VolCo: Volumetric Contact for High-Fidelity Human Grasp Generation

Zhuo Chen[1]
zxc417@student.bham.ac.uk

Yihua Cheng[2, 1*]
yihua.cheng@bit.edu.cn

Aleš Leonardis[1]
a.leonardis@bham.ac.uk

Hyung Jin Chang[1]
h.j.chang@bham.ac.uk

[1]University of Birmingham, [2]Beijing Institute of Technology

Abstract

Accurate contact modeling is fundamental to understanding hand–object interaction, yet existing contact representations are typically restricted to object surfaces and rely on hand-crafted rules to recover contact details, leading to severe penetrations and implausible results. To better exploit the rich detail in motion-capture data, we introduce **Vol**umetric **Co**ntact (VolCo), a representation that expands surface points to a set of 3D volumetric grids. VolCo encodes 3D contact that allows precise hand part recovery, and is organized in an inherent hierarchy: local contact details within each volume and global hand geometry across all volumes. Our framework, VolCoDiff, employs two modules to capture local and global features following this hierarchy. For local contact details, we use a 3D variational autoencoder to model the possible hand configurations conditioned on the local object signed distance field (SDF). For global hand geometry, we design a prior-guided diffusion model that learns the distribution of compressed latent features aggregated from the volumetric grids. We evaluate our method on two benchmark datasets and demonstrate state-of-the-art performance in penetration and stability, indicating the capability to generate tight grasps with much less severe penetrations. Our code is available at `https://github.com/chzh9311/volco`.

1 Introduction

Hands are the primary means by which we interact with our surroundings. Recently, the research community has dedicated significant attention to modeling object grasping [40, 47, 44, 21, 23] due to its broad applications in AR/VR [11], robotics [22, 41], and human-computer interaction [12].

As a fundamental problem in this field, human grasp generation aims to synthesize plausible human hand grasps for a given object template, where MANO [32] is used as the mainstream hand representation. MANO parameters are predicted directly [34, 17, 49, 39] or fitted from 3D fields [19, 43, 27] or contact maps indicating the contact regions, hand part orientations, etc.,[47, 24, 45, 23, 27, 6].

Existing methods mostly rely on hand-crafted rules in test-time-adaptation (TTA) to control penetrations [10]. Such TTA has limited adaptivity since the rules rely on object surface shapes. As shown in Fig. 1, penetration losses based on nearest-neighbor (NN) queries is non-robust, changing gradient directions when the point crosses the medium line. Moreover, slight penetrations are expected at contact regions with rigid models like MANO, but this method cannot control the exact penetration depth. Consequently, some generated grasps still show severe penetrations and limited plausibility.

*Corresponding author

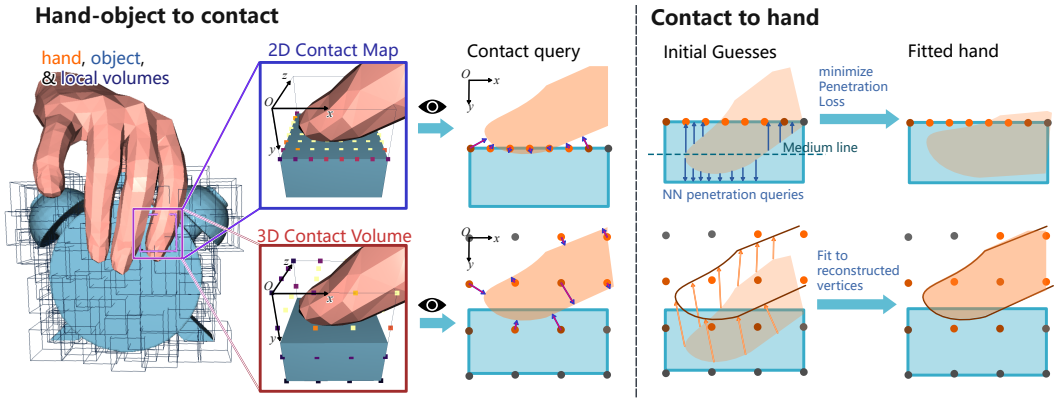


Figure 1: 3D contact volumes vs. 2D contact maps. Previous contact maps query points only on object surfaces, which can retain only a small fraction of hand surface. Thus it requires nearest-neighbor (NN) queries to avoid penetrations, which can easily stick in local minima. Contact volumes, however, sample points around the object surfaces and is able to retain the whole local hand configuration in the volume. Consequently, contact volumes can deterministically and precisely reconstruct local hand part and guide the pose fitting without instable penetration losses.

Our target is to retain the fine-grained contact details from training data to the largest extent, thus reducing severe penetrations and improving geometric plausibility. Our insight is that the points to reason contact, i.e.,contact points, should not only lie *on* the object surface, but should be scattered *around* the surface. In this way, the contact can cover much larger area of the hand and preserve the exact penetration depth, thus no rules with NN queries are required. Further, as 3D grid representations [34, 43, 27] are limited by computational cost [25], these points should be hierarchically organized to find the balance between computational cost and granularity.

To this end, we introduce **Vol**umetric **Co**ntact (VolCo), a novel contact representation constructed from local volumetric grids whose centers are uniformly sampled on the object surface, inspired by Mosaic-SDF [42]. While Mosaic-SDF employs such local volumes to represent general object geometry via signed distance values, VolCo extends the paradigm to contact modeling of a fixed topology (hand) by capturing also the correspondences. Specifically, each point in a volume encodes the contact likelihood and the correspondence using Continuous Surface Embeddings (CSE) [28]. Furthermore, VolCo aggregates point-level features within each local volume. This coherent hierarchy allows the separate modeling of local configurations and global hand priors. Based on the hierarchy, we present a volume-wise latent diffusion model i.e.,VolCoDiff, consisting of a 3DVAE i.e.,VolumeVAE and a latent diffusion model based on Scene Diffuser [15].

VolumeVAE aggregates the common prior over local volumes conditioned on the SDF values, modeling *what hand surface configurations are geometrically possible within each volume*. Since each volume is small (2^3cm^3), the feasible configuration space is limited in scale and thus possible to be compressed to a compact latent code. VolumeVAE abstracts away the fine-grained surface geometry, allowing the diffusion model to focus solely on the global hand prior.

The latent diffusion model is trained to capture *what combinations of local hand configurations constitute a plausible grasp*. This objective is governed by two key principles. First, all local hand parts must be *reachable* by a valid hand. To enforce this, an auxiliary branch is incorporated in the diffusion UNet to predict the hand pose, which is subsequently mapped to VolCo latent space to supervise the generated contact. Second, the grasp should be *stable*. Due to the joint encoding of SDFs and contact likelihood, VolCo allows the explicit contact force inference via a spring-damper model, and thus a force-aware stability loss can be applied [6]. Together, these two principles encourage the generation of physically plausible hand part combinations, while VolumeVAE ensures all local penetrations remain within plausible bounds. Finally, the generated grasp is obtained by initializing with the hand-branch prediction and subsequently fitting it to the generated VolCo representation.

In experiment, we benchmark hand pose reconstruction from contact and compare VolCo with previous representations. VolCo shows the highest precision, indicating its capability to preserve the maximum hand information. We also benchmark the grasp generation capability on two public datasets, and our method achieves the highest stability, lowest penetration depth, and highest plausible

contact area. It indicates the state-of-the-art grasping generation quality. In all, our contribution can be summarized in threefold:

1. We propose VolCo, a novel volume-based hierarchical contact representation capable of recovering the fine details of hand-object interaction;
2. We construct a VolCoDiff, a latent diffusion model for grasp generation that takes advantage of the hierarchical structure of VolCo by modeling local contact details using VAE and global hand geometry using diffusion;
3. We conduct experiments on pose reconstruction from contact and grasp generation, demonstrating state-of-the-art performance on both tasks.

2 Related Work

Contact Modeling. Contact modeling has shown great potential in hand-object estimation and generation [17, 10]. Early approaches define contact as per-point likelihood on the object surface [10, 17, 36], providing guidance on grasp location but leaving a substantial information gap to complete hand meshes. Later, the contact maps get enriched by including part / anchor label [40, 27, 49], hand continuous surface embedding [28, 44], part orientations[24], and contact forces [6]. Nevertheless, all such representations remain confined to the object surface and lack explicit modeling of penetration depth, leaving its recovery hand-crafted rules.

3D Hand-Object Representations. Some recent methods use implicit functions [19, 5] or 3D volumetric grids (Bases Point Set [34, 27] or SDF [43, 5]) to represent both hands and objects, thus relating them in a common feature space. Such representations effectively capture relations between hands and objects, but are limited in fitting to known templates or low resolution [25]. On contrary, VolCo is rooted on known templates and finds the balance of fine-details and computational cost.

Human Grasp Generation. Human Grasp Generation methods have evolved significantly in recent years. Early methods [18, 24, 46] utilize cVAE [20] as the backbone, but are limited by the capacity and posterior collapse [1] of VAEs. Subsequent research [33, 44, 49, 3, 23] gradually shifts to diffusion models [14], with latent diffusion models [31] recently taking dominance. The compressed latent space can reduce the number of parameters [49] and restrict pose plausibility [21]. Nevertheless, these methods lay their attention on global hand pose, ignoring the contact details. Another line of work try to recover the contact first, then fit the hand to the contact maps [40, 17, 47, 24, 27, 6]. However, due to the incomplete and non-deterministic nature of the contact representations, the optimization requires NN based penetration losses [32] and risks getting stuck in local minima.

In summary, prior methods typically requires test-time adaptation with non-differentiable penetration losses for better generation quality. We attribute this limitation to the poor information preserving capability of existing contact maps, which VolCo is designed to address.

3 Methodology

Our framework is shown in Fig. 2. VolCo and hand segments within local volumes are mutually convertible. Based on this feature, VolCoDiff is trained to generate the latent and decoded to VolCo, which recovers the hand configurations around the object and gets fitted by MANO.

3.1 Volumetric Contact Definition

From Hand-Object to VolCo. Consider an arbitrary point $\mathbf{x} \in \mathbb{R}^3$ carrying the information of the hand triangular mesh. The contact is represented by $[c_\mathbf{x}, \mathbf{e}_\mathbf{x}]$ where $c_\mathbf{x} \in \mathbb{R}$ is the contact likelihood and $\mathbf{e}_\mathbf{x} \in \mathbb{R}^4$ is the 4-D Continuous Surface Embedding (CSE) [28] vector of the nearest point on the hand surface. The CSE is obtained following [44]. Suppose the nearest point position is $\mathbf{y}$ and the distance is d. The vertices of the triangle this point belongs to and the corresponding CSEs are $(\mathbf{v}_i, \mathbf{e}_i), i = 1, 2, 3$, then the contact is defined in ContactOpt [10] style and the CSE vector is calculated using the barycentric coordinate to keep continuity as $\mathbf{y}$ moves along the object surface.

$$c_\mathbf{x} = \min\{d_0^2/d^2, 1\};\ \mathbf{e}_\mathbf{x} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \mathbf{e}_3] \left([\mathbf{v}_1 \quad \mathbf{v}_2 \quad \mathbf{v}_3]^{-1} \mathbf{y} \right). \tag{1}$$

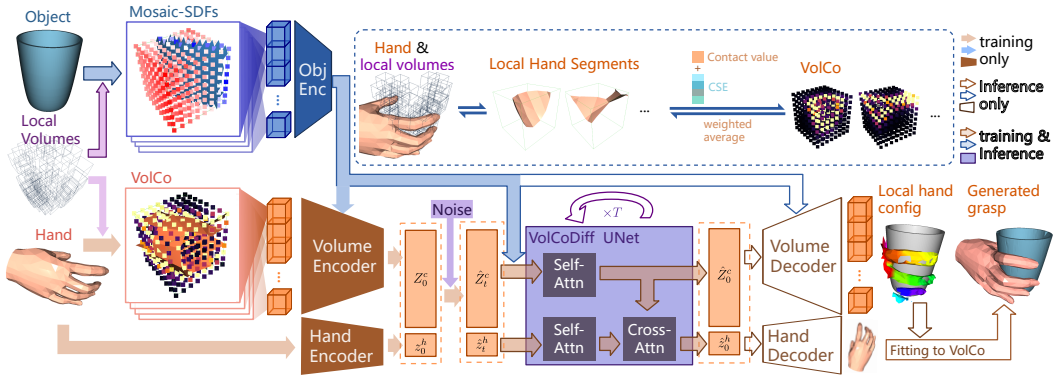


Figure 2: **VolCo and VolCoDiff framework**. Within the local volumes sampled from the object surface, the hand mesh and VolCo representation are bidirectionally transferable. In each volume, a uniform grid is sampled, and each grid point encodes the contact likelihood and CSE, forming VolCo. Conversely, local hand configurations can be reconstructed from each volume using weighted average and concatenated to recover the hand geometry. VolCoDiff operates on the latent codes extracted from VolCo via VolumeVAE, with an auxiliary hand latent vector for supervision. The generated VolCo restores high-fidelity hand meshes around the object, with the hand decoder output subsequently fitted to VolCo to produce the final grasp.

