GSHOI Denoiser: Denoising Gaussian Hand-Object Interaction for Photorealistic Rendering

Lizhi Zhao
Beihang University

Xuequan Lu
The University of Western Australia

Bin Hu
Beihang University

Wei Ke
Macao Polytechnic University

Lili Wang *
Beihang University

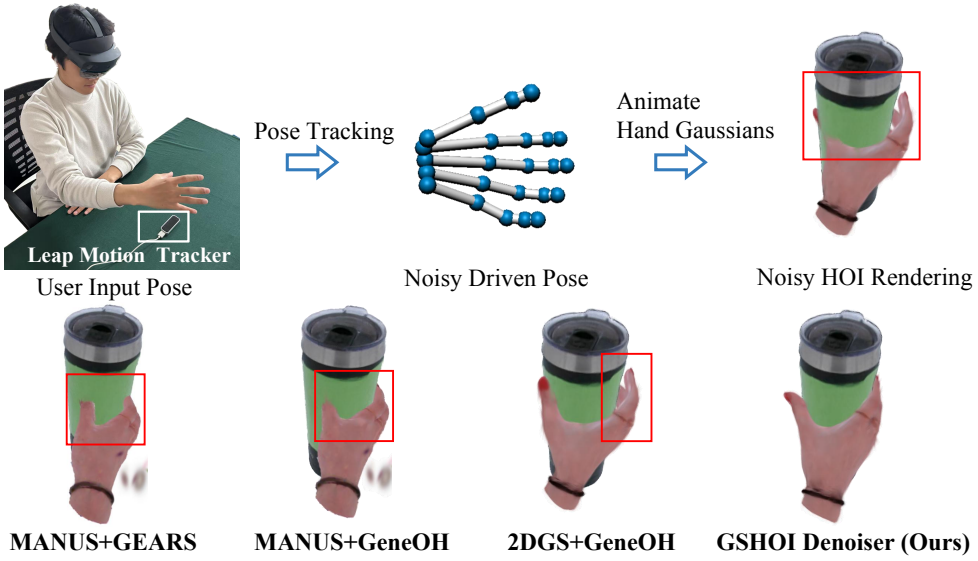


Figure 1: Top: Given a user's mid-air hand pose, the Leap Motion tracker provides a coarsely estimated, noisy driven pose, which animates the pretrained hand Gaussians to produce an hand-object interaction (HOI) rendering with penetration and unstable-grasp artifacts. Bottom: We compare our method with HOI rendering and denoising methods, including MANUS+GEARS [15, 29], MANUS+GeneOH [11], and 2DGS+GeneOH [5]. The renderings of MANUS+GEARS and MANUS+GeneOH contain severe penetrations. 2DGS+GeneOH involves a gap between the finger and object and thus lacks stability. Our GSHOI Denoiser effectively removes input noise and accurately poses fingers on the object surface, enabling photorealistic HOI rendering.

Abstract

Many VR/AR applications require the photorealistic rendering of hand–object interactions. Virtual hands are driven by users' hand poses captured via motion tracking to interact with virtual objects. The driven pose can be very noisy due to the constraints of tracking hardware and computation accuracy. This noise may lead to distorted hand poses and penetration artifacts during rendering. In this paper, we introduce the Gaussian Hand-Object Interaction Denoiser, the Gaussian splatting-based hand-object interaction denoising method, which effectively denoises the input twisted and penetrated hand poses to produce photorealistic results. We first propose the innovative joint-to-Gaussian surface representation, which accurately models the spatial relationships between hand skeleton joints and object Gaussians while highlighting hand–object penetrations and generalizing well to new hand poses and objects. Then, we propose a geometry-aware de-penetration algorithm that eliminates penetrations by detecting intersections between skeleton bones and object Gaussians and reposing any penetrated fingers onto the estimated underlying surface of the object. Experiments demonstrate that our method not only effectively reduces hand–object penetration depth but also produces more realistic rendering quality compared to the state-of-the-art methods MANUS+GEARS, MANUS+GeneOH, and 2DGS+GeneOH. The user study results show that our method significantly improves the users' visual perceptual experience regarding penetration and stability metrics. Project page: *https://github.com/ZhaoLizz/GSHOIDenoiser*

Index Terms: Virtual Reality, Gaussian Splatting, Hand-Object Interaction

1 Introduction

Hand-object interaction is common in daily life and plays a crucial role in AR/VR applications. Typically, virtual hands are driven by users' real hand poses, captured through 3D motion tracking technologies to interact with the virtual objects. However, due to the accuracy limitations of tracking hardware and computation, motion capture results often contain noise, leading to distorted hand poses and undesired HOI penetrations [11].

Researchers have proposed denoising techniques to address erroneous motion tracking results and enhance HOI plausibility. Liu et al. introduced GeneOH [11], which takes a sequence of noisy hand skeletons and object mesh as inputs, then models HOI relation as the distance between joints and object vertices, and outputs

*Corresponding author.

Lizhi Zhao, Bin Hu, and Lili Wang are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, 100191, Beijing, China. E-mails: {lizhizhao, pencil, wanglily}@buaa.edu.cn

X. Lu is with the Department of Computer Science and Software Engineering, The University of Western Australia, Crawley WA 6009, Australia. E-mail: bruce.lu@uwa.edu.au.

Wei Ke is with the Faculty of Applied Sciences, Macao Polytechnic University. E-mail: wke@mpu.edu.mo.

denoised hand skeletons via a diffusion model. GeneOH achieves promising denoising results, but it lacks the capability to reconstruct and render photorealistic HOI appearances. Pokhariya et al. proposed MANUS [15], a Gaussian splatting-based HOI rendering method that can reconstruct rigid objects and animatable hands and render photorealistic HOI images. Nevertheless, MANUS cannot perceive the relationships between hand and object Gaussians, and therefore, cannot denoise erroneous pose inputs. Instead, it relies on accurate hand poses captured from a high-end camera array, which are not accessible to typical AR/VR applications. It is non-trivial to apply GeneOH on MANUS properly as Gaussian representation differs greatly from the mesh data. The mesh vertices lie accurately on the surface, while Gaussians are coarsely distributed around the underlying surface. As a result, hand denoising for Gaussian-based HOI presents special challenges, in particular: 1) How to design a representation to model the accurate spatial relations between the hand skeleton joints and the coarsely distributed object Gaussians? 2) How to estimate and further eliminate penetrations between Gaussian-based hands and objects?

In this paper, we propose the Gaussian Hand-Object Interaction Denoiser (GSHOI Denoiser), a Gaussian splatting-based HOI denoising framework, that addresses the aforementioned challenges. Our method is capable of denoising the input twisted hand poses and HOI penetrations, generating photorealistic HOI renderings. Our method involves two key innovative components. Firstly, we propose the Joint-to-Gaussian Surface (J2GS) representation, which can model the accurate spatial relations between hand skeleton joints and object Gaussians, highlight HOI penetrations, and generalize well to new hand poses and new object Gaussians, by splatting the coarsely distributed object Gaussians onto skeleton joints as smooth local surfaces. Secondly, we propose the Geometry-aware De-penetration (GDP) algorithm to eliminate the HOI penetrations, by estimating the intersected geometries of skeleton bones and object Gaussians, and reposing any penetrated fingers on the underlying surface of object Gaussians.

We compare our GSHOI Denoiser with state-of-the-art (SOTA) HOI rendering and denoising methods MANUS+GEARS, MANUS+GeneOH, and 2DGS+GeneOH on MANUS-Grasp dataset. The experimental results demonstrate that our method not only reduces hand-object penetration depth by 57.3%, but also produces more realistic rendering quality, improving 8.3% in Peak Signal-to-Noise Ratio (PSNR), reducing 17.4% in Learned Perceptual Image Patch Similarity (LPIPS). Fig. 1 shows the comparison of renderings between our method and SOTA methods, showcasing notable improvement in penetration and grasp stability.

To summarize, our technical contributions are as follows:

- We introduce GSHOI Denoiser, the first 3D Gaussian HOI rendering denoising framework that denoises the input twisted hand poses and HOI penetrations, and produces photorealistic renderings of hand object interaction.
- We present a novel HOI representation of joint-to-Gaussian surface, which captures the accurate HOI spatial relationships by splatting the coarsely distributed object Gaussians onto skeleton joints as smooth local surfaces.
- We propose a novel geometry-aware de-penetration method to eliminate the HOI penetrations, by estimating the intersected geometries between skeleton bones and object Gaussians, and reposing penetrated fingers onto the underlying surface of object Gaussians.

2 Related Work

In this section, we introduce recent hand object interaction reconstruction and denoising methods related to our approach.

