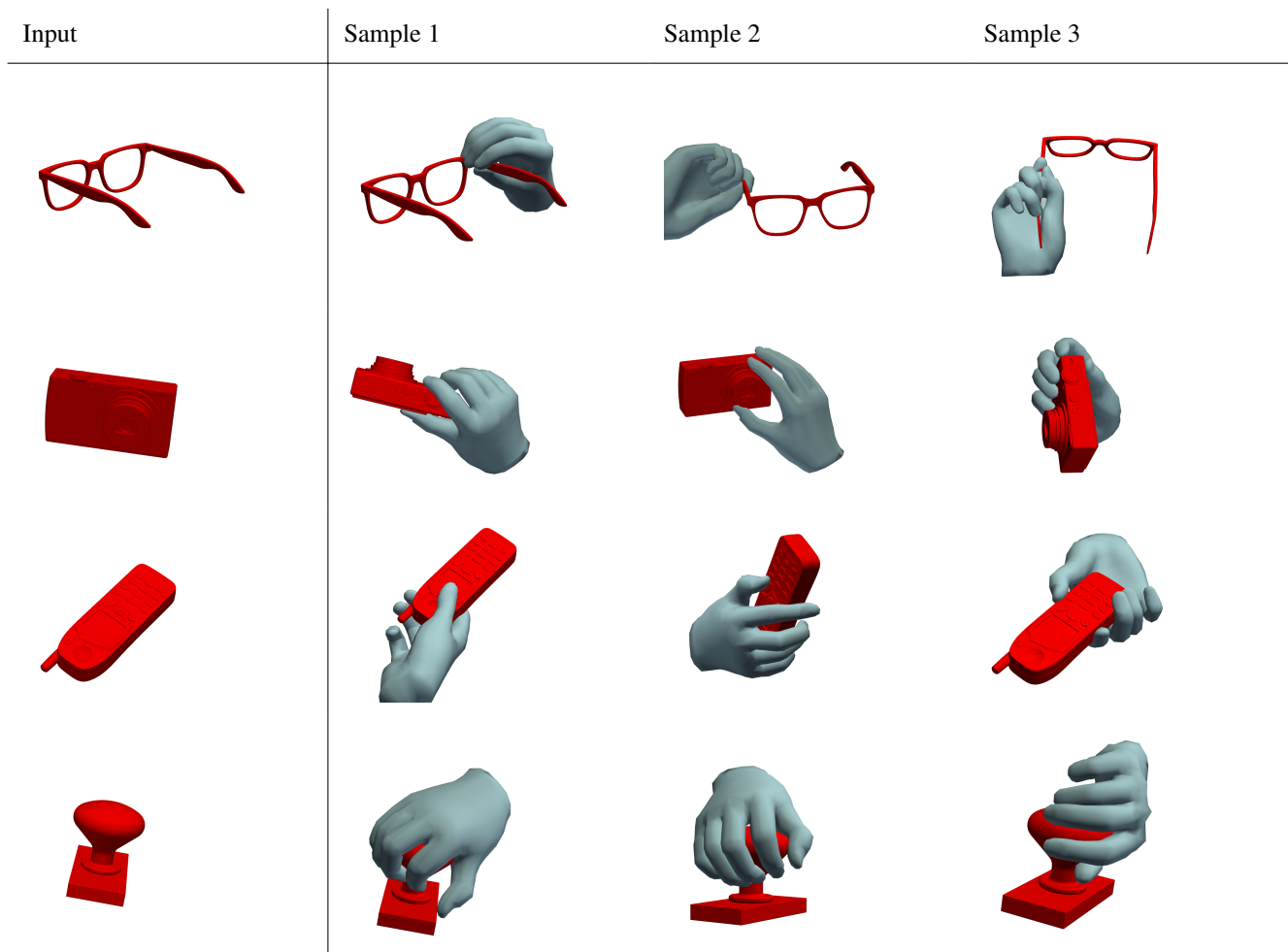


Figure 12. Qualitative evaluation of our method, *JointDiffusion*, trained on the object modality of input for grasp synthesis. Each sample is generated from the same input, the object mesh in canonical pose. *JointDiffusion* produces plausible grasps with minimal mesh penetration and consistent finger contacts.