We then discretize the representation and calculate point-wise contact vectors as illustrated in Fig. 3. Following [42], we sample N points $\{\mathbf{p}_i\}_{i=1}^N$ from the object surface, and cover the object surface by expanding each point to a local volume with half size s. Then the volume is discretized with a resolution of k. The volumetric grid points are defined by $\mathcal{G}_i = \{\mathbf{x} = \mathbf{p}_i + s \cdot (2\mathbf{j} - k - 1)/(k-1)\}_{\mathbf{j}=(i_1,i_2,i_3)\in[1,k]^3}$

Define the unit distance as half the grid gap $d_0 = s/(k-1)$, and VolCo for this grid can be written as

$$C_i = \{[c_\mathbf{x}, \mathbf{e}_\mathbf{x}]\}_{\mathbf{x}\in\mathcal{G}_i}. \tag{2}$$

From VolCo to Hand. One key advantage of VolCo the capability to reconstruct the hand part within the volume. For each point contact $(\hat{c}_\mathbf{j}, \hat{\mathbf{e}}_\mathbf{j}) \in G_i^h$, we find the triangle $\mathbf{t}_\mathbf{j}$ on the mesh with the nearest CSE (defined by the average CSE of 3 vertices) to $\hat{\mathbf{e}}$. Then, according to Eq. (1), the corresponding weight of the nearest point to the 3 vertices can be equivalently calculated by $\mathbf{w}_\mathbf{j} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \mathbf{e}_3]^{-1}\hat{\mathbf{e}}_\mathbf{j} \in \mathbb{R}^3$. The weight vector can then be extended to $\bar{\mathbf{w}}_\mathbf{j} \in \mathbb{R}^{778}$ covering all vertices by padding the rest indices with 0. For a specific vertex h that appears at least once in the nearest triangle query process, i.e.,$\mathbf{v}_h \in \cup_{\mathbf{j}\in[1,k]^3}\mathbf{t}_\mathbf{j}$, the position can be deterministically estimated as the weighted average of all volume points using

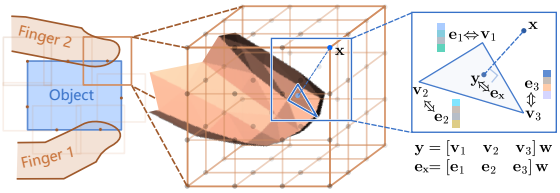


Figure 3: VolCo hierarchy and calculation. Around each of N object surface points, a $k \times k \times k$ volume grid is sampled, where each grid point carries the contact likelihood as a function of d and the CSE of its nearest hand surface point $\mathbf{y}$.

$$\hat{\mathbf{v}}_h = \frac{\sum_{\mathbf{j}\in[1,k]^3}\hat{c}_\mathbf{j}\bar{w}_{\mathbf{j}h}\mathbf{x}_\mathbf{j}}{\sum_{\mathbf{j}\in[1,k]^3}\hat{c}_\mathbf{j}\bar{w}_{\mathbf{j}h} + \varepsilon}, \quad \varepsilon = 10^{-8}. \tag{3}$$

In our experiment, we set $k = 8$ by default, while $k = 4$ setting is also reported to study the influence of resolution on efficiency and performance. s is fixed to 1cm since natural hand deformations cannot exceed 1cm. Fixed scale also facilitates capturing consistent hand local configurations regardless of the object scale. Empirically, $N = 128$ volumes roughly cover all objects in YCB benchmark [4].

3.2 Modeling Contact Prior using VolumeVAE

The hierarchical structure of VolCo allows us to decouple the modeling of local contact details and global geometry. This section mitigates the local hierarchy, answering the question: **How to model the distribution of feasible local hand configurations given the SDF values?**

Since local hand configurations are limited in shape variations and only occupy a part of the contact feature space, we propose to use a 3D VAE, i.e.,VolumeVAE, following the practice in 3D U-Net [7] and 3DVAEs [2], to compress the contact volume $C \in \mathbb{R}^{k\times k\times k\times(1+4)}$ into a compact 128-dimensional latent vector conditioned on the SDF values at each point $D = \{\text{SDF}_o(\mathbf{x})\}_{\mathbf{x}\in\mathcal{G}}$.

Losses. Formally, VolumeVAE models the conditional distribution $p(C \mid D)$ via a standard VAE formulation with reparameterization, where the encoder outputs (μ, σ) and the decoder outputs $\hat{C}$ from the sampled latent. Generally, VAE supervision is done by reconstruction loss $\mathcal{L}_{rec}$ and KL-Divergence loss $\mathcal{L}_{KL}$. In our case, $\mathcal{L}_{KL} = D_{\text{KL}}\left(q_\theta(z \mid C, D) \,\|\, p(z)\right)$ where $p(z) = \mathcal{N}(0, 1)$. However, the common reconstruction loss cannot precisely reconstruct the part hand surface because the contact and correspondences need to be treated differently.

For contact, a simple MSE loss is enough: $\mathcal{L}_c = \|c_i - \hat{c}_i\|_2$. According to Eq. (3), we also regularize the CSE loss using contact values: $\mathcal{L}_{cse} = c_i\|\mathbf{e}_i - \hat{\mathbf{e}}_i\|_1$. Ideally, a reconstruction loss of hand vertices should be applied, yet in practice the nearest triangular face lookup is not differentiable. Therefore, we propose the *CSE weight loss*, which supervises the 3-dimensional barycentric weights based on ground truth triangle. Conceptually, we want $\hat{\mathbf{w}} = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3\end{bmatrix}^{-1}\hat{\mathbf{y}}$. Due to the differentiability issue, we equivalently have $\hat{\mathbf{w}} = \begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3\end{bmatrix}^{-1}\hat{\mathbf{e}}$ according to Eq. (1), leading to a continuous loss $\mathcal{L}_{csew} = \|\hat{\mathbf{w}} - \mathbf{w}\|_1$. The final loss for training the VolumeVAE is:

$$\mathcal{L}_{vae} = \lambda_c\mathcal{L}_c + \lambda_{cse}\mathcal{L}_{cse} + \lambda_{csew}\mathcal{L}_{csew} + \beta\mathcal{L}_{KL}. \tag{4}$$

In experiments, we set $\lambda_c = 5.0, \lambda_{cse} = 1.0, \lambda_{csew} = 0.5$. β is set to 10^{-6} following [31].

3.3 Prior-Guided Latent Diffusion

VolumeVAE encodes rich information about possible local hand configurations, so the next step is to answer: **How to model the distribution of possible combinations of local hand configurations?**

Latent diffusion models [31] has recently shown great potential in human grasp generation [38, 49, 21]. We follow this practice, but with one key difference: the latent space is the concatenation of the volume-wise latent code, following the global hierarchy of VolCo.

To minimize the range of latent vectors, we set all volumes without hand vertices inside to zero, which are mapped to a constant latent. The resulting latent is a matrix $Z^c \in \mathbb{R}^{N\times 128}$. This volume-wise latent space integrates naturally with Scene Diffuser [15], a point-conditioned diffusion framework. Inspired by [23], we use the flattened Mosaic-SDF as the feature vector and utilize the PointNet [29] encoder to extract the per-point features. These features are used as the positional encoding in the attention-based UNet.

Auxiliary Branch with Hand Prior Guidance. Generated VolCo should be valid, i.e.,whether they can fit in a valid hand mesh (provided by MANO [32]). Since the VolCo carries complete information to recover the hand, an auxiliary branch to generate hand based on VolCo latents is added. Also, to make sure the generated hand fall in the plausible manifold, the target is a latent $z^h \in \mathbb{R}^{16}$ extracted from an MLP-based VAE model. The final latent space is obtained by concatenation: $z = [z^h, Z^c]$.

In general, diffusion models a Markov noising process $\{z_t\}_{t=1}^N$ and z_0 is the clean sample. Using object SDFs D as the condition, the reverse denoising process is to extract $p(z_0|D)$ by gradually cleaning z_t. We construct the standard diffusion loss following [35] by predicting the clean latent $\hat{z}_0 = f_\phi(z_t, t, D)$ and supervise by the standard loss $\mathcal{L}_{diff} = \|z_0 - f_\phi(z_t, t, D)\|_2^2$.

To allow the hand prior to flow into VolCo latent generation, we develop a single-directional dataflow. The features from the object and VolCo latents serve as the condition and fed into the hand generation branch by cross-attention [37]. With the predicted hand mesh $\langle\hat{V}, \mathcal{F}\rangle$, we can reconstruct the contact via Eq. (2) and then the latent using VolumeVAE decoder. Then the difference loss is calculated by comparing the reconstructed latent $\hat{Z}^c = \{\mu_i + \epsilon\sigma_i\}_{i=1}^{k^3}$ with the predicted $\hat{Z}_0^c$ by $\mathcal{L}_{cons} = \|\hat{Z}^c - \hat{Z}_0^c\|_2^2$.

The predicted hand mesh is also supervised by reconstruction loss $\mathcal{L}_{rec} = \|\hat{V}_0 - V_0\|_2^2$. Since the nearest point lookup is non-differentiable, the latent reconstruction does not carry gradient. The reconstructed hand supervises the generated contact but not vice versa.

Stability Loss Using Contact Forces. Another advantage of VolCo is that the contact force can be explicitly inferred. Suppose k is the object stiffness and δ is the penetration depth. According to the linear spring-damper model [16] for static contact ($\dot{\delta} = 0$): $F = k\delta$. We can explicitly calculate the contact force for each point, as each point of VolCo encodes both the contact value c and the penetration depth as object SDF. Regard two points as adjacent if their distance is smaller than half the grid gap. Assume $\mathcal{A}_{\mathbf{x}}$ is the adjacent point set of point $\mathbf{x} \in \mathcal{G}_i$. We also add a margin $m = 1\text{mm}$ and slack variables $\Delta d_{\mathbf{x}} \in [-1\text{mm}, 1\text{mm}]$ to allow error in penetrations. The force at this point is

$$F(\mathbf{x}) = 1/(|\mathcal{A}_{\mathbf{x}}| + 1)kc_{\mathbf{x}} \cdot \min\{\text{SDF}_o(\mathbf{x}) - m + \Delta d_{\mathbf{x}}, 0\} \cdot \nabla\text{SDF}_o(\mathbf{x}), \tag{5}$$

where $1/(|\mathcal{A}_{\mathbf{x}}| + 1)$ averages the force to the adjacent points, $c_{\mathbf{x}}$ regularizes the force value using contact, $\nabla\text{SDF}_o(\mathbf{x})$ points the force towards the SDF descending direction. The stability loss $\mathcal{L}_{stability}$ is improved from [6] by adding the slack variables. We refer the readers to Appendix. A for detailed formulation.