2.1 Hand Object Interaction Reconstruction

Hand-Object Interaction reconstruction aims to estimate the pose and/or reconstruct the geometries of objects and articulated hands from input images [23, 4, 10, 12]. Yang et al. proposed the Contact Potential Field (CPF) [21], which models the HOI relation using a spring-mass system to estimate the poses of pre-scanned objects and MANO hands [18] from images. Additionally, Yang et al. introduced ArtiBoost [20], an online HOI synthesis method designed to enhance the diversity of the HOI dataset. Recent approaches focus on reconstructing HOI meshes without relying on pre-scanned object and hand templates. Fan et al. presented HOLD [2], a category-agnostic method that reconstructs the geometry of both unknown hands and objects through a compositional articulated implicit model. Ye et al. proposed a diffusion network to recover the neural 3D representation of object shapes and the time-varying motion of hand articulation for HOI reconstruction [23]. Furthermore, Ye et al. introduced G-HOP [22], a diffusion-based generative prior for HOI reconstruction, which represents the hand using a skeleton distance field and aligns it with the learned object signed distance field (SDF). In addition to 3D-aware approaches, Ye et al. introduced Affordance Diffusion [24], a large-scale 2D diffusion model that generates HOI images with diverse backgrounds based on coarse hand-grasping layout conditions. However, Affordance Diffusion lacks precise pose control over synthesized hand poses. HO-NeRF [16] models the hand and object appearance with neural radiance fields [13]. These methods reconstruct high-fidelity HOI from captured grasp images. However, their reliance on such image inputs restricts applicability in scenarios in which users' physical hands interact with virtual objects. In this paper, we focus on generating plausible HOI without the captured grasp image.

Gaussian splatting exhibits a new 3D representation for reconstruction and rendering [8, 5, 1, 26, 1]. Zhao et al. proposed GaussianHand [26], which employs 3D Gaussian splatting [8] to reconstruct an animatable hand avatar with realistic appearance rendering capabilities. Pokhariya et al. introduced MANUS [15], which pioneers HOI appearance reconstruction by representing the hand using articulated Gaussians and the object with static Gaussians. Simply concatenating the hand and object Gaussians enables the rendering of both HOI and contact areas.

Previous Gaussian-based hand object reconstruction and rendering methods assume input accurate driven poses, while can not denoise erroneous poses. In this paper, we propose the GSHOI Denoiser, achieving both high-quality HOI rendering and denoising.

2.2 Hand Object Interaction Denoising

HOI pose estimation methods often involve erroneous interaction noise [14, 25, 27, 31], including hand twisting and penetration, etc. HOI Denoising aims to understand the 3D scene and remove the noise to produce a perceptually realistic sequence [32, 30, 6, 7]. Zhou et al. proposed TOCH [28], a spatio-temporal representation modeling the ray-casting relation between hands and objects for refining incorrect mesh-based HOI sequences. Zhou et al. further proposed GEARS [29] to denoise the HOI sequence by attaching bounding boxes on hand joints to query the neighborhood object surface vertices and extracting the object's local geometry features with a spatio-temporal network. Liu et al. proposed GeneOH [11], a contact-centric HOI denoising method that learns the manifold of Euclidean distance and trajectory consistency representation of HOI sequence with a diffusion model, which can denoise the input noisy sequences by first diffusing them to a whitened noise space and then cleaning the HOI sequence via the trained denoiser. Both methods focus on mesh geometry only, neglecting HOI appearance modeling. Previous HOI denoising methods work well for hands and objects meshes but do not consider realistic rendering. Adapting these methods for Gaussian-based HOI rendering is non-trivial because mesh vertices lie precisely *on* thin surfaces, whereas

Gaussians are coarsely distributed around an underlying surface. In this paper, we aim HOI denoising for Gaussian splatting rendering.

3 METHOD

3.1 The Pipeline of GSHOI Denoiser

We formulate the Gaussian HOI rendering denoising problem as follows. Given the pretrained Gaussians of a rigid object, and a sequence of *twisted penetrated* driven hand poses in object's canonical coordinate frame, along with pretrained skeleton articulated hand Gaussians, we aim to denoise the hand pose sequence without accessing ground truth (GT) images to achieve clean hand poses, and render photorealistic HOI images from arbitrary view as:

$$\begin{aligned} \hat{\theta}_f &= Denoiser\left(\mathcal{G}_o, \mathcal{G}_h(\theta_f)\right), \\ I_f &= Splat\left(cat(\mathcal{G}_o, \mathcal{G}_h(\hat{\theta}_f)), v\right), \end{aligned} \tag{1}$$

where $\mathcal{G}_o$ is the rigid object Gaussians, $\mathcal{G}_h(\cdot)$ is articulated hand Gaussians, θ_f is the f-frame noisy input pose parameters, $Denoiser$ is a HOI denoising function, $\hat{\theta}_f$ denotes the f-frame cleaned pose parameters. v is an arbitrary viewpoint, and $cat(\cdot,\cdot)$ denotes the concatenation operator, $Splat$ denotes the Gaussian splatting rendering function.

Fig 2 shows the pipeline of our GSHOI Denoiser, which includes 4 steps: HOI Gaussian initialization, constructing the joint-to-Gaussian surface representation to capture the accurate HOI spatial relationships (Section 3.2), hand pose denoising with the diffusion model, and using the geometry-aware de-penetration method to estimate and eliminate HOI penetrations (Section 3.3).

First, we initialize the hand and object Gaussians. We use the 2D Gaussians [5] to represent objects and hands. The rigid object Gaussians can be easily trained from multi-view object images with camera viewpoint annotations. For hand Gaussians, we first initialize the Gaussians on the registered canonical MANO [18] surface and pre-compute the skinning weights of MANO as a spatial weight volume, which stores the interpolated skinning weights of arbitrary query points. Given an annotated hand pose, we use the rigid kinematic chain to calculate bone transformations and then animate the canonical hand Gaussians to the posed space by:

$$\begin{aligned} \mathcal{B}(\theta) &= \{B_i\}_{i=1\ldots n_b}, \\ x_p^i &= \left(\sum_{k=1}^{n_b} \mathcal{W}(x_c^i)_k B_k\right) x_c^i, \end{aligned} \tag{2}$$

where θ is the driven hand pose, $\mathcal{B}(\cdot)$ denotes n_b transformations and $B_i \in SE(3)$ denotes transforming the i-th joint from the canonical frame to the posed frame. $\mathcal{W}(\cdot)$ is the skinning weights volume, $x_c^i \in \mathbb{R}^3$ denotes position property of the i-th Gaussian of $\mathcal{G}_h$, and x_p^i denotes the position of the i-th Gaussian in the posed space. $\mathcal{W}(x_c^i)_k \in \mathbb{R}$ denotes the skinning weights at position x_c^i of the k-th joint. The posed hand Gaussians are then rendered and trained with GT images.

Second, given pretrained hand and object Gaussians and noisy hand poses, we construct our proposed joint-to-Gaussian surface representation to depict HOI spatial relations.

Third, we apply a diffusion model [11] to achieve denoised hand skeleton joints from the noisy joint-to-Gaussian surface representation. Specifically, in the training phase, the diffusion model first diffuses the clean HOI representation to a whitened noisy space by gradually adding noise to it. Then the diffusion model learns a denoiser network to predict the added noise to project it back to the clean space. The denoiser network is trained with a regression loss to minimize the difference between the added and predicted noise.

In the inference phase, we input our HOI representation of the noisy input poses and pass it through the denoiser network. The network eliminates the noise step-by-step to output the denoised hand skeleton joints sequence, which acts as the supervision to optimize the noisy hand poses. We regress skeleton joints from the hand Gaussians in a simple yet effective manner. We assign each canonical MANO vertex a region ID [21]. Then, each canonical hand Gaussian x_c is assigned a region ID based on the nearest MANO vertex, dividing the hand Gaussians into 17 anatomical parts. Since the parts are organized hierarchically, we select 10% of each part's Gaussians that are closest to the part's parent as adjacencies, and compute their centers as the corresponding skeleton joint. Finger tip joints are defined as the centers of the farthest adjacencies. For the carpal and thumb metacarpal joints, we use the center of the full part. The palm metacarpal joint is discarded to maintain consistency with the 16 joints of the MANO model. We construct a regression matrix to conduct the canonical joints regression process. With the calculated skeleton joints of the posed hand Gaussians, we optimize the noisy hand poses to achieve denoised hand poses as:

$$\begin{aligned} J_f &= \mathcal{E} \cdot \mathcal{G}_h(\theta_f).pos, \\ \theta_f^d &= \arg\min_{\theta_f} MSE\left(\bar{J}_f, J_f\right), \end{aligned} \tag{3}$$

where θ_f is the f-frame noisy hand pose, $\mathcal{E} \in \mathbb{R}^{n_h \times 16}$ denotes the joints regression matrix and n_h is the number of hand Gaussians. pos denotes the position property of Gaussians, J_f is regressed joints of posed hand Gaussians. MSE is the mean square error loss, $\bar{J}_f$ is the f-frame denoised hand skeleton joints from the diffusion model, and θ_f^d is the optimized denoised hand pose.

