

Beyond the Contact: Discovering Comprehensive Affordance for 3D Objects from Pre-trained 2D Diffusion Models

Hyeonwoo Kim^{1*}, Sookwan Han^{1*}, Patrick Kwon², and Hanbyul Joo¹

¹ Seoul National University

² Naver Webtoon AI

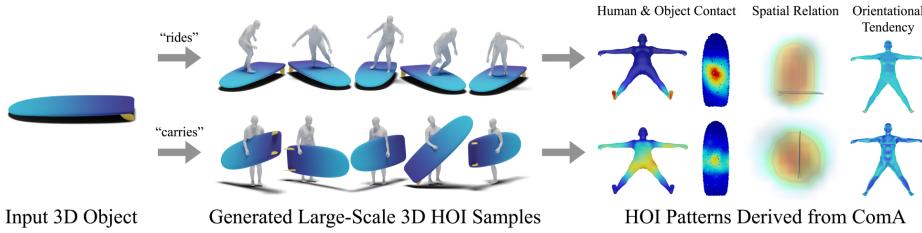


Fig. 1: Given a 3D object, we generate numerous 3D Human-Object Interaction (HOI) samples using text prompts, and learn a novel affordance representation called *Comprehensive Affordance* (ComA) which models both contact and non-contact HOI patterns.

Abstract. Understanding the inherent human knowledge in interacting with a given environment (*e.g.*, affordance) is essential for improving AI to better assist humans. While existing approaches primarily focus on human-object contacts during interactions, such affordance representation cannot fully address other important aspects of human-object interactions (HOIs), *i.e.*, patterns of relative positions and orientations. In this paper, we introduce a novel affordance representation, named *Comprehensive Affordance* (ComA). Given a 3D object mesh, ComA models the distribution of relative orientation and proximity of vertices in interacting human meshes, capturing plausible patterns of contact, relative orientations, and spatial relationships. To construct the distribution, we present a novel pipeline that synthesizes diverse and realistic 3D HOI samples given any 3D object mesh. The pipeline leverages a pre-trained 2D inpainting diffusion model to generate HOI images from object renderings and lifts them into 3D. To avoid the generation of false affordances, we propose a new inpainting framework, *Adaptive Mask Inpainting*. Since ComA is built on synthetic samples, it can extend to any object in an unbounded manner. Through extensive experiments, we demonstrate that ComA outperforms competitors that rely on human annotations in modeling contact-based affordance. Importantly, we also showcase the potential of ComA to reconstruct human-object interactions in 3D through an optimization framework, highlighting its advantage in incorporating both contact and non-contact properties.

Keywords: Affordance · Human-Object Interaction · Diffusion Model

*Indicates equal contribution

1 Introduction

Humans possess an innate ability to perceive the functionalities provided by the environment or objects and efficiently utilize them—a concept known as affordance [18]. Affordance incorporates a variety of patterns typically observable in human-object interactions (HOIs), including not only plausible physical contact, but also spatial and orientational non-contact relationships. For example, while using a laptop, the orientation and distance patterns between human face and the screen can be considered part of affordance, alongside the hand regions touching the laptop keyboard. Teaching such affordance knowledge to machines is crucial for enabling them to better understand or mimic human behaviors. Consequently, numerous studies emerged to implement such affordance knowledge into AI and robotics [1, 21, 24, 39, 42, 65, 83, 91, 93].

Most previous approaches however, primarily focus on a limited spectrum of affordances. Some approaches represent the human and object regions in contact using contact scores on 2D images [1, 42], 3D objects [91], and 3D humans [24, 83] using direct supervision from human annotations. Others represent affordances by inferring [65] or generating [39, 93] plausible human-object pairs in contact-oriented object categories. While a few recent studies tackle learning the spatial relations [21, 92] between a human and an object, these methods mainly focus on object categories where physical contact is dominant than non-contact patterns.