Finally, with the weights $\lambda_{cons} = \lambda_{rec} = \lambda_{stability} = 0.1$ in experiments, the loss in training the diffusion model is written as

$$\mathcal{L}_{ldm} = \mathcal{L}_{diff} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{stability}\mathcal{L}_{stability}. \tag{6}$$

3.4 Finetuning the Local Details.

Although the hand branch reconstructs a feasible hand, it usually does not align exactly with generated VolCo. Thus, finetuning is usually necessary for high-quality output. Different from literature [24, 43, 6], our goal is nothing but finding the best fit to VolCo near the initial pose provided by hand branch.

From the predicted VolCo, we can obtain correspondence weights $\hat{w}_{\mathbf{x}h}$ between grid point $\mathbf{x}$ and vertex h in Eq. (3), and the contact likelihood $\hat{c}_{\mathbf{x}}$. Suppose M volumes are predicted as in contact (not all-zero), then the reconstruction loss is defined based on the weighted sum of MSE between each grid point and its predicted nearest point on the hand mesh.

$$\mathcal{L}_{vrec} = \frac{1}{Mk^3} \sum\nolimits_{i=1}^{M} \sum\nolimits_{\mathbf{x}\in\mathcal{G}_i} \hat{c}_{\mathbf{x}} \text{MSE}\left(\mathbf{x}, \sum\nolimits_{h=1}^{778} \hat{w}_{\mathbf{x}h}\hat{\mathbf{v}}_h\right). \tag{7}$$

Second, for $N - M$ volumes that are predicted as non-contact, a simple repulsive loss is applied to fit those non-contact volumes. For any $\mathbf{x} \in \mathcal{G}_i$, estimate the contact using its nearest hand vertex as $\tilde{c}_{\mathbf{x}}$. Then the repulsive loss is $\mathcal{L}_{rep} = -\sum_{i=1}^{M} \sum_{\mathbf{x}\in\mathcal{G}_i} \tilde{c}_{\mathbf{x}} \min\{SDF_o(\mathbf{x}), 0\}/(Nk^3 - Mk^3)$.

With the pose from the auxiliary branch as the initial state, a regularization term is applied to punish large modifications: $\mathcal{L}_{reg} = \|\hat{\theta} - \theta_0\|$. Our final loss function for the optimization process is

$$\mathcal{L}_{optim} = \mathcal{L}_{vrec} + \lambda_{rep}\mathcal{L}_{rep} + \lambda_{reg}\mathcal{L}_{reg}. \tag{8}$$

4 Experiment

In this section, we conduct experiments on 2 tasks: reconstructing from ground truth contact and human grasp generation, both done on two public datasets.

Datasets. GRAB [34] is used as our training set and one of the test set due to its high quality motion capture data. HO3D [13] contains 10 objects from YCB dataset [4] and are not seen when training on GRAB. This dataset is used for benchmarking the adaptivity to out-of-domain objects.

Implementation Details. VolumeVAE and VolCoDiff are both trained using AdamW [26] with learning rates 2×10^{-4} and 10^{-4} for 20 and 100 epochs, respectively. Pose fitting uses AdamW for 1,000 iterations. Our experiments are done on a linux workstation with Intel(R) Xeon(R) w7-3445

Table 1: Quantitative results of hand reconstruction from ground truth contact on GRAB test set. '↑' after a criterion means the higher the better, while '↓' means the opposite. 2D contact maps use the same number of sampled points (4096) as our sparse setting, yet our contact still demonstrates much better performance across all criteria.

Method	Corr	Contact	EPE(mm) ↓	F@5mm ↑	F@15mm ↑	AUC ↑
ContactOpt [10]	–	Map	89.42	0.125	0.365	0.067
ContactGen [24]	Part	Map + D	59.38	0.405	0.773	0.619
ManiDext [44]	CSE	Map	11.19	0.616	0.923	0.785
Sparse (64×4^3)	CSE	Volume	6.70$^{\downarrow 40.1\%}$	0.739$^{\uparrow 20.0\%}$	0.970$^{\uparrow 5.1\%}$	0.868$^{\uparrow 10.6\%}$
Normal (128×8^3)	CSE	Volume	**6.33**$^{\downarrow 43.4\%}$	**0.752**$^{\uparrow 22.1\%}$	**0.971**$^{\uparrow 5.2\%}$	**0.875**$^{\uparrow 11.4\%}$

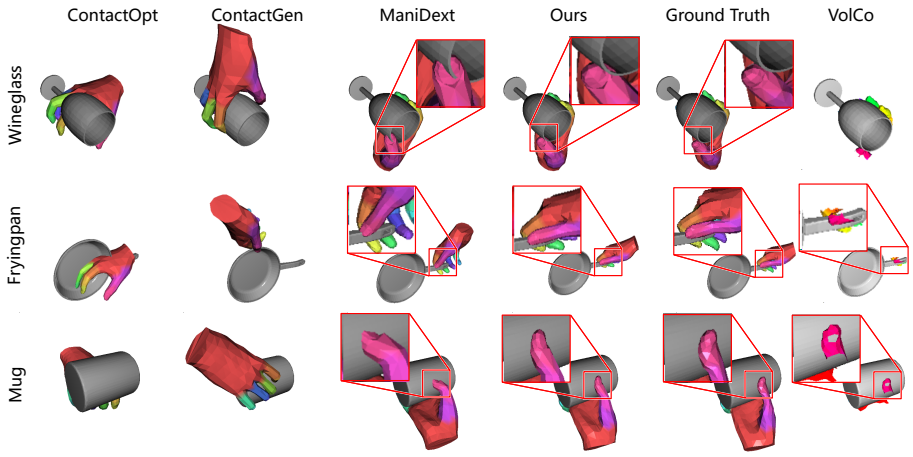


Figure 4: Reconstruction from previous contact maps vs. VolCo. The rightmost column represents VolCo using the reconstructed local vertices. VolCo demonstrates the best capability to restore details.

CPU, 64GB RAM, and two 4090 GPUs. Please refer to Appendices B and C.1 for model architectures and data pre-processing details.

4.1 Reconstruction from VolCo

We test the correspondence and contact representations together with comparisons to previous representations on GRAB test set every 64 frames. Three criteria are utilized following previous research [48, 13, 24]: 1) *End Point Error (EPE)*—average Euclidean distance of mesh vertices; 2) *F-score*—harmonic average of precision and recall between vertices of two meshes under certain threshold(5mm and 15mm); 3) *Area Under the ROC Curve (AUC)* [9] expanding from 0 to 5cm. All hands are aligned to the objects. The resulting poses are obtained via optimization, initialized by the best 6D poses that fit the average hand to the contact. See Appendix C.3 for initialization details.

In Tab. 1, our method achieves the best performance across all metrics, indicating superior information preservation. The improvement from part labels to CSE confirms that point-wise correspondence is critical for reconstruction. The transition from map to volume representation reduces EPE by 40.1% at the same point count, highlighting the importance of sampling along the surface normal direction.

Fig. 4 illustrates the reason for improved precision. First, VolCo itself can recover local hand details without MANO as shown in the last column. This makes the fitted model precise at local details, restoring the slight penetrations; Second, as shown in 'Fryingpan' example, such contact encodes the exact spatial positions of more points with higher confidence around the object surface, which makes it possible to recover a tight grip around a slim bar—where the contact area is very small.

Table 2: Quantitative results of grasp synthesis on GRAB Dataset. '↑' after a criterion means the higher the better, while '↓' means the opposite. Best result in each column is marked in **bold** and the second best is underlined. Our method achieves the best stability, smallest PD and largest CA.

Method	SD (cm)↓	PD (cm)↓	IV (cm^3)↓	CA (cm^2) ↑	CR ↑	Entropy ↑	Cluster Size ↑
GrabNet [34]	$1.02_{\pm 2.07}$	$0.40_{\pm 0.34}$	$7.43_{\pm 9.89}$	$25.5_{\pm 13.5}$	**1.00**	**2.79**	2.57
ContactGen [24]	$2.32_{\pm 3.47}$	$0.52_{\pm 0.49}$	$3.32_{\pm 3.46}$	$19.9_{\pm 13.3}$	0.99	2.76	**4.28**
FastGrasp [38]	$2.55_{\pm 3.71}$	$0.53_{\pm 0.66}$	$\mathbf{2.02}_{\pm 2.19}$	$13.3_{\pm 7.42}$	0.86	2.77	0.41
FAGrasp [6]	$0.61_{\pm 1.37}$	$0.73_{\pm 0.67}$	$4.74_{\pm 4.71}$	$30.6_{\pm 11.1}$	**1.00**	2.65	3.84
Ours	$\mathbf{0.52}^{\downarrow 14.8\%}_{\pm 0.64}$	$\mathbf{0.27}^{\downarrow 32.5\%}_{\pm 0.28}$	$3.78_{\pm 4.05}$	$\mathbf{31.1}^{\uparrow 1.6\%}_{\pm 12.9}$	**1.00**	2.77	3.61

Table 3: Quantitative results of grasp synthesis on HO3D Dataset. '↑' after a criterion means the higher the better, while '↓' means the opposite. Best result in each column is marked in bold. Our method shows strong adaptivity to out-of-domain objects.

Method	SD (cm)↓	PD (cm)↓	IV (cm^3)↓	CA (cm^2)↑	CR ↑	Entropy ↑	Cluster Size ↑
GrabNet [34]	$1.61_{\pm 2.84}$	$1.04_{\pm 0.78}$	$8.30_{\pm 6.69}$	$31.4_{\pm 17.7}$	**1.00**	**2.87**	3.18
GraspTTA [17]	$4.17_{\pm 4.21}$	$1.12_{\pm 0.92}$	$\mathbf{3.40}_{\pm 3.38}$	$16.6_{\pm 15.5}$	0.72	2.76	0.22
ContactGen [24]	$3.41_{\pm 4.31}$	$0.98_{\pm 0.78}$	$4.01_{\pm 2.93}$	$19.9_{\pm 12.3}$	0.95	2.77	**4.56**
FastGrasp [38]	$2.49_{\pm 3.35}$	$1.18_{\pm 0.88}$	$9.54_{\pm 19.7}$	$18.1_{\pm 15.2}$	**1.00**	2.76	0.34
FAGrasp [6]	$1.21_{\pm 1.93}$	$1.93_{\pm 1.33}$	$45.4_{\pm 93.6}$	$38.4_{\pm 21.8}$	**1.00**	2.76	4.17
Ours	$\mathbf{1.06}^{\downarrow 12.3\%}_{\pm 1.68}$	$\mathbf{0.70}^{\downarrow 28.6\%}_{\pm 0.64}$	$10.9_{\pm 27.3}$	$\mathbf{45.4}^{\uparrow 18.2\%}_{\pm 29.2}$	**1.00**	2.85	3.85

4.2 Grasp Generation Analysis

Metrics. Following previous research [17, 24, 38, 6], we utilize multiple criteria to evaluate stability, geometric plausibility and diversity of the generation. 1) *Simulation Displacement (SD)* measures the displacement of object center of mass in PyBullet [8] simulation; 2) *Penetration Depth (PD)* measures the average of all samples' deepest penetration depth; 3) *Intersection Volume (IV)* is calculated by voxelizing the object meshes to $1\ mm^3$ and counting the overlapping voxels. 3) *Contact Ratio (CR)* reflects the ratio of samples in contact. To indicate the tightness of grasps, we introduce *Contact Area (CA)* to measure the valid contact area on the object surface (See Appendix. C.4). 4) *Entropy* and *Cluster Size* measure the diversity of generation.

Quantitative Analysis. The results on both datasets are reported in Tab. 2 and 3 along with 1-sigma error bars. On the in-domain GRAB dataset, our method achieves the lowest SD and PD with the largest CA, exceeding the previous state-of-the-art by 14.8%, 32.5%, and 1.6%, respectively. This demonstrates that the generated grasps are stable, plausible, and contact-rich — whereas previous methods often collapse to similar poses in exchange for less penetration. On the out-of-domain HO3D dataset, our method surpasses the previous state-of-the-art by 12.3% (SD), 28.5% (PD), 18.2% (CA), demonstrating strong generalization to unseen objects. Despite HO3D objects being generally larger than those in GRAB, our method retains strong adaptability to these scales.