Fourth, we propose the Geometry-aware De-penetration method to further eliminate the HOI penetrations and render the hand object Gaussians as photorealistic HOI images.

3.2 Joint-to-Gaussian Surface Representation

We present joint-to-Gaussian surface representation to model the accurate spatial relations between hand skeleton joints and the discrete coarsely distributed object Gaussians, while highlighting the HOI spatial penetrations.

Our key insight behind achieving the above goals is to splat the discrete object Gaussians onto hand skeleton joints as sample points on underlying smooth local surfaces and construct the relative spatial vectors between skeleton joints and the sample points. We formulate our J2GS representation $\mathcal{R}$ as:

$$\mathcal{R}(\mathcal{J}, \mathcal{G}_o, \mathcal{A}) = \mathcal{J} - unproject(Splat^d(\mathcal{G}_o, \mathcal{V}), \mathcal{V}), \tag{4}$$

where $\mathcal{V}$ denotes the set of our proposed joint cameras, $Splat^d$ denotes the Gaussian depth rendering function, $unproject(\cdot,\cdot)$ denotes the screen-to-world transformation depending on cameras $\mathcal{V}$, and $\mathcal{A}$ denotes a Gausaian subset named anchors. We will detail the design of our J2GS below.

Given hand skeleton joints sequence and object Gaussians, we begin constructing our J2GS representation by randomly selecting some object Gaussians as anchors from the subset of Gaussians that are close to the hand skeleton trajectory within a threshold as:

$$\mathcal{A} = select(\|\mathcal{G}_o - \cup(\mathcal{J})\|_2 < t_c, \mathcal{G}_o).shuffle()[:n_a], \tag{5}$$

where $\mathcal{J}$ denote the joints sequence, n_a is the number of anchors, $\mathcal{A} = \{\mathbf{a}_i\}_{i=1\ldots n_a}$ denote the anchors, $t_c = 10\,mm$ is the distance threshold. $select(\cdot,\cdot)$ denotes the conditional selection operator, with the condition and choice list as inputs, $\|\cdot\|_2$ denotes the Euclidean distance, $\cup(\mathcal{J})$ denotes the union of hand skeletons across f-frames as the temporal trajectory, and $shuffle()[:n_a]$ randomly select n_a Gaussians.

Then, we propose the joint cameras to capture the local underlying surfaces around the anchors. We attach the n_a cameras with a

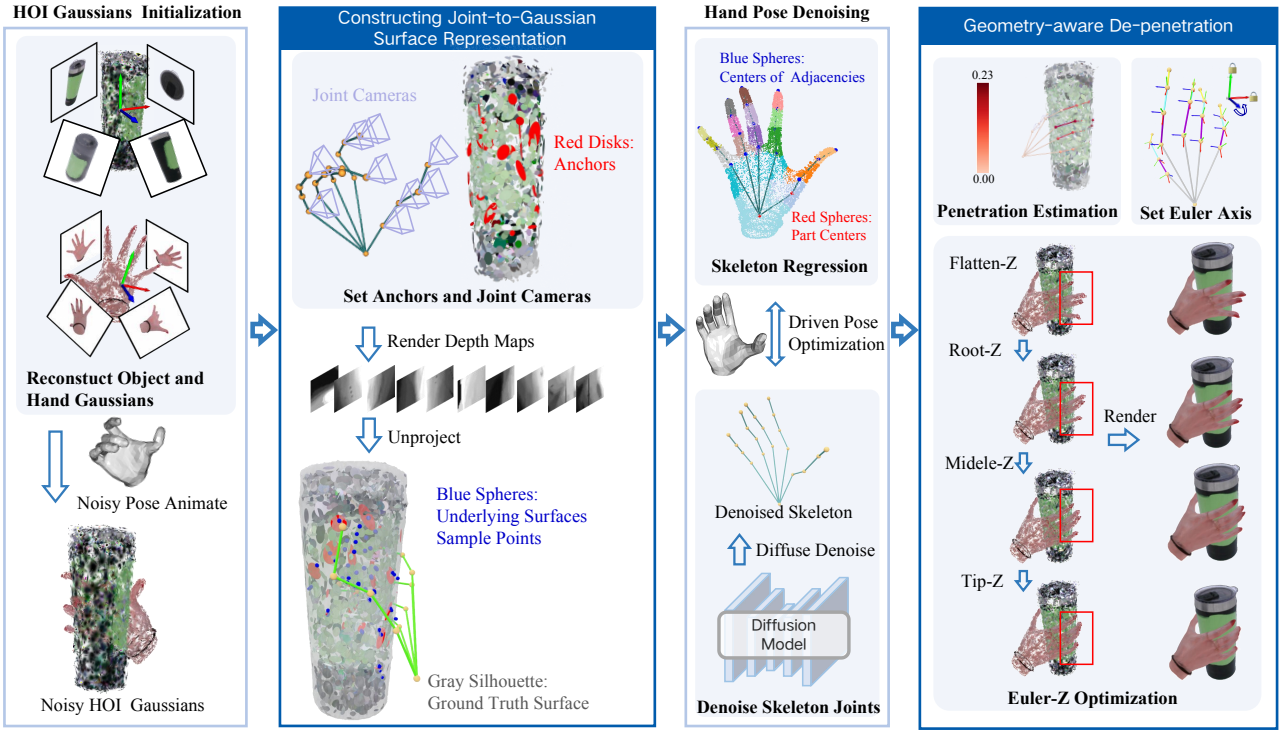


Figure 2: The pipeline of our GSHOI Denoiser method. Constructing J2GS: Disks denote object Gaussians, with anchors colored in red. Joint cameras are attached to the skeleton and oriented to look at the anchors. Depth maps are rendered and unprojected as underlying surface sample points (visualized as blue spheres). Hand Pose Denoising: Skeleton regression calculates the skeleton joints from posed hand Gaussians. We optimize the driven pose supervised by the denoised skeleton from the diffusion Model. GDP: The color bar denotes the estimated skeleton bone penetration depth. To eliminate penetration, we optimize the z-axis Euler angles from the root to the tip bones.

very small field of view (FoV) on each joint and set the extrinsic of each camera to look at a corresponding anchor as:

$$camera(p_0, p_1) = \left[\begin{array}{c|c} lookat(p_0, p_1) & p_0 \\ \hline \mathbf{0} & 1 \end{array}\right].inverse(), \quad (6)$$

where $lookat(p_0, p_1)$ returns the rotation matrix that orients the z-axis forward from p_0 to p_1, and $inverse()$ returns the inverse matrix. Therefore, the $camera(p_0, p_1)$ function returns the extrinsic matrix for positioning the camera at p_0 and orienting it with the z-axis pointing toward p_1.

Our joint cameras are denoted as:

$$\begin{aligned} \mathbf{v}_{f,i,k} &= camera(J_f^k, \mathbf{a}_i.pos), \\ \mathcal{V} &= \{\mathbf{v}_{f,i,k}\}_{f=1\ldots F, i=1\ldots n_a, k=1\ldots n_j}, \end{aligned} \quad (7)$$

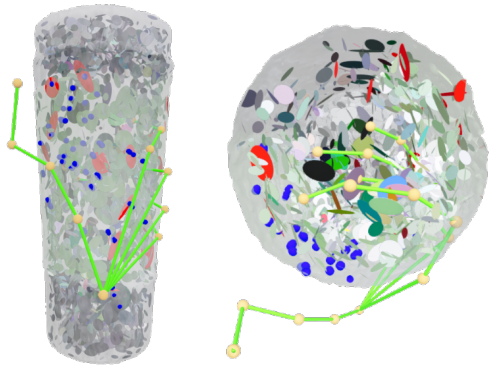
where F denotes the frame length of the joints sequence.

Next, we estimate the object's local underlying surfaces near each anchor and construct the spatial vectors as our J2GS representation. As shown in Eq. 4, we splat object Gaussians onto each joint camera to create local depth maps, where each pixel reflects the Euclidean distance between a joint and a sample point on the continuous underlying surface of object Gaussians near an anchor. We then unproject each pixel to the world coordinate and construct the spatial vectors pointing from sample points to joints, forming our J2GS representation.