In this paper, we introduce *Comprehensive Affordance* (ComA), a novel representation of an affordance encompassing the various aspects of human-object interaction, including orientational tendencies and relative positions. The motivation for our design arises from the observation that there exist typical 3D relations between entire body parts and objects during HOIs. For instance, human faces, torsos, and arms often exhibit specific distances and orientations with some variability, when interacting with certain object categories, as shown in Fig. 2. ComA is designed to model these patterns by constructing a distribution between each pair of object surface points and human surface points in 3D. This distribution represents the likelihood of human body parts displaying specific spatial and orientational relations with object points, as shown in Fig. 3.

Learning ComA requires various 3D HOI samples showing the way humans use the target object in 3D, from which pairwise distributions of 3D HOI relations can be constructed. Manually annotating such cues is challenging, compared to the annotations focusing solely on contact regions [42, 65, 83, 89, 91, 92]. As a key solution, we present a pipeline to synthesize a large-scale synthetic 3D HOI samples leveraging a pre-trained 2D diffusion model. Interestingly, we observe that a pre-trained 2D diffusion model captures the affordance knowledge in the form of 2D images, as shown in Fig. 2, where the images are synthesized from text prompts. However, leveraging these 2D HOI cues to construct ComA for the given 3D object (which requires 3D HOI samples) presents the following challenges: (1) the diffusion model needs to be applied for inserting plausible humans without altering the original shape and pose of the target object, and (2) 3D HOI samples should be lifted from the synthesized 2D images. To address these challenges, we present *Adaptive Mask Inpainting* to insert humans without

**Text Prompt:** 1 person using a laptop**Text Prompt:** 1 person using a telescope

Fig. 2: Typically, people (1) view the screen (2) from relatively distant distance while using a laptop (Left), whereas they (1) peer into it (2) from a close distance while using a telescope (Right). Pre-trained diffusion model has a knowledge of these (1) orientational and (2) spatial relation between human and object during interaction.

altering the appearance of target objects in 2D, along with a 3D lifting pipeline from 2D HOI cues.

We demonstrate the efficacy of our approach by learning ComA on 100 3D object meshes in diverse categories collected from various sources [2, 10, 30, 54, 75, 87]. We compare ComA with existing approaches [83, 91] on the BEHAVE dataset [2], demonstrating that ours learned from synthetic cues outperforms competitors trained on manual annotations for contact-based affordance. Additionally, we verify the advantage of ComA against contact-only representations, in the scenario of reconstructing HOIs via an optimization framework. We will release the results and our source code¹.

In summary, our main contributions can be summarized as follows: (1) we introduce a new representation ComA, which models the distribution of both explicit contact and non-contact patterns during HOIs; (2) we present a pipeline to build ComA from numerous synthetic samples by leveraging a pre-trained 2D diffusion model, scalable to any 3D object without requiring laborious annotations; and (3) we propose Adaptive Mask Inpainting, a method for preserving the original context of an image when using inpainting diffusion model.

2 Related Work

Learning Visual Affordance. Affordance was introduced by Gibson [18] as a set of functionalities that the environment furnishes to an agent (*e.g.*, human, animal). The concept extends to the vision and robotics area, where the focus is on teaching embodied agents to interact with scenes [1, 16, 41, 58, 88], mimicking human-object [2, 8, 11, 14, 24, 29, 34, 44, 56, 65, 67, 71, 78, 96, 97], or hand-object [3–5, 80, 101] interactions. Earlier methods focused on learning action category labels [6, 26, 40] with bounding boxes but lacked a detailed description of affordance. Due to the limited affordance information available from text descriptions or labels, some studies depicted affordance as contact regions or heatmaps on objects [1, 42, 91] and humans [24, 83]. However, learning affordances beyond the contact was challenging since existing datasets created

¹<https://github.com/snuvclab/coma>

through a multi-camera system [2, 30], motion capture [23], and manual process [13, 32, 83, 91] focused only on annotating contact information rather than capturing the holistic aspect of HOI. Following studies generate plausible HOI samples in 2D [39, 93] or 3D [89, 98] which has the potential for learning such broader HOI knowledge, but do not address the problem directly. In contrast, our method generates 3D HOI samples and extracts ComA for modelling beyond the contact.