C.5. Multimodal model: grasp refinement & synthesis

We further explore our model expressiveness by jointly training two context encoders along with the diffusion backbone of *JointDiffusion*, as opposed to separately trained models for object conditioning and noisy hand-object pair conditioning. Fig. 13 shows qualitative results of the grasp synthesis from this model, while Fig. 14 shows qualitative results of the grasp denoising task for the same model. We trained two multimodal models: one on the ContactPose [8] dataset, and one on the OakInk [46] dataset which we only evaluate on grasp synthesis.

A quantitative evaluation on the denoising task is shown on Tab. 7, and one on the generation task is shown on Tab. 6. For the latter, the increase in simulation displacement (SD) for our method with contact fitting suggests that some hand penetration is helpful to a stable grasp. Note that the synthetic nature of most OakInk samples results in incorrect vertex normals, adversely affecting our penetration regularization loss and performance. This could be solved with a different approach to penetration regularization, such as via the signed distance function.

Table 6. Evaluation of our multimodal model on static grasp generation against two state-of-the-art methods on two benchmarks. *JointDiffusion* outperforms GraspTTA [22] on the ContactPose benchmark [8] and is on par with GrabNet [43] on the OakInk benchmark [46]. We used reported metrics for GrabNet [43] from the OakInk paper [46] and sampled one grasp per dataset sample for our method on both benchmarks. Best results are in bold.

| Method | ContactPose [8] | | OakInk [46] | |
|-----------------------|-------------------------|-------------|-------------------------|-------------|
| | IV (cm ³) ↓ | SD (cm) ↓ | IV (cm ³) ↓ | SD (cm) ↓ |
| GraspTTA [22] | 5.17 | 3.81 | - | - |
| GrabNet [43] | - | - | 6.60 | 1.21 |
| <i>JointDiffusion</i> | 5.13 | 5.80 | 5.98 | 5.84 |

Table 7. Evaluation of our approach on static grasp refinement against two SOTA methods and our baseline on the Perturbed ContactPose benchmark. * means reported figures. Our multimodal model is marked with †. Both *JointDiffusion* variants were evaluated with one non-cherry-picked generated grasp per sample. While our baseline yields better reconstruction accuracy in absolute pose, our full model *JointDiffusion* shows greater contact accuracy and outperforms ContactOpt [16] and TOCH [53] on almost all metrics. The multimodal version still outperforms these baselines on contact-based metrics and IV score for grasp refinement, while also being able to do grasp synthesis. Best results are in bold, second best are underlined.

| Method | MPJPE (mm) ↓ | R-MPJPE (mm) ↓ | IV (cm ³) ↓ | F1 (%) ↑ | Precision (%) ↑ | Recall (%) ↑ |
|-------------------------|--------------|----------------|-------------------------|--------------|-----------------|--------------|
| Perturbed data | 83.02 | 21.55 | 6.99 | 1.55 | 1.88 | 2.74 |
| ContactOpt [16] | 32.88 | <u>28.17</u> | 12.83* | 17.27 | 13.24 | 34.30 |
| TOCH [53] | 26.96 | 29.24 | 10.14 | 22.23 | 21.46 | 25.09 |
| <i>JointDiffusion</i> | 27.69 | 23.54 | <u>6.04</u> | 27.20 | 25.21 | <u>32.80</u> |
| <i>JointDiffusion</i> † | 35.45 | 33.10 | 5.62 | <u>24.88</u> | <u>23.87</u> | 29.24 |