Our intersection volumes are not the least among all methods, which is desired as a result of tighter grasps. Larger CA indicate that our method generate grasps with more valid contact. Since slight penetrations are allowed, larger CA is usually accompanied with larger IV. Over 28% improvement in PD on both datasets indicates a larger portion of penetrations are plausible. The preference for tighter grasps is also shown in Qualitative analysis in Fig. 5.

Qualitative Analysis. We compare our method qualitatively with previous state-of-the-art methods in penetrations [24, 38] and stability [6] in Fig. 5. The figure highlights the contact details, where previous methods tend to generate loose grasps due to aggressive penetration losses or implausibly deep penetrations especially in thin objects like wineglass and mug. This is due to the non-differentiable penetration loss with NN query. Our method, on contrast, finds a good balance of grasping steadiness

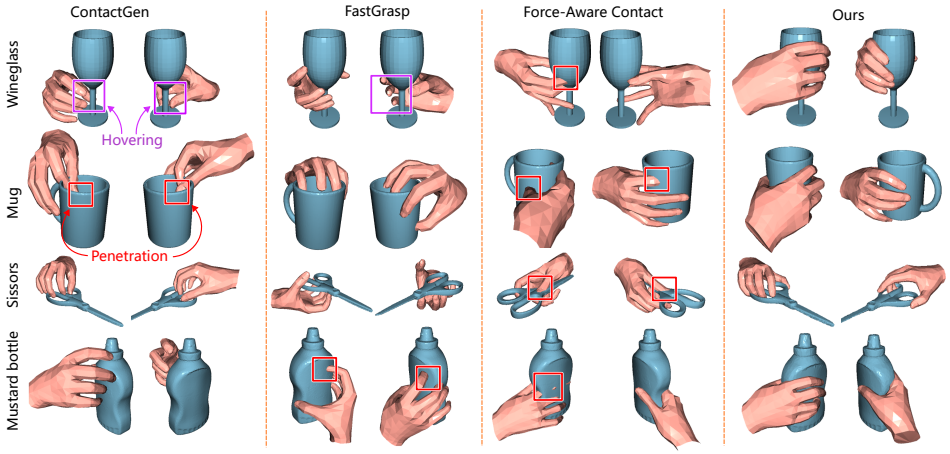


Figure 5: Qualitative comparison with previous methods [24, 38, 6]. Hovering fingers are marked with purple boxes while penetrations are marked in red. While previous methods tend to produce excessively loose or penetrating grasps, our method achieves a physically plausible balance.

Table 4: Ablation study on key design choices. The best are marked in **bold** and the second best are underlined. VolCo brings obvious improvement, and the stability loss is more effective in reducing SD with geometric plausibility ensured by prior guidance and initialization.

No.	VolCo	Prior Guidance	Stability Loss	Initia-lization	SD (cm) ↓	PD (cm) ↓	IV (cm^3) ↓	CR ↑	CA ↑
1	✗	✗	✗	✗	1.72	0.42	4.64	0.92	24.7
2	✓	✗	✗	✗	1.28	0.42	2.34	**1.00**	19.5
3	✓	✗	✓	✗	1.24	0.39	**2.11**	**1.00**	19.1
4	✓	✓	✗	✗	1.63	0.23	4.16	**1.00**	26.5
5	✓	✓	✗	✓	0.78	0.26	3.61	**1.00**	28.5
6	✓	✓	✓	✗	1.10	**0.21**	3.53	**1.00**	30.8
7	✓	✓	✓	✓	**0.52**	0.27	3.78	**1.00**	**31.1**

and penetration—penetrations are common but mostly not severe, just like in the original motion capture data, indicating higher geometric plausibility. This also explains the lower penetration depth and higher penetration volumes, as our model prefers tight grasps with larger contact areas. More generated grasps are available in Appendix D.

4.3 Ablation Study

This section studies the contribution of VolCo and the key design choices, i.e.,hand prior guidance and explicit force-aware stability loss. 'Prior Guidance' indicates if the auxiliary branch is used during training, while 'Initialization' means if the pose is initialized with the predicted pose. The first experiment use point-based contact maps following [23].

Point-based contact yields weak stability due to incomplete contact details and hand-crafted rules. Switching to VolCo ($1 \rightarrow 2$) obviously improves stability. Adding prior guidance ($2 \rightarrow 4; 3 \rightarrow 5$) greatly decreases penetration depth, and increases IV and CA, while adding initial guess ($4 \rightarrow 5; 6 \rightarrow 7$) further improves stability and CA, highlighting the dual role of the auxiliary branch: it regularizes generated VolCo during training to allow being fitted by MANO with less penetrations, and provides better initialization during testing to fit larger contact areas. Interestingly, the stability loss reduce displacement obviously only when there is prior guidance ($4 \rightarrow 6; 5 \rightarrow 7$), suggesting geometric plausibility is a prerequisite for physical stability.

Table 5: Per sample runtime breakdown and performance, tested on GRAB dataset. We provide a spars version (VolCoDiff-S) to study the influence of volume resolutions. †GraspTTA is not trained on GRAB, so the SD and PD results are not strictly comparable.

	Generator Framework	Inference time (s)	Optim. Iterations	Optim. time (s)	Total time (s)	SD (cm)	PD (cm)
GrabNet [34]	VAE	**0.14**	-	-	**0.14**	1.02	0.40
GraspTTA† [17]	VAE	0.17	1000	6.83	7.00	3.35	0.64
ContactGen [24]	VAE	0.34	1200	43.6	44.0	2.32	0.52
FastGrasp [38]	Diffusion	8.60	-	-	8.60	2.55	0.53
FAGrasp [6]	VAE	0.415	1200	39.1	39.5	0.61	0.73
Point-Contact Diff [23]	Diffusion	2.90	1000	**4.69**	7.59	1.72	0.42
VolCoDiff-S (64×4^3)	Diffusion	5.48	1000	4.89	10.4	0.67	0.32
VolCoDiff (128×8^3)	Diffusion	5.28	1000	5.39	10.7	**0.52**	**0.27**

4.4 Runtime Analysis

In this section, we analyze the trade-off between efficiency and performance against prior methods. Table. 5 reports a detailed runtime breakdown, with all experiments run on a single RTX 4090 GPU. Our method achieves the best simulation displacement and penetration depth among all compared methods. VAE-based methods need only a single forward pass at inference, while diffusion-based methods require iterative denoising, so the latter are slower. Among diffusion-based methods, ours has the second-fastest inference time. Its optimization time is also the second fastest overall, behind only the point-contact baseline [23]. Additionally, reducing the resolution from 128×8^3 to 64×4^3 has little effect on efficiency, suggesting the preference for the normal resolution.

However, the point-contact baseline shows obvious drawback in batched inference. Although our method has more parameters (74.04M vs. 24.41M) and higher peak memory (358.1M vs. 191.4M), it achieves substantially lower TFLOPs (2.01 vs. 13.2). The major source of computation is the UNet attention blocks which scales as $O(N)$ in token count, not the volumetric representation due VAE compression. Thus our inference time barely changes as the number of samples increases from 1 to 10 (5.28s→5.47s), but the time of baseline increases substantially (2.90s→19.57s).

5 Conclusion

We propose VolCo, a novel volumetric contact representation that faithfully preserves fine-grained details in hand-object contact. Building on VolCo, we present VolCoDiff, a volume-wise latent diffusion model, in which the local details are handled by VolumeVAE and the global composition of hand configurations is mitigated by diffusion. The experiments demonstrate the highest precision in reconstruction from contact and the state-of-the-art geometric plausibility and stability in grasp generation, emphasizing the value of introducing hierarchical contact volumes.

Limitations. Our method is currently limited in controllability and scale adaptivity. It prefers only tighter grasps, and the fixed number and scale of volumes limit the coverage of object scales. Future work will explore ways to control VolumeVAE latents and apply adaptive volume numbers and scales.

Broader Impact. Our method can help with improving AR/VR experiences and dexterous robotic deployment, but risks fostering the misuse of robots in harmful contexts.

6 Acknowledgements

This research was funded by the China Scholarship Council - University of Birmingham PhD Scholarship programme (No. 202406230091) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2020-II201789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Samuel R Bowman et al. "Generating sentences from a continuous space". In: *20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*. Association for Computational Linguistics (ACL). 2016, pp. 10–21.

[2] Andrew Brock et al. "Generative and Discriminative Voxel Modeling with Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Aug. 2016. DOI: 10.48550/arXiv.1608.04236. arXiv: 1608.04236 [cs].

[3] Junuk Cha et al. "Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 1577–1585. DOI: 10.48550/arXiv.2404.00562.

[4] Yu-Wei Chao et al. "DexYCB: A Benchmark for Capturing Hand Grasping of Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9044–9053. DOI: 10.48550/arXiv.2104.04631.

[5] Zerui Chen et al. "gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 12890–12900.

[6] Zhuo Chen et al. "Force-Aware 3D Contact Modeling for Stable Grasp Generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Nov. 2025. DOI: 10.48550/arXiv.2511.13247. eprint: 2511.13247 (cs).

[7] Özgün Çiçek et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, Oct. 2016, pp. 424–432. ISBN: 978-3-319-46722-1. DOI: 10.1007/978-3-319-46723-8_49.

[8] Erwin Coumans and Yunfei Bai. *PyBullet, a Python module for physics simulation for games, robotics and machine learning*. http://pybullet.org. 2016–2021.

[9] Tom Fawcett. "An Introduction to ROC Analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2005.10.010.

[10] Patrick Grady et al. "ContactOpt: Optimizing Contact to Improve Grasps". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, June 2021, pp. 1471–1481. ISBN: 978-1-6654-4509-2. DOI: 10.1109/CVPR46437.2021.00152.

[11] Patrick Grady et al. "PressureVision++: Estimating Fingertip Pressure from Diverse RGB Images". In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 8683–8693. DOI: 10.1109/WACV57701.2024.00850.

[12] Lin Guo, Zongxing Lu, and Ligang Yao. "Human-machine interaction sensing technology based on hand gesture recognition: A review". In: *IEEE Transactions on Human-Machine Systems* (2021), pp. 300–309.

[13] Shreyas Hampali et al. "HOnnotate: A Method for 3D Annotation of Hand and Object Poses". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3196–3206. DOI: 10.48550/arXiv.1907.01481.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. DOI: 10.48550/arXiv.2006.11239.

[15] Siyuan Huang et al. "Diffusion-based Generation, Optimization, and Planning in 3D Scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[16] K. H. Hunt and F. R. E. Crossley. "Coefficient of restitution interpreted as damping in vibroimpact". In: *Journal of Applied Mechanics* 42.2 (1975), pp. 440–445.

[17] Hanwen Jiang et al. "Hand-Object Contact Consistency Reasoning for Human Grasps Generation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11107–11116. DOI: 10.48550/arXiv.2104.03304.

[18] Korrawe Karunratanakul et al. "A Skeleton-Driven Neural Occupancy Representation for Articulated Hands". In: *Proceedings of the International Conference on 3D Vision*. IEEE, Sept. 2021. DOI: 10.48550/arXiv.2109.11399. arXiv: 2109.11399.

[19] Korrawe Karunratanakul et al. “Grasping Field: Learning Implicit Representations for Human Grasps”. In: *2020 International Conference on 3D Vision (3DV)*. Nov. 2020, pp. 333–344. DOI: 10.1109/3DV50981.2020.00043.

[20] Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. 2014. DOI: 10.48550/arXiv.1312.6114.

[21] Muchen Li et al. “LatentHOI: On the Generalizable Hand Object Motion Generation with Latent Hand Diffusion.” In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 17416–17425. DOI: 10.1109/CVPR52734.2025.01623.