In detail, the shape of our J2GS represetation is $\mathcal{R} \in \mathbb{R}^{(F\times n_a\times n_j)\times(h\times w)\times 3}$, where h, w denotes the height and width of the depth maps. We also define the J2GS distance as the norm of J2GS representation averaged across the $h \times w$ dimension as:

$$\mathbf{D}(\mathcal{R}) = \|\mathcal{R}\|_2.mean(dim = 1) \in \mathbb{R}^{(F\times n_a\times n_j)}, \quad (8)$$

where $\mathbf{D}$ is the J2GS distance, and $mean(\cdot)$ is the average function along a given dimension.



(a) Side View (b) See-through Top View

Figure 3: Visualization of our J2GS representation. Yellow spheres and cyan cylinders represent hand skeleton joints and bones, respectively. Blue spheres denote the underlying surface sample points. Disks are 2D Gaussians, where red disks are anchors. The semi-transparent gray silhouette represents the object's GT mesh surface.

Fig. 3(a) shows the visualization of our J2GS representation. The skeleton joints and bones of a penetrated driven pose are rendered as yellow spheres and cyan cylinders. We render the object Gaussians as disks with each Gaussian's position, rotation, scale, and color property. For each anchor, we enlarge the scale property by and set red color for visualization clarity. The blue spheres denote our sample points on the underlying surface of object Gaussians around each anchor. We also render the object's GT mesh as a semi-transparent gray silhouette. Fig. 3(b) are the same scene of 3(a) rendered in see-through mode by setting a larger near clipping plane to reveal the interior of the GT mesh. These visualizations confirm two key findings. 1) Most Gaussians, including anchors, lie *inside* the GT surface *coarsely*. Simply using spatial vectors

pointing from the position of anchor Guassians to the joints leads to a penetrated and coarse HOI representation. 2) Our underlying surface sample points, which are sampled by splatting the coarse Gaussians on joint cameras, lie *on* or *near* the GT surface around the anchors. Based on these sample points, our J2GS representation can model the accurate spatial HOI relations.

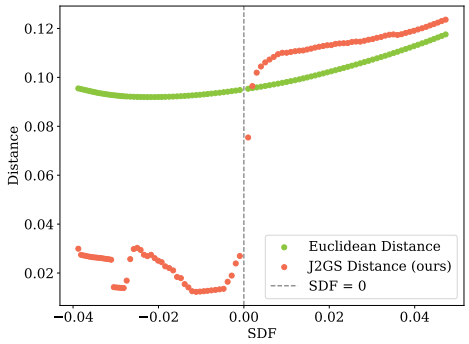


Figure 4: Relationship between a query point's SDF and its distance to an anchor. We sample query points on a ray passing the interior and exterior of the object. For each point we compute its SDF value to the GT mesh, and its Eculidean and J2GS distance to an anchor.

To verify that our J2GS representation can highlight the HOI spatial penetrations, we select query points along the normal direction of a random vertex of the GT mesh. We then compute both the Euclidean and our J2GS distance to a randomly selected anchor. Figure 4 illustrates the relationship between each query point's SDF value and its Euclidean or J2GS distance to the anchor. The Euclidean distance remains continuous for points both outside (SDF > 0) and inside (SDF < 0) the GT mesh. In contrast, our J2GS distance exhibits a clear distribution discrepancy between the external and internal regions. This discrepancy stems from the fact that the object Gaussians are trained using outside-in camera setups, which maintain correct Gaussian depth rendering only when the camera is positioned roughly outside the GT surface and shows coarser surface estimation results within the interior. The external and internal distance distribution discrepancy confirms that our J2GS representation effectively highlights HOI spatial penetrations, benefiting the diffusion model in denoising penetrated hand skeleton joints. Moreover, using a mesh proxy of object Gaussians is not suitable for HOI representation since extracting a mesh from Gaussians via marching cubes inevitably introduces reconstruction errors, causing misalignment between the mesh proxy and the underlying surface of Gaussians. As a result, the denoised HOI that appears accurate on mesh can still produce noisy renderings. Our method uses a unified representation for both rendering and geometric denoising.

3.3 Geometry-aware De-penetration

We propose the Geometry-aware de-penetration algorithm to first estimate the penetrated geometries of HOI, then repose the penetrated fingers appropriately on the underlying surface of object Gaussians. As shown in Algorithm 1, our GDP algorithm takes the f-th frame denoised hand pose θ_f^d, the object Gaussians $\mathcal{G}_o$, and a rendering viewpoint v as inputs, and outputs the de-penetrated hand pose $\hat{\theta}_f$ and corresponding HOI image I_f. We first initialize the optimizable hand pose $\hat{\theta}_f$ and its corresponding skeleton joints $\hat{J}_f$, and decompose [21] the hand pose as the Euler angles $\hat{\Psi}_f \in \mathbb{R}^{15\times3}$ of 15 skeleton bones to control the fingers more intuitively (Lines 1-2). Next, we estimate and eliminate the penetrations of 5 fingers (Lines 3-16). For the l-th finger, we take 3 bones' Euler angles from $\hat{\Psi}_f$ and estimate each bone's penetration depth with function `PeneDepth` (Lines 4-5). If any bone is penetrated (i.e., $\mathbf{d} > 0$), we reset the three bones' Euler z-axis angle as 0 to flatten the finger (Lines 6-7). Then, we iteratively increase the z-axis angle for the root ψ_{root}, then update $\hat{\Psi}_f$ and skeleton joints, then update the penetration depth with function `UpdatePD` until any part of the finger penetrates (Lines 8-10). This operation poses the root bone from flatten to bent until HOI contact occurs. We repeat this process for the middle bone's angle ψ_{mid} until the middle or tip bone penetrates, and we do the same for the tip bone (Lines 11-16). Finally, we compose the optimized Eulers $\hat{\Psi}_f$ as $\hat{\theta}_f$ and render posed hand Gaussians alongside the object Gaussians to produce the HOI image I_f (Lines 17-19).

Algorithm 1: Geometry-aware De-penetration

Input: denoised pose θ_f^d, object Gaussians $\mathcal{G}_o$, viewpoint v.
Output: de-penetrated hand pose $\hat{\theta}_f$, HOI image I_f.

1 $\hat{\theta}_f \leftarrow \theta_f^d$, $\hat{J}_f \leftarrow J_f^d$;
2 $\hat{\Psi}_f \leftarrow decompose(\hat{\theta}_f) \in \mathbb{R}^{15\times3}$;
3 **for** $l \leftarrow 1$ **to** 5 **do**
4 $\quad\{\psi_{root}, \psi_{mid}, \psi_{tip}\} \leftarrow \hat{\Psi}_f.finger(l)$;
5 $\quad\{\mathbf{d}_{root}, \mathbf{d}_{mid}, \mathbf{d}_{tip}\} \leftarrow$ `PeneDepth`$(\hat{J}_f,\ l)$;
6 $\quad$**if** $(\mathbf{d}_{root} > 0)||(\mathbf{d}_{mid} > 0)||(\mathbf{d}_{tip} > 0)$ **then**
7 $\quad\quad\psi_{root}.z,\ \psi_{mid}.z,\ \psi_{tip}.z \leftarrow \mathbf{0}$;
8 $\quad\quad$**while** $(\mathbf{d}_{root} > 0)||(\mathbf{d}_{mid} > 0)||(\mathbf{d}_{tip} > 0)$ **do**
9 $\quad\quad\quad\psi_{root}.z \leftarrow \psi_{root}.z + \delta_z$;
10 $\quad\quad\quad\{\mathbf{d}_{root}, \mathbf{d}_{mid}, \mathbf{d}_{tip}\} \leftarrow$ `UpdatePD`$(\psi_{root}.z, \psi_{mid}.z, \psi_{tip}.z, l)$;
11 $\quad\quad$**while** $(\mathbf{d}_{mid} > 0)||(\mathbf{d}_{tip} > 0)$ **do**
12 $\quad\quad\quad\psi_{mid}.z \leftarrow \psi_{mid}.z + \delta_z$;
13 $\quad\quad\quad\{\mathbf{d}_{root}, \mathbf{d}_{mid}, \mathbf{d}_{tip}\} \leftarrow$ `UpdatePD`$(\psi_{root}.z, \psi_{mid}.z, \psi_{tip}.z, l)$;
14 $\quad\quad$**while** $(\mathbf{d}_{tip} > 0)$ **do**
15 $\quad\quad\quad\psi_{tip}.z \leftarrow \psi_{tip}.z + \delta_z$;
16 $\quad\quad\quad\{\mathbf{d}_{root}, \mathbf{d}_{mid}, \mathbf{d}_{tip}\} \leftarrow$ `UpdatePD`$(\psi_{root}.z, \psi_{mid}.z, \psi_{tip}.z, l)$;
17 $\hat{\theta}_f = compose(\hat{\Psi}_f)$;
18 $I_f = Splat\left(cat(\mathcal{G}_o, \mathcal{G}_h(\hat{\theta}_f)), v\right)$
19 **return** $\hat{\theta}_f, I_f$