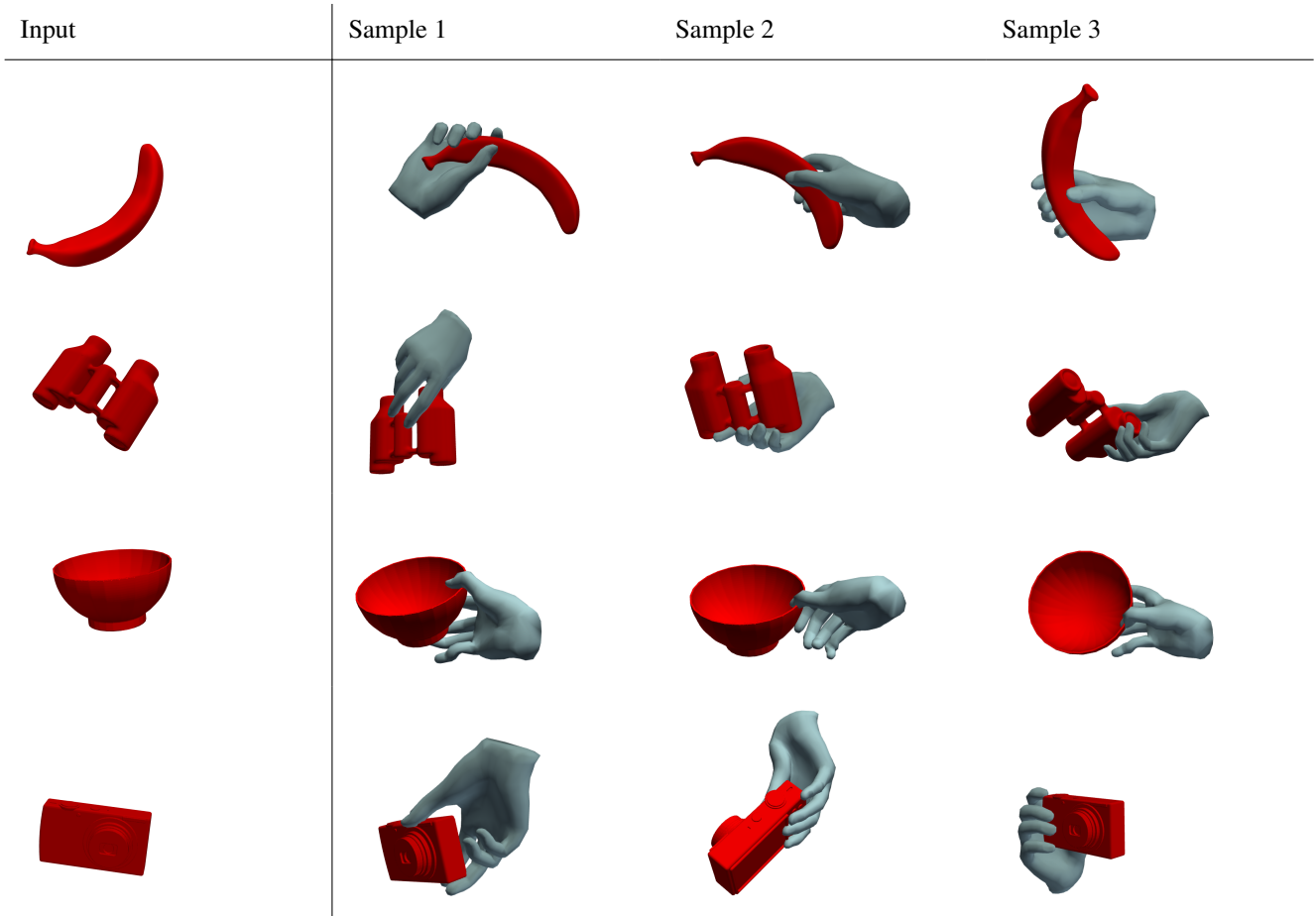


Figure 13. Qualitative evaluation of our multimodal *JointDiffusion*, trained on both object and noisy hand-object pair modalities, in the grasp synthesis setting. Each sample is generated from the same input, the object mesh in canonical pose. *JointDiffusion* produces plausible grasps with minimal mesh penetration and consistent finger contacts.