[22] Samuel Li et al. “ShapeGrasp: Zero-Shot Task-Oriented Grasping with Large Language Models through Geometric Decomposition”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2024), pp. 10527–10534.

[23] An-Lun Liu, Yu-Wei Chao, and Yi-Ting Chen. “Task-Oriented Human Grasp Synthesis via Context- and Task-Aware Diffusers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 10375–10385.

[24] Shaowei Liu et al. “ContactGen: Generative Contact Modeling for Grasp Generation”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 20552–20563. ISBN: 979-8-3503-0718-4. DOI: 10.1109/ICCV51070.2023.01884.

[25] Zhijian Liu et al. “Point-Voxel CNN for Efficient 3D Deep Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[26] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[27] Théo Morales, Omid Taheri, and Gerard Lacey. “A Versatile and Differentiable Hand-Object Interaction Representation”. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE/CVF, Nov. 2024. DOI: 10.1109/WACV61041.2025.00013. arXiv: 2409.16855 [cs].

[28] Natalia Neverova et al. *Continuous Surface Embeddings*. Nov. 2020. DOI: 10.48550/arXiv.2011.12438. arXiv: 2011.12438 [cs].

[29] Charles R Qi et al. “Pointnet: Deep learning on point sets for 3d classification and segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[30] Charles R Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *arXiv preprint arXiv:1706.02413* (2017).

[31] Robin Rombach et al. “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[32] Javier Romero, Dimitrios Tzionas, and Michael J. Black. “Embodied Hands: Modeling and Capturing Hands and Bodies Together”. In: *ACM Transactions on Graphics* 36.6 (Dec. 2017), pp. 1–17. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/3130800.3130883. arXiv: 2201.02610 [cs].

[33] Soshi Shimada et al. “MACS: Mass Conditioned 3D Hand and Object Motion Synthesis”. In: *2024 International Conference on 3D Vision (3DV)*. Mar. 2024, pp. 1082–1091. DOI: 10.1109/3DV62453.2024.00082.

[34] Omid Taheri et al. “GRAB: A Dataset of Whole-Body Human Grasping of Objects”. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 581–600. ISBN: 978-3-030-58548-8. DOI: 10.1007/978-3-030-58548-8_34.

[35] Guy Tevet et al. “Human Motion Diffusion Model”. In: *The Eleventh International Conference on Learning Representations*. ICLR, Oct. 2022. DOI: 10.48550/arXiv.2209.14916. arXiv: 2209.14916 [cs].

[36] Tze Ho Elden Tse et al. “S^2Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning”. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Vol. 13661. Cham: Springer Nature Switzerland, 2022, pp. 568–584. ISBN: 978-3-031-19768-0. DOI: 10.1007/978-3-031-19769-7_33.

[37] Ashish Vaswani et al. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[38] Xiaofei Wu et al. “FastGrasp: Efficient Grasp Synthesis with Diffusion”. In: *International Conference on 3D Vision (3DV)*. Nov. 2024. DOI: 10.48550/arXiv.2411.14786. arXiv: 2411.14786 [cs].

[39] Guo-Hao Xu et al. “Dexterous Grasp Transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 17933–17942. DOI: 10.48550/arXiv.2404.18135.

[40] Lixin Yang et al. “CPF: Learning a Contact Potential Field To Model the Hand-Object Interaction”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11097–11106.

[41] Ruihan Yang et al. *EgoVLA: Learning Vision-Language-Action Models from Egocentric Human Videos*. 2025. arXiv: 2507.12440 [cs.RO]. URL: https://arxiv.org/abs/2507.12440.

[42] Lior Yariv et al. “Mosaic-SDF for 3D Generative Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 4630–4639.

[43] Yufei Ye et al. “G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 1911–1920. DOI: 10.48550/arXiv.2404.12383.

[44] Jiajun Zhang et al. “ManiDext: Hand-Object Manipulation Synthesis via Continuous Correspondence Embeddings and Residual-Guided Diffusion”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025), pp. 1–15. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2025.3588302.

[45] Zhongqun Zhang et al. “NL2Contact: Natural Language Guided 3D Hand-Object Contact Modeling with Diffusion Model”. In: *Proceedings of the 2024 European Conference on Computer Vision*. ECVA, July 2024. DOI: 10.48550/arXiv.2407.12727. arXiv: 2407.12727 [cs].

[46] Juntian Zheng et al. “CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 585–594. DOI: 10.48550/arXiv.2303.15469.

[47] Keyang Zhou et al. “TOCH: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement”. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Vol. 13663. Cham: Springer Nature Switzerland, 2022, pp. 1–19. ISBN: 978-3-031-20061-8. DOI: 10.1007/978-3-031-20062-5_1.

[48] Christian Zimmermann et al. “FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 813–822.

[49] Binghui Zuo et al. “GraspDiff: Grasping Generation for Hand-Object Interaction With Multimodal Guided Diffusion”. In: *IEEE Transactions on Visualization and Computer Graphics* (2024), pp. 1–13. ISSN: 1941-0506. DOI: 10.1109/TVCG.2024.3466190.

A Stability Loss With Slack Variables

In this section, we briefly restate the stability loss in the literature [6], and how it gets improved by slack variables.

Suppose a grasp with an object of mass m, inertia matrix I, and surface friction ratio μ in a scene with gravitational acceleration $\mathbf{g}$. For a contact point $\mathbf{x}$, the normal force applied is denoted by $\mathbf{F}_\mathbf{x} = F_\mathbf{x}\mathbf{n}_\mathbf{x} \in \mathbb{R}^3$, where $\mathbf{n}_\mathbf{x}$ is the normal direction. With the object SDF, the direction is formally defined as $\mathbf{n}_\mathbf{x} = -\dot{\text{SDF}}_o(\mathbf{x})$. Suppose $\mathbf{b}_\mathbf{x}, \mathbf{t}_\mathbf{x}, \mathbf{n}_\mathbf{x}$ form a regular basis and $\bar{\mathbf{b}}_\mathbf{x} = (\mathbf{x} - \mathbf{p}_{CoM}) \times \mathbf{b}_\mathbf{x}$ etc.,, then the rotational and angular accelerations are calculated by

$$\mathbf{a} = \mathbf{g} + \frac{1}{m}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} F_{\mathbf{x}}\left(\mathbf{n}_{\mathbf{x}} + \mu\delta_{\mathbf{x}}\mathbf{b}_{\mathbf{x}} + \mu\gamma_{\mathbf{x}}\mathbf{t}_{\mathbf{x}}\right); \tag{9}$$

$$\boldsymbol{\alpha} = I^{-1}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} F_{\mathbf{x}}\left(\bar{\mathbf{n}}_{\mathbf{x}} + \mu\delta_{\mathbf{x}}\bar{\mathbf{b}}_{\mathbf{x}} + \mu\gamma_{\mathbf{x}}\bar{\mathbf{t}}_{\mathbf{x}}\right), \tag{10}$$

where $\delta_{\mathbf{x}}, \gamma_{\mathbf{x}} \in [-1, 1]$ are variables to determine friction.

The stability loss is ideally

$$E_{stability} = \|\mathbf{a}\|^2 + \|\boldsymbol{\alpha}\|^2. \tag{11}$$

Following the paper [6], to formulate Eq. (11) to a differentiable loss term, $\delta_{\mathbf{x}}, \gamma_{\mathbf{x}}$ should be eliminated. This is achieved by comparing the possible upper and lower bounds of each dimension of the acceleration. Use the subscript $j = 1, 2, 3$ to represent the jth element in the vector, then we have

$$\begin{aligned}\mathbf{g}_j + \frac{1}{m}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} F_{\mathbf{x}}\left(\mathbf{n}_{\mathbf{x}j} - \mu|\mathbf{b}_{\mathbf{x}j}| - \mu|\mathbf{t}_{\mathbf{x}j}|\right) &\triangleq L_j \le \mathbf{a}_j \\ &\le U_j \triangleq \mathbf{g}_j + \frac{1}{m}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} F_{\mathbf{x}}\left(\mathbf{n}_{\mathbf{x}j} + \mu|\mathbf{b}_{\mathbf{x}j}| + \mu|\mathbf{t}_{\mathbf{x}j}|\right). \end{aligned} \tag{12}$$

$$\begin{aligned}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} F_{\mathbf{x}}\left(I_{j\cdot}^{-1}\bar{\mathbf{n}}_{\mathbf{x}} - \mu|I_{j\cdot}^{-1}\bar{\mathbf{b}}_{\mathbf{x}}| - \mu|I_{j\cdot}^{-1}\bar{\mathbf{t}}_{\mathbf{x}}|\right) &\triangleq \mathcal{L}_j \le \boldsymbol{\alpha}_j \\ &\le \mathcal{U}_j \triangleq \sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} F_{\mathbf{x}}\left(I_{j\cdot}^{-1}\bar{\mathbf{n}}_{\mathbf{x}} + \mu|I_{j\cdot}^{-1}\bar{\mathbf{b}}_{\mathbf{x}}| + \mu|I_{j\cdot}^{-1}\bar{\mathbf{t}}_{\mathbf{x}}|\right). \end{aligned} \tag{13}$$

So the stability loss is designed to punish only the cases when 0 fall out of the feasible range.

$$\mathcal{L}_{stability} = \sum_{j=1}^{3}\left(\max\{L_j, 0\} + \max\{\mathcal{L}_j, 0\} - \min\{U_j, 0\} - \min\{\mathcal{U}_j, 0\}\right) \tag{14}$$

This formulation, however, cannot be directly applied in our setting. The key concern is that this formula assumes the forces are precise, yet the spring-damper force model $F_{\mathbf{x}} = kd$ is an approximation and random errors in penetration depth in motion capture data are unavoidable. As stated in the main paper, our solution is to add a slack variable to loosen the constraint on penetration depth, so any depth within $\pm\Delta d$ error is considered correct. In other words, the upper and lower bounds in Eq. (12) and (13) can be expanded by replacing the constant $F_{\mathbf{x}}$ with a function of $\Delta d_{\mathbf{x}}$, e.g.,,

$$L_j(\Delta d) = \mathbf{g}_j + \frac{1}{m}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} k(d_{\mathbf{x}} + \Delta d_{\mathbf{x}})\left(\mathbf{n}_{\mathbf{x}j} - \mu|\mathbf{b}_{\mathbf{x}j}| - \mu|\mathbf{t}_{\mathbf{x}j}|\right) \tag{15}$$

$$= L_j + \frac{1}{m}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} k\Delta d_{\mathbf{x}}\left(\mathbf{n}_{\mathbf{x}j} - \mu|\mathbf{b}_{\mathbf{x}j}| - \mu|\mathbf{t}_{\mathbf{x}j}|\right) \tag{16}$$

$$\ge L_j - \frac{1}{m}\sum_{i=1}^{N}\sum_{\mathbf{x}\in\mathcal{G}_i} k\Delta d\left|\mathbf{n}_{\mathbf{x}j} - \mu|\mathbf{b}_{\mathbf{x}j}| - \mu|\mathbf{t}_{\mathbf{x}j}|\right| \triangleq L'_j. \tag{17}$$

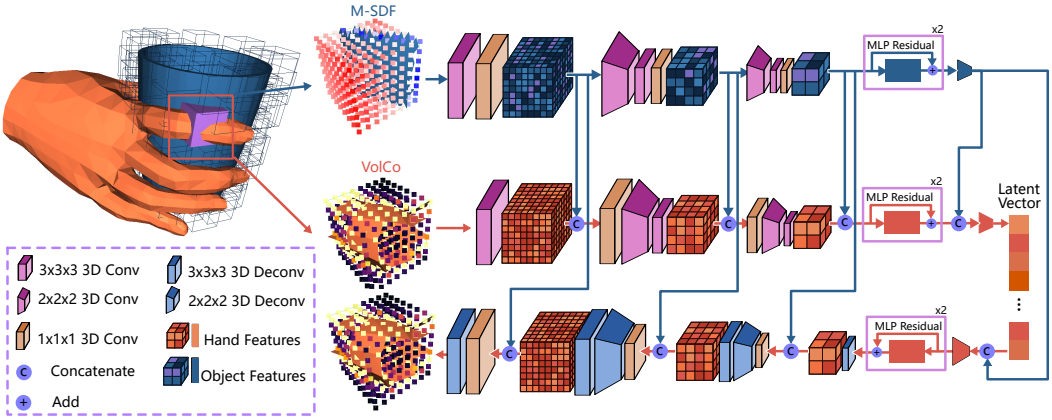


Figure 6: VolumeVAE architecture

The same works for $U'_j, \mathcal{L}'_j, \mathcal{U}'_j$.