Algorithm 2: Penetration Depth Estimation

Input: f-th frame joints $\hat{J}_f$, finger index l.
Output: penetration depth of the l-th finger's bones

1 **Function** `PeneDepth`$(\hat{J}_f,\ l)$**:**
2 $\quad\mathbf{J}_l \leftarrow \hat{J}_f.finger(l) \in \mathbb{R}^{4\times3}$;
3 $\quad\mathbf{e}_{root} \leftarrow \mathbf{J}_l[2] - \mathbf{J}_l[1];\ \mathbf{v}_{root} \leftarrow camera(\mathbf{J}_l[2], \mathbf{J}_l[1])$;
4 $\quad\mathbf{e}_{mid} \leftarrow \mathbf{J}_l[3] - \mathbf{J}_l[2];\ \mathbf{v}_{mid} \leftarrow camera(\mathbf{J}_l[3], \mathbf{J}_l[2])$;
5 $\quad\mathbf{e}_{tip} \leftarrow \mathbf{J}_l[4] - \mathbf{J}_l[3];\ \mathbf{v}_{tip} \leftarrow camera(\mathbf{J}_l[4], \mathbf{J}_l[3])$;
6 $\quad\mathbf{d}_{root} \leftarrow \mathbf{D}\left(\overline{Splat}^d(\mathcal{G}_o, \mathbf{v}_{root}, \|\mathbf{e}_{root}\|_2)\right)$;
7 $\quad\mathbf{d}_{mid} \leftarrow \mathbf{D}\left(\overline{Splat}^d(\mathcal{G}_o, \mathbf{v}_{mid}, \|\mathbf{e}_{mid}\|_2)\right)$;
8 $\quad\mathbf{d}_{tip} \leftarrow \mathbf{D}\left(\overline{Splat}^d(\mathcal{G}_o, \mathbf{v}_{tip}, \|\mathbf{e}_{tip}\|_2)\right)$;
9 $\quad$**return** $\{\mathbf{d}_{root}, \mathbf{d}_{mid}, \mathbf{d}_{tip}\}$;

As shown in Algorithm 2, we define the `PeneDepth` function to estimate the penetration depth of the root, middle, and tip bones of a finger. Given the f-th frame skeleton $\hat{J}_f$ and the finger index l as inputs, `PeneDepth` outputs the penetration depth of the l-th finger's 3 bones. We first get the finger joints $\mathbf{J}_l$ (Line 2), and build the vector of root $\mathbf{e}_{root}$, middle $\mathbf{e}_{mid}$, and tip $\mathbf{e}_{tip}$ bones by connecting each paried joints, and set 3 cameras $\mathbf{v}_{root}, \mathbf{v}_{mid}, \mathbf{v}_{tip}$ along each bone with Eq. 6 (Lines 3-5). We then render 3 depth maps $\{\mathbf{d}_{root}, \mathbf{d}_{mid}, \mathbf{d}_{tip}\}$ using the Gaussian depth rendering function

with the far plane clipping as $\overline{Splat}^{d}(\mathcal{G}, \mathbf{v}, d_{far})$, where d_{far} denotes the distance from the camera to the farplane (Lines 6-8). By setting the far plane as the length of each bone, we can estimate the penetration depth of each bone with our proposed J2GS distance in Eq. 8, since the depth maps are rendered with clipping any Gaussians outside the end of the bone. If no Gaussians are rendered along the bone, the depth is set to 0.

Algorithm 3 shows the detail of `UpdatePD` function, which takes the z-axis Euler angles of 3 bones of the l-th finger as inputs and outputs the updated penetration depth of the 3 bones. We first update the hand Euler angles' corresponding values with input (Line 2). Then we compose $\hat{\Psi}_f$ as hand pose $\hat{\theta}_f$ and calculate the updated skeleton joints $\hat{J}_f$ of the hand Gaussians (Lines 3-4). Finally, we compute the penetration depth by the `PeneDepth` function with $\hat{J}_f$ and l as inputs and return the updated results (Line 5).

Algorithm 3: Penetration Depth Updating

Input: z-axis Euler angle of root ψ_{root}, middle ψ_{mid}, and tip ψ_{tip} bones of the l-th finger.
Output: The updated penetration depth of the 3 bones.

1 **Function** `UpdatePD`$(\psi_{root}, \psi_{mid}, \psi_{tip}, l)$**:**
2 $\quad \hat{\Psi}_f.finger(l) \leftarrow \{\psi_{root}, \psi_{mid}, \psi_{tip}\}$;
3 $\quad \hat{\theta}_f \leftarrow compose(\hat{\Psi}_f)$;
4 $\quad \hat{J}_f \leftarrow \mathcal{R} \cdot \mathcal{G}_h(\hat{\theta}_f).pos$;
5 $\quad$ **return** `PeneDepth`$(\hat{J}_f, l)$;

4 Experiments

4.1 Experimental Settings

Dataset. We conduct both comparison and ablation studies on the MANUS-Grasp dataset [15], which provides HOI image sequences from 38 camera views of 3 subjects and 35 objects, along with per-frame MANO hand pose annotations. It also includes separately captured hand and object images for reconstruction use. We use Subject0 and Subject1 for our experiments. To assess the generalization of HOI denoising methods to new objects, we leave grasp sequences of *cube*, *books1*, *fruits1*, *color1*, *color2*, *color3*, and *color4* as the test set for Subject0, and *tech2*, *color1*, *color2*, *color3*, and *color4* as the test set for Subject1; sequences of all other objects serve as the training set. We retain the final 40 frames of each grasp sequence in both the training and test sets. We remove the background of grasp images using SegmentAnything2 [17] to evaluate the HOI rendering quality. Objects *color1-4* are provided with GT contact maps of hands for evaluation.

Metrics. We introduce 8 evaluation metrics. Penetration Depth (PD) in mm measures HOI penetration by averaging SDF values of the hand Gaussians located inside the object's GT mesh. PSNR, Structural Similarity Index Measure (SSIM), and LPIPS evaluate the HOI rendering. Finally, mean Intersection over Union (mIoU), F1, and accuracy measures the overlap area, the overall similarity, and the contact classification accuracy between predicted and T contact maps, respectively. Mean Per-Joint Position Error (MPJPE) in mm quantifies the average distance between the predicted and GT skeleton joints [28, 11].

Comparison methods. We compare our GSHOI Denoiser with the SOTA HOI rendering and denoising methods MANUS [15], MANUS+GEARS [29], and MANUS+GeneOH [11], as well as 2DGS+GeneOH. As GEARS and GeneOH operate on mesh vertices, we adapt them to Gaussians by treating Gaussian positions as vertices. MANUS drives the hand Gaussians with a kinematic chain. We regress the chain parameters from noisy input poses and denoised poses of GEARS and GeneOH to drive the MANUS hand.

4.2 Implementation Details

Our hands and objects Gaussians are implemented with 2DGS [5]. The diffusion model follows GeneOH [11] exactly. GeneOH provides a pre-trained checkpoint on the synthetic mesh-based HOI dataset GRAB [19]. We fine-tune the checkpoint on the MANUS-Grasp dataset for about 2 hours for both ours and GeneOH on a single NVIDIA RTX 4090 GPU. We set the HOI sequence length to 10 frames. We render the depth maps in 32×32 resolution. Rendering a batch of 5250 depth maps takes only 0.015 seconds. The diffusion model consists of 200 iterative steps, each taking 0.02 seconds. The geometry-aware de-penetration converges in about 300 iterations, with each iteration taking around 0.015 seconds. We perturb each sequence by adding random noise sampled from the normal distribution on the MANO translation, rotation, and pose parameters with the standard deviations set to 0.01, 0.05, 0.4, respectively, to set the input MPJPE around 23.0 following GeneOH. The random perturbations result in not only HOI penetrations but also unnatural hand poses, unstable grasps, and temporal jitter.