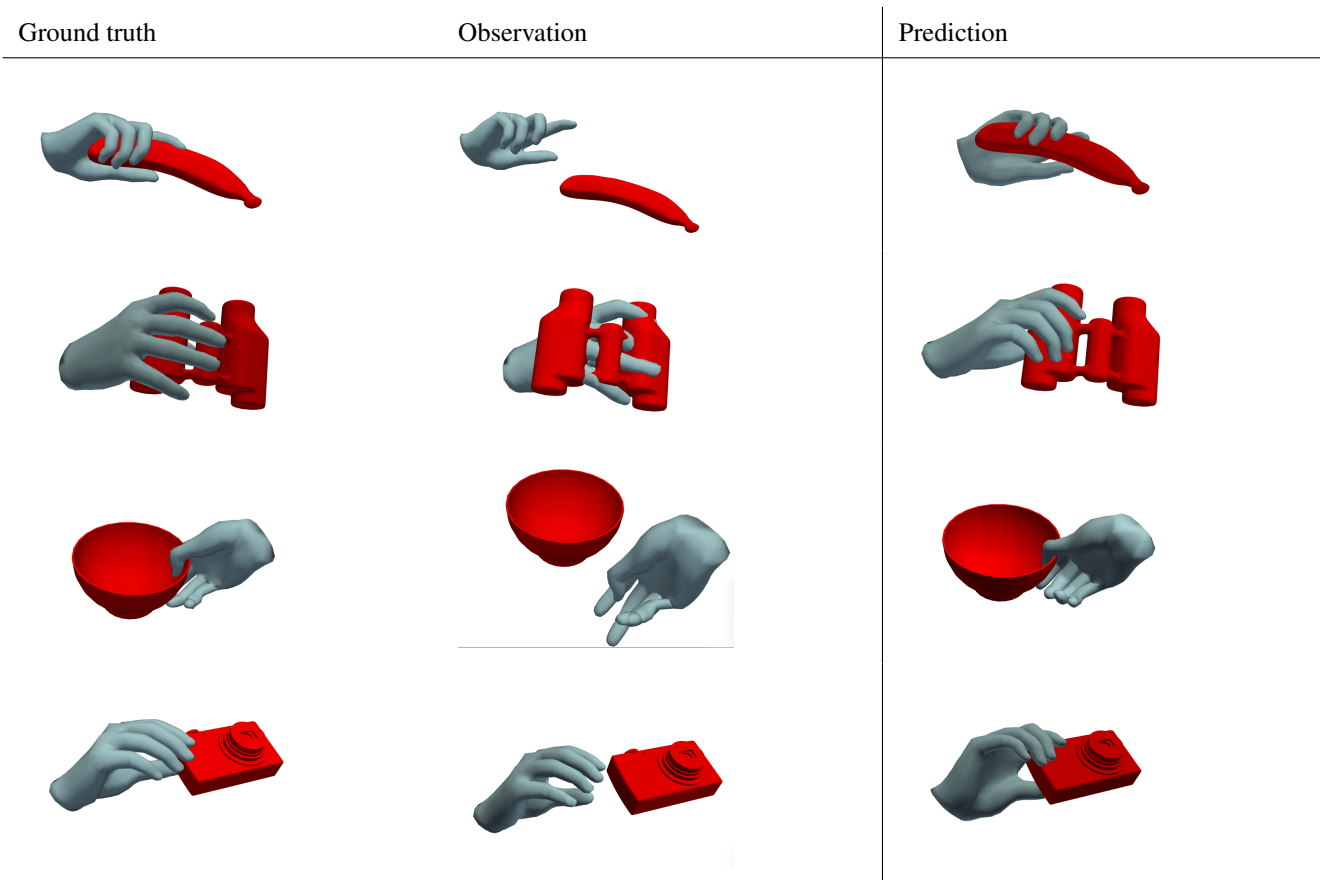


Figure 14. Qualitative evaluation of our multimodal *JointDiffusion*, trained on both object and noisy hand-object pair modalities, in the grasp refinement setting. *JointDiffusion* produces plausible grasps with minimal mesh penetration and respects finger contacts from the ground-truth mesh.