Substitute $L_j, \mathcal{L}_j, U_j, \mathcal{U}_j$ with $L'_j, \mathcal{L}'_j, U'_j, \mathcal{U}'_j$ in Eq. (14), then we get a new stability loss which can be applied to our case.

We set the limit of the slack variables to $\Delta d = 1\text{mm}$.

B Model Architectures

This section provide details regarding the architectures of VolumeVAE, VolCoDiff UNet, and hand VAE, in support for reproducing the results.

B.1 VolumeVAE

VolumeVAE adopts a symmetric 3D VAE architecture, illustrated in Fig. 6. The input VolCo and object SDF volumes are passed through two downsampling stages, flattened, and processed by residual MLP blocks followed by a projection layer to yield a 1D feature vector. Multi-scale object features are injected into the contact features at corresponding scales via channel concatenations. Due to the small input dimensionality ($8^3 \times n_{channel}$), the input is designed to minimize information loss: downsampling layers are implemented by 3D convolutional layers with a kernel size of 2, and all activations are LeakyReLU with a negative slope of 0.2.

B.2 VolCoDiff

The UNet of VolCoDiff is shown in Fig. 7. The object SDF volumes are flattened to a $8^3 = 512$-dimensional vector, which are used as feature for each object volume center and processed by a PointNet [29] encoder. The hand branch receives the information from the object and the contact by cross-attention, while the contact branch does not receive information from the hand branch.

B.3 Hand VAE

The hand VAE consists of a MLP encoder that receives the flattened hand vertex positions as input and outputs a 32-dimensional latent vector, where the first 16 dimensions represent the mean and the remaining 16 the log-standard-deviation. The latent gets reparameterized to a sampled 16-dimensional latent and decoded by a 2-layer MLP decoder to a 61-dimensional vector $(\mathbf{t} \mid \boldsymbol{\theta} \mid \boldsymbol{\beta})$ representing the MANO parameter of the hand.

The training of hand VAE is done on GRAB training set. The loss terms include a reconstruction loss $\mathcal{L}_{rec}$ and KL-divergence loss. The reconstruction loss is calculated based on the hand vertices that are predicted from MANO $\{\mathbf{v}_i\}_{i=1}^{778} = \text{MANO}(\mathbf{t} \mid \boldsymbol{\theta} \mid \boldsymbol{\beta})$, i.e.,

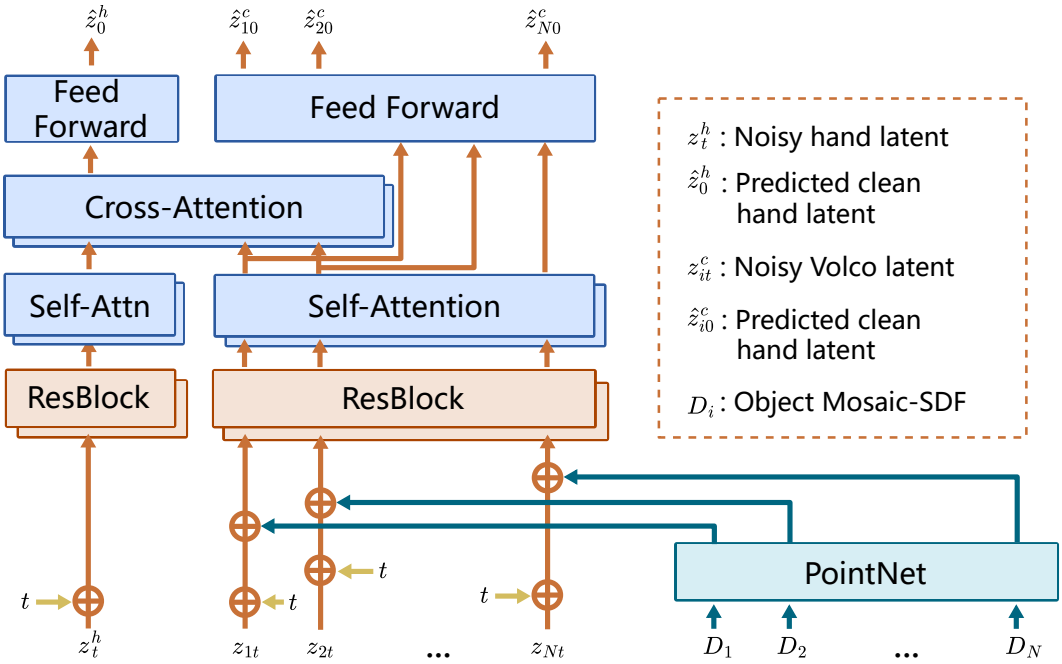


Figure 7: The UNet architecture of VolCoDiff.

$$\mathcal{L}_{rec} = \frac{1}{778}\sum_{i=1}^{778} MSE(\mathbf{v}_i, \hat{\mathbf{v}}_i) \tag{18}$$

With the KL-divergence loss $\mathcal{L}_{KL} = D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)||\mathcal{N}(0, I))$, the loss to train the hand VAE is simply

$$\mathcal{L}_{handvae} = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} \tag{19}$$

where $\lambda_{KL} = 10^{-6}$.

The model is trained with AdamW optimizer [26] and 10^{-4} learning rate for 100 epochs.

C Experimental Details

This section provides details to reproduce our result in the main paper. Our experiments are done on a linux machine with Intel(R) Xeon(R) w7-3445 CPU, 64GB RAM, and two 4090 GPUs.

C.1 Data Pre-processing

For training, we adopt the GRAB dataset [34] with the same pre-processing pipeline as GrabNet, yielding 323,622 frames of closely interacting hand-object pairs. The validation and test sets are obtained by downsampling every 4 and 64 frames, comprising 7,757 and 1,012 frames, respectively. The training set is used directly for training the diffusion model, and the test set is used in the experiment of reconstruction from contact. Further, the training set is augmented by a randomly selected rotation from all 24 axis-aligned ones which is applied to the local volumes. This keeps the rotated volumes still axis-aligned so no interpolations are needed. The resulting Mosaic-SDF and VolCo can be efficiently obtained by reordering the grid points.

However, the training of VolumeVAE requires special attention. The key concern is that one training sample of VolumeVAE should be one single volume with object SDF and contact values. So we construct a local volume dataset based on the sampled points. Each object surface is sampled 1,024 points, so totally 331.4M local volumes are available for training. The dataset exceeds the capacity of

a compact VAE model, and a substantial portion of the data carries redundant or repeated information. We therefore apply two-level downsampling during data pre-processing.

To mitigate the similar poses between continuous frames, we downsample the training set every 5 frames, and validation and test set every 10 frames; To reduce the redundancy of nearby sampled points, we also downsample the object points each frame by 40% (training set) and 20% (validation and testing). Since in our VolCoDiff, non-contact volumes are set to 0, so fitting to those volumes in training is not necessary. Thus, we only take the sampled volumes with at least one hand vertex inside to construct the dataset. Finally, we arrive at a training set of 7,320,315 volumes, a validation set of 165,651 and a test set of 273,315.

C.2 Contact-Map Based Method

The contact-map based method that we utilize in ablation study follows the architecture of [23], which differs from VolCoDiff in three points: 1. No hand branch; 2. The object encoder is PointNet++ [30], and the input are the point positions and normal vectors; 3. The input contact for each point is the vector consisting of the contact and CSE $(c \mid \mathbf{e})$ instead of the latent vector. This model is trained by the standard diffusion loss $\mathcal{L}_{diff}$, building up by the contact value loss $\mathcal{L}_c$ and the cse value loss weighted by contact like in VolumeVAE $\mathcal{L}_{cse}$:

$$\mathcal{L}_{diff} = \mathcal{L}_c + \lambda_{cse}\mathcal{L}_{cse}. \tag{20}$$

where $\lambda_{cse} = 0.05$ for more focus on the contact.

C.3 Initialization for Reconstruction From Contact

The methods in Tab. 1 obtain the final result by refining an initial pose. This initialization is provided based on the correspondence representations. Note that all initial poses are the average pose, the difference is only in the way to obtain the global 6D pose.

If the contact representation does not encode correspondence [10]), then the initial 6D pose is simply set to 0.

If the correspondence is represented by hand part [24], the initialization can be regarded as aligning the part centers on the rigid hand to those calculated from the contact map. The part centers of the hand are represented by anchors from [40], with picking one anchor to stand for each part. The part centers from the contact is the average of all point positions weighted by the contact likelihood. Denote the hand part centers as $\{\mathbf{p}_i\}_{i=1}^{16}$ and the object part centers as $\{\mathbf{q}_i\}_{i=1}^{16}$, then the initial orientation R_0 and translation $\mathbf{t}_0$ are calculated by point cloud registration.

Further, if the correspondence is represented by CSE, then we can calculate the corresponding point on the hand surface for each point in VolCo, by weighted average of relative hand vertices using weights in Eq. (3). Then R_0 and t_0 are also calculated by point registration using such point pairs.

C.4 Contact Area Metric Definition

Contact Area (CA) is proposed to measure the surface area of the object mesh that is in contact with the hand. Formally, an object vertex v_i is considered in contact if its distance to the nearest point on the hand mesh is below a threshold δ and the object surface normal at v_i opposes the hand surface normal at the corresponding closest point, i.e.,

$$\mathcal{C} = \left\{ v_i \in \mathcal{V}_{\text{obj}} \mid d(v_i, \mathcal{M}_{\text{hand}}) < \delta \ \wedge \ \mathbf{n}_i^{\text{obj}} \cdot \mathbf{n}_i^{\text{hand}} < 0 \right\}, \tag{21}$$

where $\mathcal{V}_{\text{obj}}$ denotes the object vertices, $d(\cdot,\cdot)$ is the closest-point distance to the hand mesh $\mathcal{M}_{\text{hand}}$, and $\mathbf{n}_i^{\text{obj}}$, $\mathbf{n}_i^{\text{hand}}$ are the surface normals at v_i and its closest hand point, respectively. The contact area is then computed as the summed area of all object faces that contain at least one contact vertex:

$$\text{CA} = \sum_{f \in \mathcal{F}_{\text{obj}}} \mathbf{1}[f \cap \mathcal{C} \neq \emptyset] \cdot A(f), \tag{22}$$

where $\mathcal{F}_{\text{obj}}$ is the set of object faces and $A(f)$ denotes the area of face f. A larger CA indicates more extensive hand-object contact, which is generally desirable for stable grasps.