4.3 Results and discussion

Quantitative results. The quantitative comparison between our GSHOI Denoiser and the SOTA methods on the MANUS-Grasp dataset for Subject0 and Subject1 is shown in Table 1. For reference, we include two additional rows: 2DGS+GT Pose, in which 2DGS of hands and objects are driven by GT pose, and 2DGS+Input Pose, where the driven pose is the noisy input pose without any denoising. We compare the MANUS, MANUS+GEARS, MANUS+GeneOH, and 2DGS+GeneOH and our GSHOI Denoiser in the subsequent rows. As shown in the table, our GSHOI Denoiser achieves the highest denoising and rendering quality on almost all metrics. Notably, for the PD metric, our method is lower than MANUS+GEARS by 2.26 (64.6%) and 0.73 (17.0%), and is lower than MANUS+GeneOH by 6.08 (83.0%) and 1.64 (31.5%) for Subject0 and Subject1, respectively. Our method also reduces PD by 1.77 (58.8%) and 0.57 (13.8%) compared to 2DGS+GeneOH, highlighting its effectiveness in eliminating HOI penetrations. Regarding rendering quality, our method outperforms MANUS+GEARS by 4.63 (22.0%) and 6.59 (41.1%) in PSNR for Subject0 and Subject1, respectively, and outperforms MANUS+GeneOH by 1.41 (5.7%) and 2.23 (10.8%), and improves 2DGS+GeneOH by 0.14 (0.5%) and 0.06 (0.2%), respectively. For MPJPE, our approach achieves 8.27 (38.7%) and 16.48 (52.5%) lower error than MANUS+GEARS on two subjects, and a 0.22 (1.6%) lower error than MANUS+GeneOH and 0.7 (5.0%) than 2DGS+GeneOH on Subject0, demonstrating more faithful skeleton motion recovery. However, the MPJPE of our method is 1.43 (9.5%) higher on Subject1 compared with MANUS+GeneOH, because our GDP method reposes skeleton joints to eliminate penetration, occasionally introducing slight deviations from GT pose.

We present comparisons of contact maps of our GSHOI Denoiser to MANUS+GeneOH for the MANUS-Grasp dataset in Table 2. The predicted contact maps are rendered by coloring hand Gaussians that come into contact with objects in white, with non-contact Gaussians being black. Following MANUS [15], we consider a hand Gaussian to be in contact if its SDF value to the object's GT mesh is below a threshold of $\tau = 0.004$ at any frame in the trajectory. These results indicate that our GSHOI Denoiser aligns more closely with the GT contact map than MANUS+GeneOH, outperforming it by 0.03 (42.7%), 0.06 (40%), and 4.75 (5.29%) in mIoU, F1, and accuracy on Subject0, respectively, and by 0.02 (23.0%), 0.035 (23.0%), and 3.00 (3.4%) on Subject1.

Visualization. The quantitative improvements in the metrics are also reflected by the visualization results of 2 subjects interacting with 6 new objects *out of the training set*, as presented in Fig. 5. We compare our GSHOI Denoiser to MANUS, MANUS+GEARS, MANUS+GeneOH, and 2DGS+GeneOH. For

Table 1: Quantitative comparisons of our GSHOI Denoiser to SOTA methods on the MANUS-Grasp Dataset.

Method	PD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MPJPE ↓	PD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MPJPE ↓
	Subject0					Subject1				
2DGS+GT Pose	1.69	29.61	0.977	0.034	4.86	2.13	29.00	0.978	0.038	4.87
2DGS+Input Pose	5.18	24.21	0.968	0.054	23.18	5.97	22.62	0.967	0.062	23.10
MANUS	3.52	23.75	0.968	0.056	22.42	5.73	21.26	0.963	0.070	24.40
MANUS+GEARS	3.50	21.10	0.963	0.075	21.36	4.29	16.89	0.948	0.113	31.40
MANUS+GeneOH	7.32	24.32	0.968	0.052	13.31	5.20	21.51	0.964	0.066	**13.49**
2DGS+GeneOH	3.01	25.59	0.970	0.046	13.79	4.13	23.78	0.969	0.055	14.68
GSHOI Denoiser (ours)	**1.24**	**25.73**	**0.971**	**0.045**	**13.09**	**3.56**	**23.84**	**0.970**	**0.054**	14.92

reference, we also include 2DGS+GT Pose and the GT image. We summarize the qualitative improvements of our GSHOI Denoiser as follows. 1) Penetration: In the first 5 rows, the perturbed input poses cause severe HOI penetration, yielding unpleasant renderings from MANUS, which do not consider HOI denoising. MANUS+GEARS tends to produce great gaps between hands and objects (third and fourth rows). MANUS+GeneOH alleviates penetration to some degree, but residual penetration still persists. For instance, the middle finger penetrates the book (first row), the thumb intersects the bottle (second row), the index finger is embedded in the apple (third row), and multiple fingers penetrate the bottle and mug (fourth and fifth rows). 2DGS+GeneOH also involves penetrations, as shown in the middle finger (first row), the little finger (second row), and the ring finger in the fourth row. In contrast, our GSHOI Denoiser successfully eliminates these penetrations by repositioning the fingers on the object surface, resulting in more accurate and visually appealing HOI renderings. 2) Pose twist: In the sixth row, the input pose has a twist posed thumb and penetrated middle and ring fingers. MANUS+GEARS alleviates the pose twist but also leaves the hand too far from the object. While MANUS+GeneOH corrects the middle and ring fingers, it fails to solve the unnatural twist of the thumb. 2DGS+GeneOH removes the twist but lifts the ring finger unnaturally. By contrast, our GSHOI Denoiser corrects the problematic fingers without interfering with other accurate fingers, avoiding any additional noise. 3) Unstable grasp: The input pose of the last two rows involves distance noises between fingers and objects, leading to unstable grasp, as shown by MANUS. MANUS+GEARS fails to move hands near the objects. MANUS+GeneOH produces over-grasped and over penetrated poses. For example, the hand grabs the inside of the mug instead of the outsides (seventh row), and the thumb is placed far away from the earphones (eighth row). 2DGS+GeneOH results contain minor penetration, as shown in the index finger in the seventh row and the distance gap of the ring finger in the eighth row. Our method accurately poses the unstable hands on the objects and shows great fidelity with the GT Pose.

We summarize the cause of penetration, pose twistion, and unstable grasp as follows. GeneOH train diffusion model to learn the distribution of HOI spatial representations, which are defined by vectors from the centers of object anchor Gaussians to skeleton joints. However, as demonstrated in Fig. 3, the Gaussians are coarsely distributed *inside* the object's GT surface, resulting in the modeled HOI distances being more extended than that of GT. Consequently, during inference, given new object Gaussians, if the selected anchors lie far inside the object surface, the resulting poses are more likely to penetrate the object. Conversely, if anchors are placed too close to the object surface, the resulting hand pose may appear large distance from the object, leading to unstable grasps. Furthermore, since Gaussians are randomly distributed, selecting anchors that produce novel spatial directions can lead to twisted hand poses. Our GSHOI Denoiser overcomes these limitations by leveraging the proposed J2GS representation, which samples points *on* the object's GT surface, thus making the distribution of HOI spatial representation uniform across the training and test set. Our GDP further eliminates penetrations, enabling our method to achieve higher-fidelity HOI renderings.

Table 2: Quantitative comparisons of our GSHOI Denoiser to MANUS+GeneOH on the contact maps of MANUS-Grasp Dataset.

Method	mIoU ↑	F1 ↑	Accuracy (%) ↑
	Subject0		
2DGS+GT Pose	0.147	0.255	94.50
2DGS+Input Pose	0.097	0.172	90.00
MANUS+GeneOH	0.082	0.150	89.75
GSHOI Denoiser (Ours)	**0.117**	**0.210**	**94.50**
	Subject1		
2DGS+GT Pose	0.142	0.237	94.00
2DGS+Input Pose	0.087	0.147	86.75
MANUS+GeneOH	0.087	0.152	89.75
GSHOI Denoiser (Ours)	**0.107**	**0.187**	**92.75**

4.4 Ablation Study

We conduct ablation studies on the MANUS-Grasp dataset to evaluate the effectiveness of our proposed components, J2GS and GDP. First, we use the 2DGS using the input noisy pose without denoising methods as our baseline. Second, we apply our J2GS representation to construct the HOI relations and train our diffusion model (J2GS). Third, we conduct experiments with our proposed GDP without the diffusion model to eliminate penetration (GDP). Fourth, we compose our J2GS with GDP as our complete GSHOI Denoiser (J2GS+GDP). Quantitative results are presented in Table 3, and qualitative results are visualized in Fig. 6.