Data Synthesis for Learning. There have been approaches to supplement data in fields lacking efficient annotation process by leveraging generative models, irrespective of the research area. Many methods employ GANs [19, 35] or diffusion models [28, 69, 77] to augment data [25, 82] for various vision tasks, including perception and representation learning [86], classification [9, 49, 81], segmentation [51, 84, 100], dense visual alignment [64], and further extending to 3D tasks, such as neural rendering [22, 99], shape reconstruction [62], and so forth. The major challenge in leveraging synthesized data (or generative models for cues) is controllability, as the generator, even when conditioned, typically produces free samples. Ali *et al.* [31] presents a method to control the viewpoint of the synthesized image by modeling the viewpoint-free latent space. CHORUS [21] applies cascaded filtering to control the quality of the generated dataset to learn 3D human-object spatial relations, similar to our approach.

Diffusion Model for Synthesizing HOI Images. While diffusion models [28, 76, 77] excel at generating realistic images [66, 69], few address the generation of 2D human-object pairs (*i.e.*, HOI images). Inpainting models [48, 69], while a common choice for human insertion, often compromise scene details, resulting in images with false affordances. Image editing techniques [27, 52, 55] may offer an alternative but tend to prioritize style changes and struggle with creating new geometry. Recently, Ye *et al.* [93] employs an additional layout network to determine the inpainting region, and Kulal *et al.* [39] trains a diffusion model on video clips to generate HOI images. In contrast to these approaches, we create HOI images by inpainting humans while maintaining the context of the original image, without additional network training. Concurrent work by Li *et al.* [43] introduces *Dynamic Mask Inpainting*, which uses attention masks to change inpainting masks over timesteps to insert human into a rendered 3D scene, while we explicitly apply segmentation model to adapt the inpainting mask.

3 Method

Given an input 3D object mesh, our goal is to learn the novel affordance representation, ComA, which models the distribution of relative proximity and orientation between human and the object surface points. Our pipeline begins with creating 2D HOI images (Sec. 3.1). These images are then lifted to 3D, resulting large-scale synthetic 3D HOI samples (Sec. 3.2). From these 3D samples, we finally learn ComA (Sec. 3.3). An overview of our method is shown in Fig. 3.

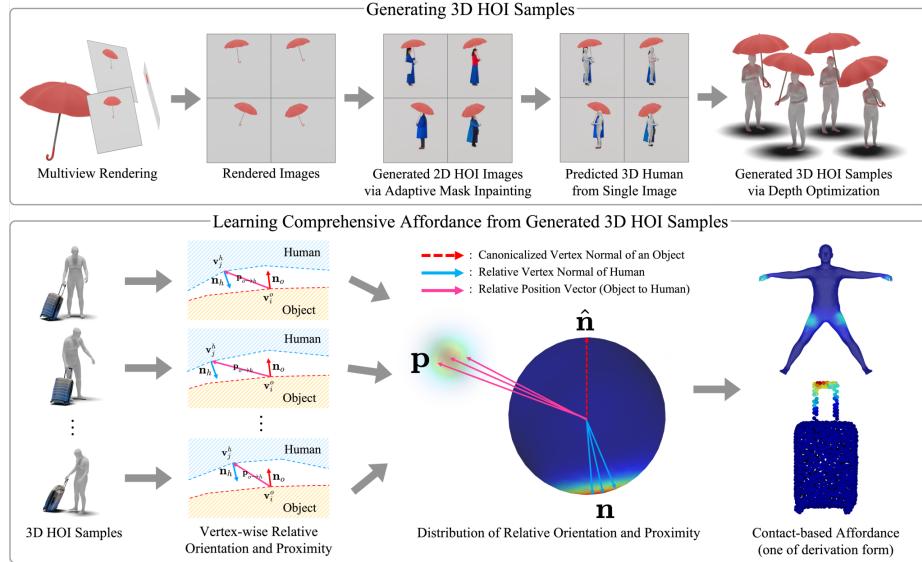


Fig. 3: Method Overview. Our method can be divided into two parts: (1) Generating 3D HOI Samples and (2) Learning ComA from Generated 3D HOI Samples. In the first step, we utilize an inpainting diffusion model with our *Adaptive Mask Inpainting* to create 2D HOI images, and generate 3D HOI samples via uplifting pipeline. In the second step, the generated 3D HOI samples are aggregated to create distributions for relative proximity and orientation, which can be derived into various affordance forms.