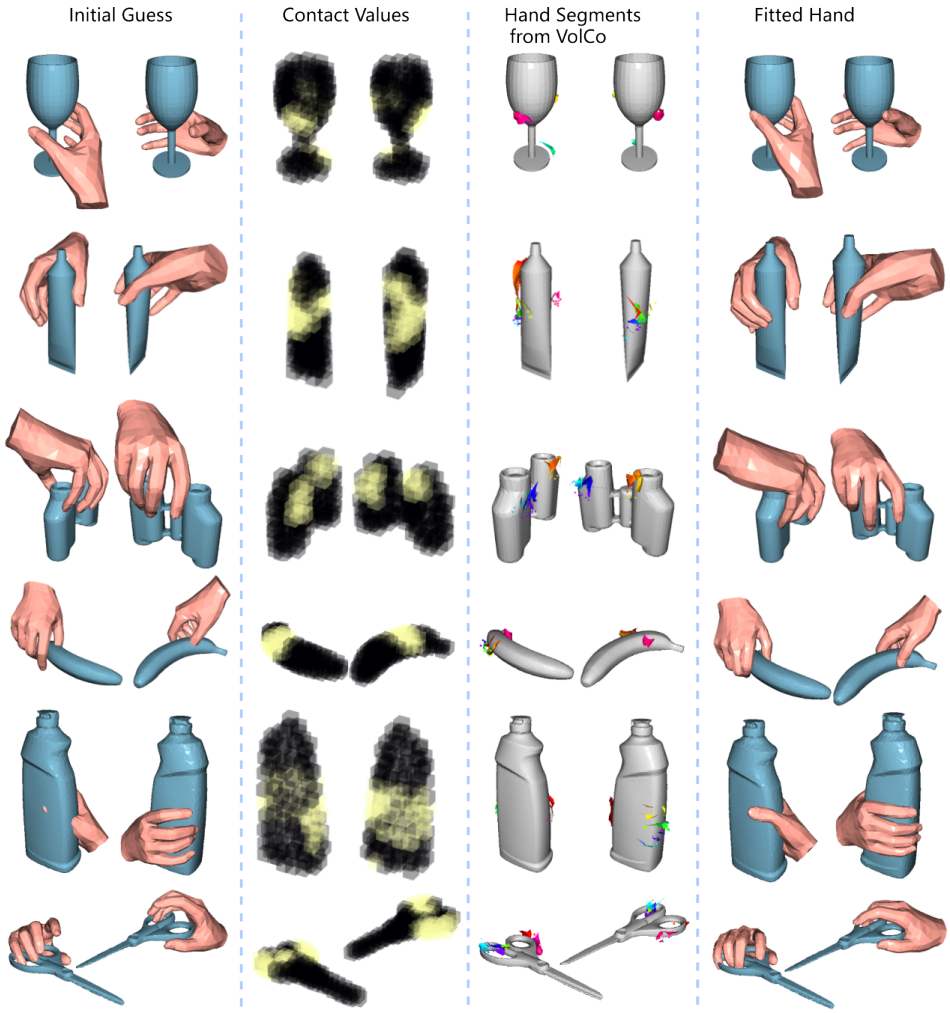


Figure 8: The initial guess, predicted VolCo, and the final grasps of several objects, with each sample shown by two viewpoints.

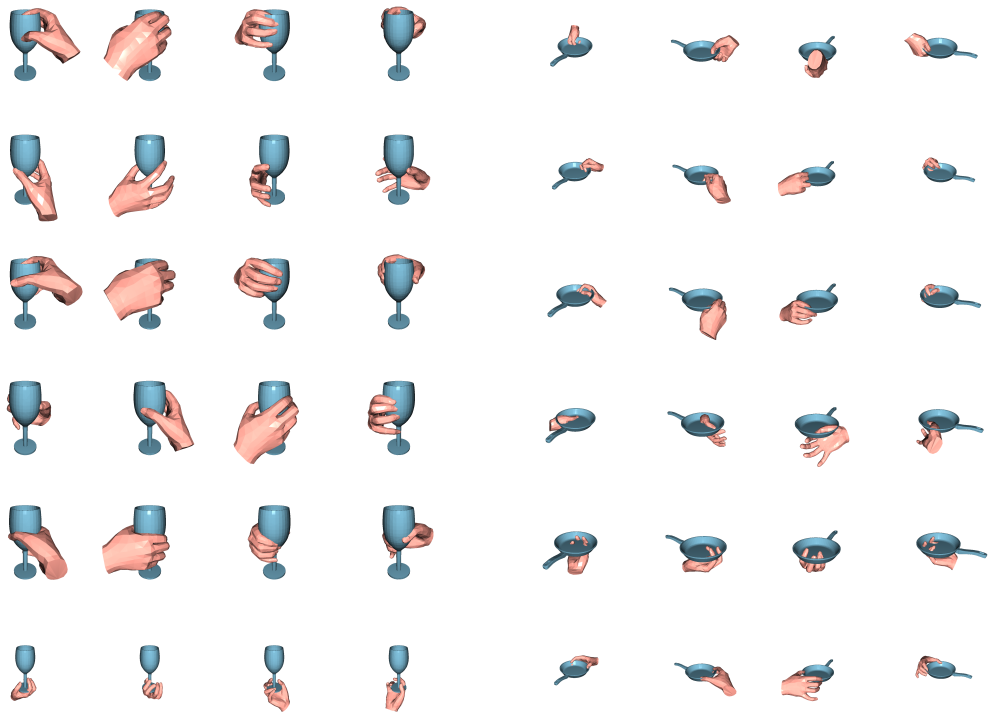
D Generated Grasp Visualizations

D.1 Initial Guess and Refinement

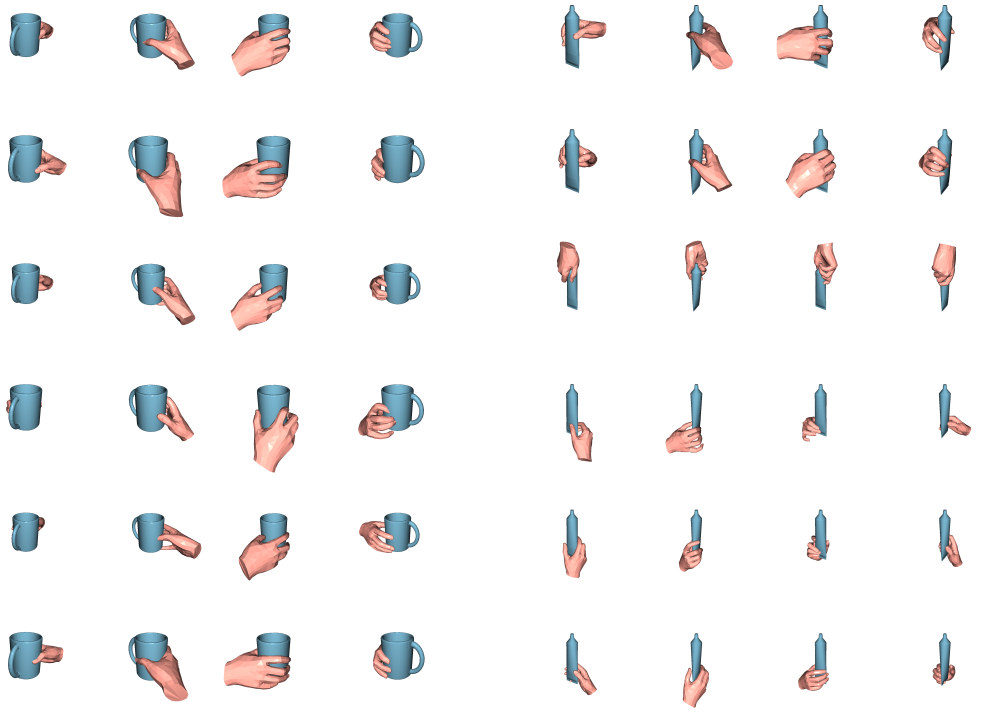
To illustrate the process of pose refinement, we show some samples including the initial guess form the hand branch, the contact values shown in volumes, the segments reconstructed from VolCo, and the final fitted hand in Fig. 8. We observe that the initial guess is usually closely related to the predicted VolCo, with only some slight displacement that results in artifacts. With the contact and repulsive losses defined in Eq. (9), the pose fitting stage efficiently converges to the optimal hand pose consistent with VolCo, for two reasons. First, The dominant loss term $\mathcal{L}_{rec}$ is essentially quadratic w.r.t. the hand vertex locations, yielding a well-behaved optimization landscape without the complexity introduced by contact calculations [10]; Second, the initial guess provides a strong initialization in the vicinity of the optimal pose, facilitating the convergence to the global minimum.

D.2 Generated Grasps

In support to the conclusions in the experiments, this section shows more generated samples. These samples are randomly selected for the highest representative. Fig. 9 shows GRAB samples Fig. 10 shows HO3D samples, with each row showing the 4 viewpoints of a sample. The figures demonstrates the diversity and high probability of plausible generation.

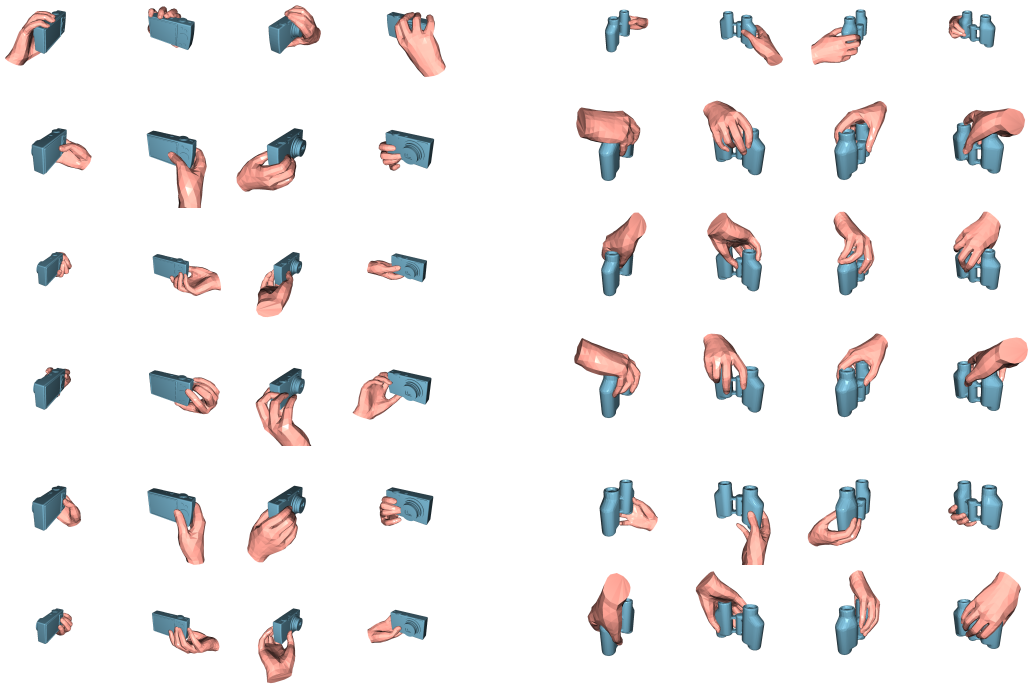


(a) Wineglass (b) Frying Pan



(c) Mug (d) Toothpaste

Figure 9: Qualitative results on GRAB dataset (2/3)



(e) Camera (f) Binoculars

Figure 9: Qualitative results on GRAB dataset (3/3).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution of this paper is a novel contact representation based on local volumes and a grasp generation framework based on it. The new method can generate tighter grasps with much less severe penetrations. This is clearly introduced in abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

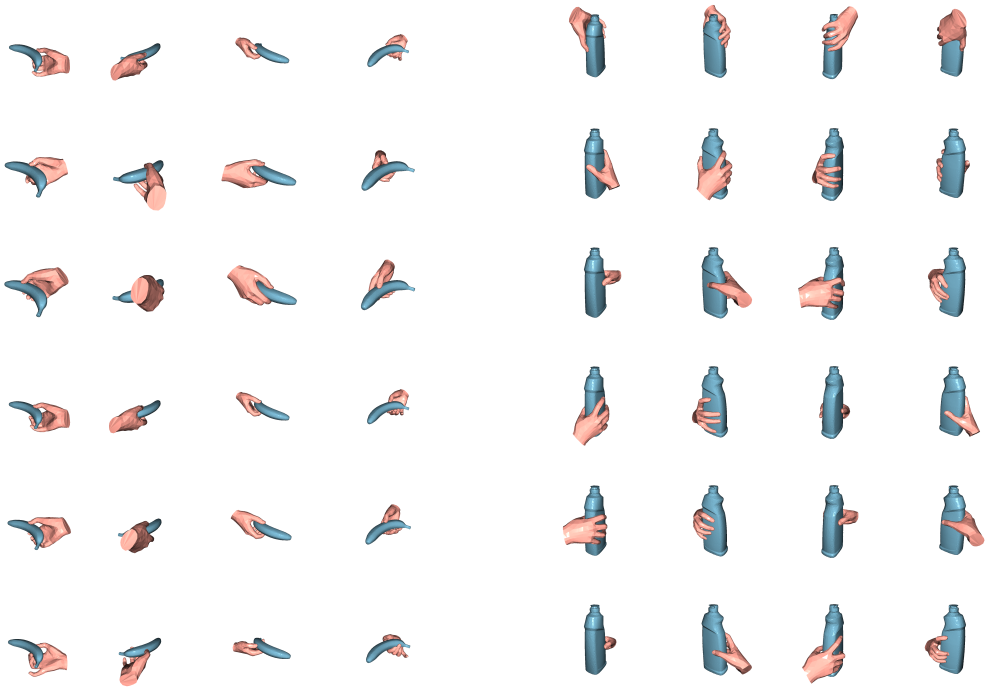
2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are introduced in the last section.