Quantitative results. As shown in Table 3, compared with the baseline, our J2GS reduces the PD by 1.33 (47.0%) for Subject0 and 1.73 (29.0%) for Subject1, and improves PSNR by 1.51(6.2%) and 1.21 (5.3%) repectively. Applying only GDP without a denoised pose from the J2GS does not yield strong performance because GDP supposes an approximately correct input pose and refines only penetrated fingers, verifying the importance of J2GS. Our complete GSHOI Denoiser (J2GS+GDP) further eliminates PD over the J2GS by a large margin of 1.59 (56.2%) for Subject0 and 0.68 (16.0%) for Subject1, and also achieves the best rendering quality in PSNR, SSIM, and LPIPS metrics for both subjects, showing the effectiveness of the proposed J2GS and GDP. Although J2GS+GDP increases MPJPE slightly (by 0.05 for Subject0 and 0.36 for Subject1) relative to J2GS. This increase arises because GDP reposes penetrated fingers. Though we only optimize the z-axis of each finger's Euler angles and fix the x and y-axes to preserve pose consistency, the optimized skeleton joints can still deviate from the GT. Compared with GDP, our J2GS+GDP decreases PD by 0.26 (17.3%) for Subject0 and 0.14 (3.8%) for Subject1, and also increases PSNR by 1.51 (6.2%) and 1.22 (5.4%), respectively.

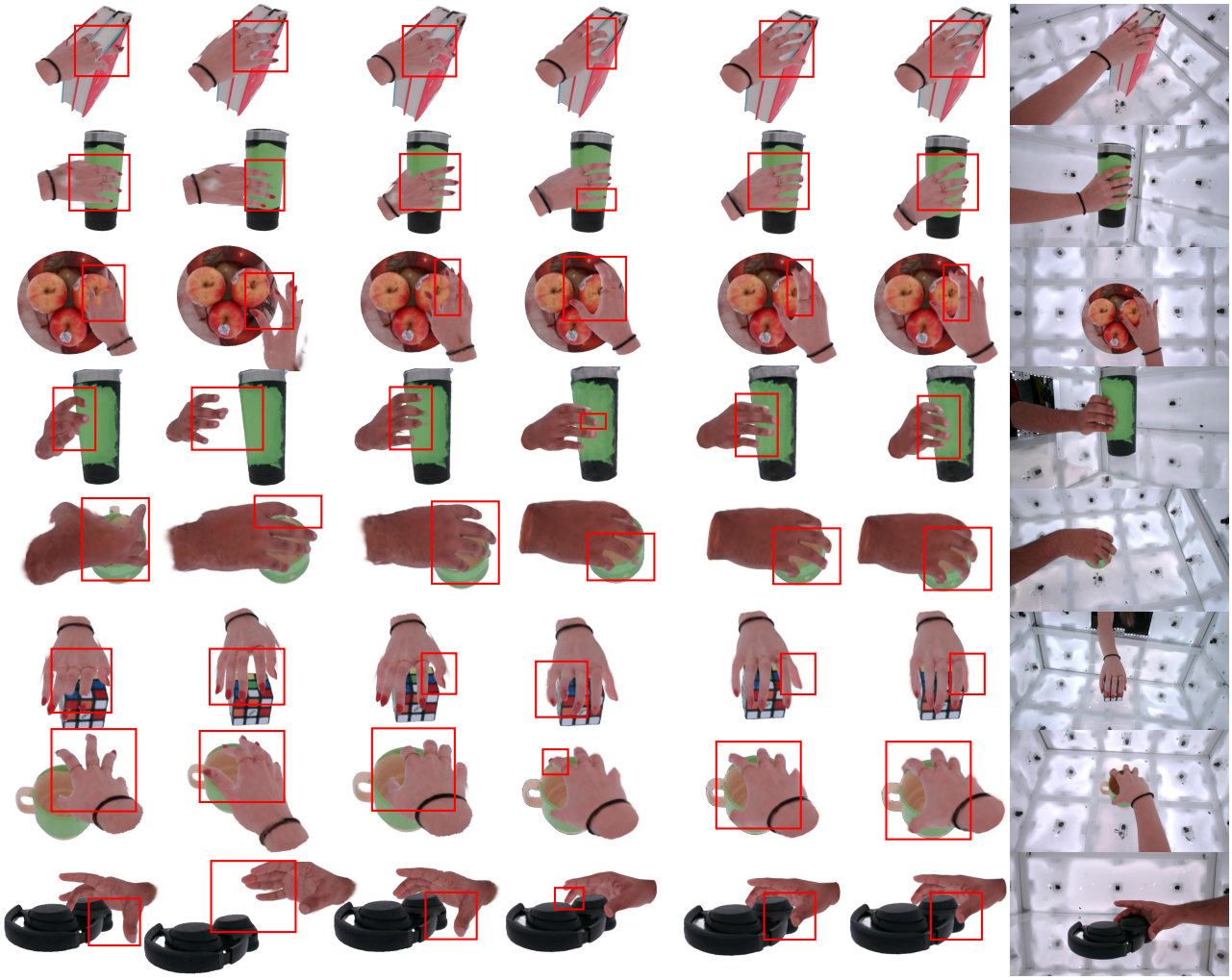


Figure 5: Visual results of MANUS, MANUS+GEARS, MANUS+GeneOH, and our GSHOI Denoiser on MANUS-Grasp Dataset.

Table 3: Ablation study for our J2GS and GDP components on the MANUS-Grasp dataset.

Method	PD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MPJPE ↓	PD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MPJPE ↓
	Subject0					Subject1				
Baseline	5.18	24.21	0.968	0.054	23.18	5.97	22.62	0.967	0.062	23.10
Baseline+J2GS	2.83	25.72	0.970	0.046	**13.04**	4.24	23.83	**0.970**	**0.054**	**14.56**
Baseline+GDP	1.50	24.22	0.968	0.054	22.62	3.70	22.62	0.966	0.062	23.49
Baseline+J2GS+GDP	**1.24**	**25.73**	**0.971**	**0.045**	13.09	**3.56**	**23.84**	**0.970**	**0.054**	14.92

Visualization. The visualization results in Fig. 6 further demonstrate the effectiveness of our proposed J2GS and GDP. The comparison results can be summarized as follows: 1) Penetration: In the first row, the input pose exhibits penetration. Our J2GS produces HOI with cleaner middle finger than the input pose, but the penetration of the index finger still remains. With the help of GDP, our J2GS+GDP fully eliminates penetration while preserving the de-penetrated hand poses similar to the J2GS results. Notably, directly applying GDP can also remove penetration, but the results deviate more from the GT Pose. 2) Unstable grasp: In the second row, the input pose incorporates an unstable grasp pose to the mug. Using J2GS positions the hand firmly to grasp the mug, resulting in a configuration closely matching the GT pose, but the index finger is slightly penetrated. Our J2GS+GDP effectively removes the penetration artifact. Using only GDP without J2GS eliminates penetration but yields a pose less similar to the GT pose.

4.5 User Study

We conducted a within-subject study to evaluate the visual perceptual quality of HOI rendering produced by our GSHOI Denoiser and MANUS. The visual perceptual quality is measured by two 5-point Likert-scale subjective metrics of penetration and instability. Lower scores indicate reduced perceptual HOI rendering noise.

Participants and Conditions. We recruited 15 participants (10 males, 5 females) between 20 and 27 years of age. The study included two conditions: GSHOI Denoiser and MANUS, both trained on the MANUS-Grasp dataset.

Task. We use the LeapMotion to capture users' hand skeleton joints and regress them as driven pose parameters. The hand Gaussians are driven by users' moving hands and rendered along with the static object Gaussians in real time. Participants are asked to grasp the virtual object with their mid-air hand poses and control the virtual hand to fit the object surface as closely as possi-

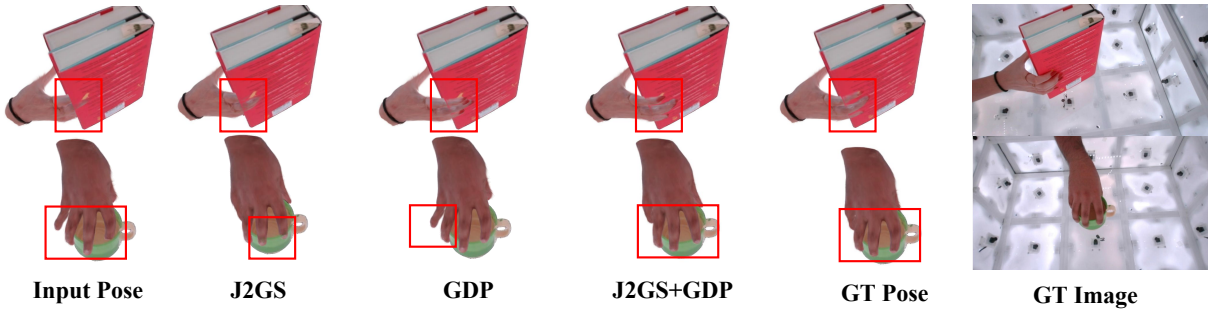


Figure 6: Visual results of ablation study for our J2GS and GDP components on the MANUS-Grasp dataset.

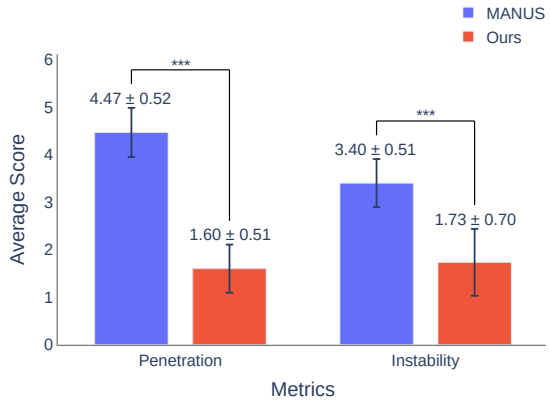


Figure 7: Statistic results of the user study for scoring the penetration and grasp stability of our GSHOI Denoiser and the MANUS. Lower penetration scores and higher stability scores denote a more plausible user experience with HOI rendering.

ble while avoiding penetration. Each participant experiences both conditions in a random order. To mitigate the effect of visual fatigue, after completing the first condition, participants are given a few minutes of rest before proceeding to the following condition. For the MANUS condition, participants can press a button to pause the hand tracking and observe the HOI rendering with a fixed hand pose. For our GSHOI Denoiser condition, when the button is triggered, we additionally apply HOI denoising process before fixing the hand pose.

Results. Fig. 7 shows the statistical results of the average score over all participants for the penetration and stability metrics. The results indicate that for the penetration metric, our GSHOI Denoiser achieves an average score with a mean of 1.06 and a standard deviation of 0.51, while MANUS achieves a higher penetration average score with a mean of 4.47 and a standard deviation of 0.52. We apply the p-value and $Cohen's\ d$ to estimate the average score differences. The p-value < 0.001 indicates a *significant* de-penetration effect of our method, and the $Cohen's\ d = 5.6 > 0.8$, indicating a $huge$ effect size. For the instability metric, our method achieves an average score with a mean of 1.73 and a standard deviation of 0.70. The p-value < 0.001 and $Cohen's\ d = 2.7 > 0.8$ indicate that our method achieves a more stable grasp pose than MANUS *significantly*. We infer the reason for our method's superior performance on the two metrics as follows. The tracking hand skeleton joints from the Leap Motion controller are very noisy. Regressing driven poses from the tracking data in real time further introduces additional noise, which makes it difficult for users to control the hand to fit on the object surface precisely via mid-air poses without any haptic feedback of physical contact. Our GSHOI Denoiser effectively denoises the HOI, producing plausible grasp poses. In contrast, MANUS does not consider any denoising strategy, usually leading to less satisfactory user experiences. After completing the questionnaire, we interviewed users regarding the physical plausibility of the renderings. Three users reported that when using MANUS, the hand-tracking noise prevented precise pose control, resulting in grasps lacking physical plausibility. When using the GSHOI Denoiser, they only needed to pose their hands roughly near the object, and the denoised grasp poses look physically plausible.

5 Conclusion

In this paper, we have introduced the Gaussian Hand-Object Interaction Denoiser, the first Gaussian splatting-based hand-object interaction denoising method, which effectively denoises the input twisted and penetrated hand poses to produce photorealistic results. We first propose the novel joint-to-Gaussian surface representation, which accurately models the spatial relationships between hand skeleton joints and object Gaussians while emphasizing hand–object penetrations and generalizing well to new hand poses and objects. We then propose a geometry-aware de-penetration algorithm that eliminates penetrations by detecting intersections between skeleton bones and object Gaussians and reposing any penetrated fingers onto the estimated underlying surface of the object. Compared to the state-of-the-art method, GeneOH+MANUS, our method not only effectively reduces penetration but also produces more realistic rendering quality. User study results show that our method can improve the user's subjective experience significantly.

Despite our method's effectiveness, it also involves limitations. First, the finger lengths of the pretrained hand Gaussians differ from those of the user's physical hand. Directly driving the hand Gaussians using pose angles tracked from the user leads to joint misalignments, especially at the fingertips. As future work, we plan to fine-tune the finger length of the hand Gaussians to better fit each user's hand, enabling more accurate pose-driven animation. Second, reconstructing hand and object Gaussians presently relies on dense-view image data with high-quality camera annotations and background segmentations. In future work, we aim to achieve hand–object reconstructions from single-view or sparse-view images by leveraging image-generation models. Third, though our method supports real-time HOI rendering, the diffusion and de-penetration processes take about 10 seconds in total. Applications using our method could initially render the noisy HOI input and incrementally update the rendering as long as the GSHOI Denoiser completes one iteration and produces an intermediate result. We suggest two directions for future speed improvement. 1) Replace the diffusion model with more efficient generative models [9, 3] capable of producing results in a single forward pass. 2) Substitute the iterative de-penetration with a closed-form solution: given the hand Gaussians with known shape, first derive the target joint positions so that each finger precisely contacts the object surface, and then recover the corresponding bone rotations via inverse kinematics.

Acknowledgments

This work is supported by the National Natural Science Foundation of China through Project 61932003, 62372026, by Beijing Science and Technology Plan Project Z221100007722004, and by National Key R&D plan 2019YFC1521102, and by the fundamental research funds for the central universities.

REFERENCES

[1] R. Fan, J. Wu, X. Shi, L. Zhao, Q. Ma, and L. Wang. Fov-gs: Foveated 3d gaussian splatting for dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[2] Z. Fan, M. Parelli, M. E. Kadoglou, M. Kocabas, X. Chen, M. J. Black, and O. Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 494–504, 2024.

[3] K. Frans, D. Hafner, S. Levine, and P. Abbeel. One step diffusion via shortcut models, 2025.

[4] M. Höll, M. Oberweger, C. Arth, and V. Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 175–182. IEEE, 2018.

[5] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

[6] J. Jiang, L. Zhao, X. Lu, W. Hu, I. Razzak, and M. Wang. Dhgcn: dynamic hop graph convolution network for self-supervised point cloud learning. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, pp. 12883–12891, 2024.

[7] J. Jiang, Q. Zhou, Y. Li, X. Zhao, M. Wang, L. Ma, J. Chang, J. Zhang, X. Lu, et al. Pcotta: Continual test-time adaptation for multi-task point cloud understanding. *Advances in Neural Information Processing Systems*, 37:96229–96253, 2024.

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[9] D. Kim, C.-H. Lai, W.-H. Liao, N. Murata, Y. Takida, T. Uesaka, Y. He, Y. Mitsufuji, and S. Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

[10] X. Liu, B. Wang, H. Wang, and L. Yi. Few-shot physically-aware articulated mesh generation via hierarchical deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 854–864, 2023.

[11] X. Liu and L. Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] X. Liu, J. Zhang, R. Hu, H. Huang, H. Wang, and L. Yi. Self-supervised category-level articulated object pose estimation with part-level se (3) equivariance. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[14] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024.

[15] C. Pokhariya, I. N. Shah, A. Xing, Z. Li, K. Chen, A. Sharma, and S. Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2197–2208, 2024.

[16] W. Qu, Z. Cui, Y. Zhang, C. Meng, C. Ma, X. Deng, and H. Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15100–15111, 2023.

[17] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[18] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[19] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.

[20] L. Yang, K. Li, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] L. Yang, X. Zhan, K. Li, W. Xu, J. Zhang, J. Li, and C. Lu. Learning a contact potential field for modeling the hand-object interaction, 2024. doi: 10.1109/TPAMI.2024.3372102

[22] Y. Ye, A. Gupta, K. Kitani, and S. Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, 2024.

[23] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023.

[24] Y. Ye, X. Li, A. Gupta, S. D. Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023.

[25] L. Zhao, X. Lu, Q. Bao, and M. Wang. In-place gestures classification via long-term memory augmented network. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 224–233. IEEE, 2022.

[26] L. Zhao, X. Lu, R. Fan, S. K. Im, and L. Wang. Gaussianhand: Real-time 3d gaussian rendering for hand avatar animation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2024. doi: 10.1109/TVCG.2024.3516778

[27] L. Zhao, X. Lu, M. Zhao, and M. Wang. Classifying in-place gestures with end-to-end point cloud learning. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 229–238. IEEE, 2021.

[28] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2022.

[29] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[30] K. Zhou, Z. Cheng, H. P. Shum, F. W. Li, and X. Liang. Stgae: Spatial-temporal graph auto-encoder for hand motion denoising. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 41–49. IEEE, 2021.

[31] K. Zhou, Y. Ma, H. P. Shum, and X. Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7749–7763, 2023.

[32] K. Zhou, H. P. Shum, F. W. Li, and X. Liang. Multi-task spatial-temporal graph auto-encoder for hand motion denoising. *IEEE Transactions on Visualization and Computer Graphics*, 30(10):6754–6769, 2023.