   Guidelines:

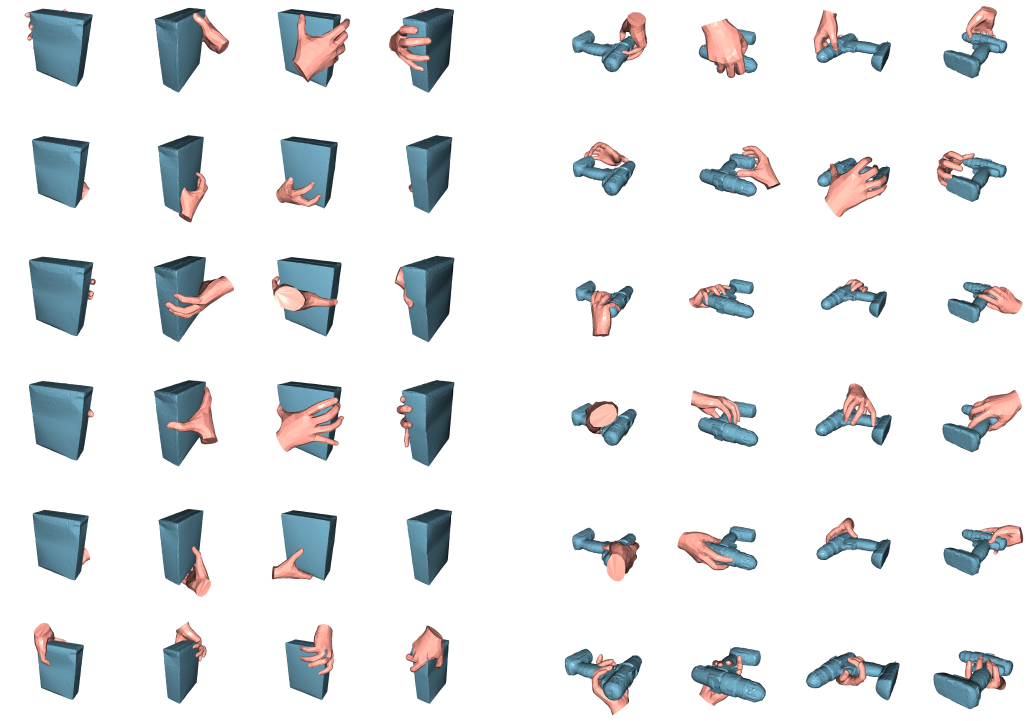


(a) Banana (b) Bleach Cleanser

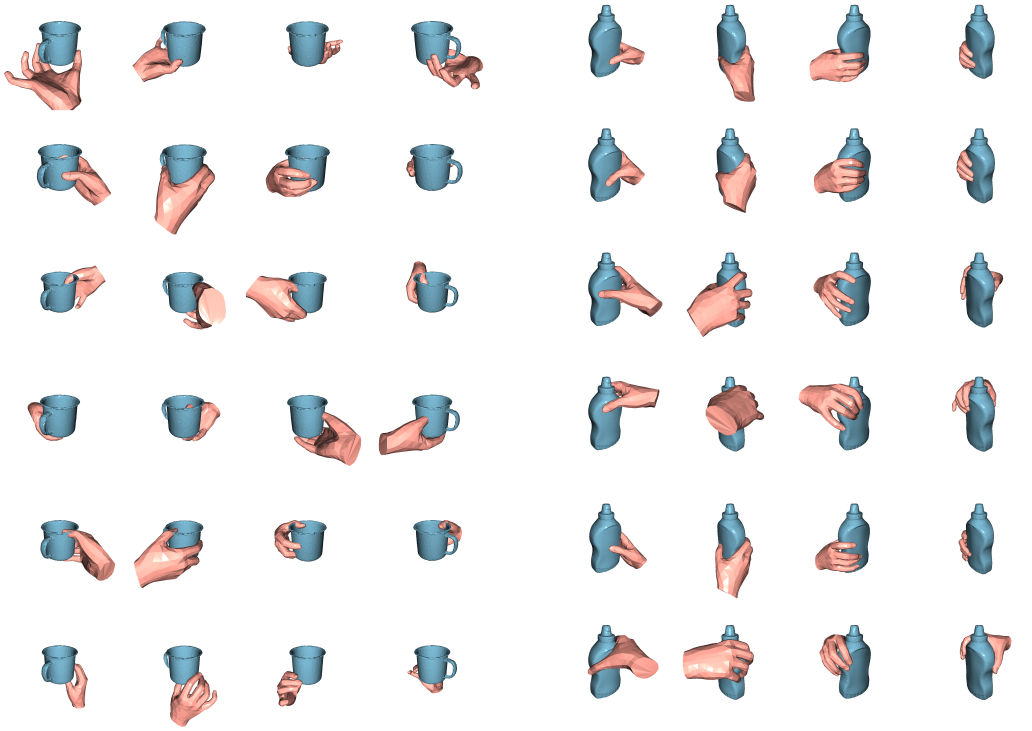
Figure 10: Qualitative results on out-of-domain HO3D dataset (1/5).

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

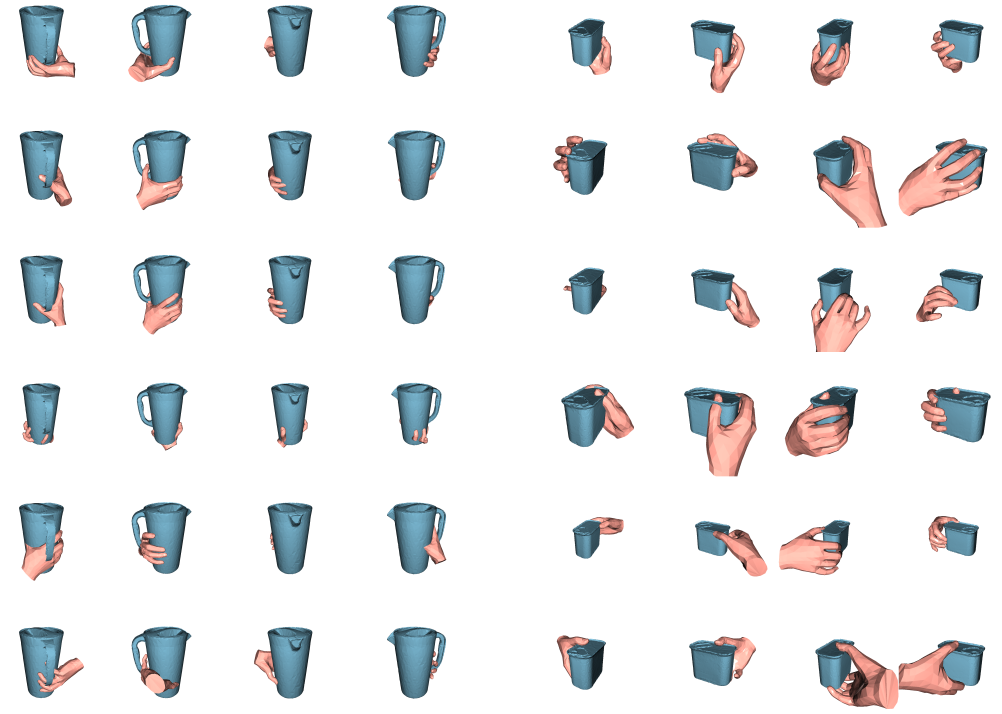


(c) Cracker Box (d) Power Drill

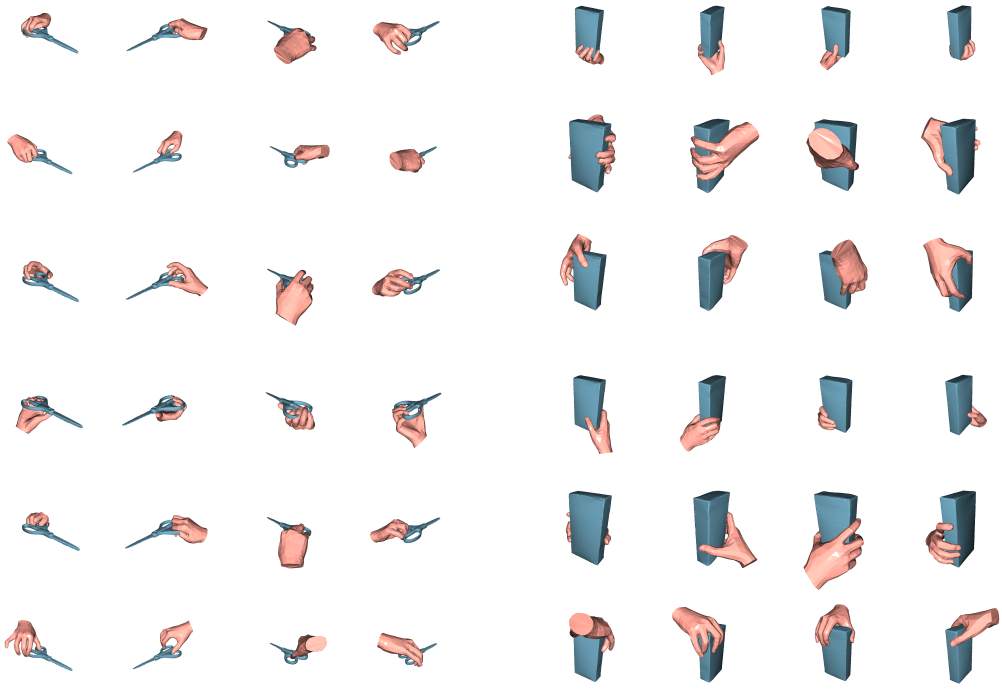


(e) Mug (f) Mustard Bottle

Figure 10: Qualitative results on out-of-domain HO3D dataset (3/5).



(g) Pitcher Base (h) Potted Meat Can



(i) Scissors (j) Sugar Box

Figure 10: Qualitative results on out-of-domain HO3D dataset (5/5).

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All hyperparameters we use in the training and testing are given in the paper. The model architecture is briefly introduced in the paper and described in detail in Appendix. B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is available at `https://github.com/chzh9311/volco`

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The hyperparameters are given in Methodology section, with brief motivations on choosing these parameters. Mostly, the hyperparameters are selected following previous works.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results reported in the main tables Tab. 2 and 3 are accompanied with error bars indicating the standard deviation (1-sigma).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The detailed computational sources are described in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper utilizes only the public datasets for experiments, which are open-access for research purposes. These are no self-collected human data in our experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The social impacts are discussed in the last section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper proposes to generate the human grasp, which aims at studying the general interactive patterns and can hardly be used to harm public safety. Though it can foster the deployment of dexterous robots that might result in harmful context, this is more related with robotic breakthroughs and are out of scope of this paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The public datasets are properly cited and we use them following their terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not contribute new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: There is no user study in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not include crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The main framework of the method is based on latent diffusion model, not LLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.