3.1 2D HOI Image Generation

Rendering Object from Multi-Viewpoints. Given a 3D object, we first render it from multiple viewpoints. We place the object and camera differently for two object types: *static* and *dynamic*. For objects that are presumed to be fixed on the ground during interaction (*static*: e.g., chair), we fix the object on the ground plane, whereas objects with dynamic pose during interaction (*dynamic*: e.g., cup) are perturbed with sampled rotation and translation. Type of the object is inferred automatically by rendering the object at any direction first and then utilizing vision-language model [61] for predicting. The rendering process can be fully automated, but we empirically find it beneficial to leverage a small amount of human labor, such as restricting the range of perturbation due to the inherent bias of pre-trained diffusion models. For rendering, we place weak perspective cameras $\{\Pi_c\}_{c=1}^N$ around the object with equal azimuth interval and same elevation, where N is set as 8 for *static*, 40 for *dynamic* objects.

Inpainting Mask Selection. After rendering the images, we select the masks to inpaint a human interacting with the object. While our inpainting pipeline is quite robust to initial masks, we found that initial mask selection is helpful to avoid failure cases such as generating hallucinated objects or ambiguous interactions. Specifically, we move sliding windows around the object and choose the windows whose Intersection over Union (IoU) with the object segmentation mask falls within a specific range. See Supp. Mat. for more details.

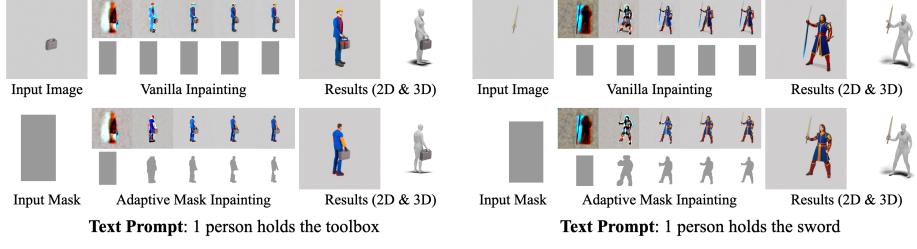


Fig. 4: Adaptive Mask Inpainting. Without Adaptive Mask Inpainting, the original object is damaged when inserting humans, resulting in false affordances.

Prompt Generation. Inspired by CHORUS [21], we automatically generate the prompts that describe the possible interactions with humans and the given object, which are taken as guidance for inpainting. Specifically, we utilize the vision-language model [61] to infer the HOI prompt using the rendered object image and predefined template. We generate 3 prompts per given object.

Adaptive Mask Inpainting for Human Insertion. We synthesize 2D images of human-object interaction by inpainting the human into rendered object images using publicly available text-conditional inpainting diffusion models [69]. However, the challenge arises as the object geometry and details within the mask region are not preserved during inpainting, resulting in false affordances. To mitigate this problem, we present Adaptive Mask Inpainting to progressively specify the inpainting region over diffusion timesteps. Let the sequence of denoised image latent be $\{x_t\}_{t=1}^T$, where x_T is the fully random noise and x_0 is the denoised latent. Inspired by the observation that the quality of the predicted denoised image \hat{x}_0 improves over progress of timestep, and creates a low-level structure of the target prompt (in this case, human) even at the early steps as shown in Fig. 4, we aim to *adapt* the inpainting region over timesteps by spatially discovering the mask from the low-level structure. Specifically, we apply PointRend [38] to decoded latent $\mathbf{D}(\hat{x}_0)$ at specific timesteps to predict the human mask, and consequently use as the mask for the next denoising step.

3.2 Lifting 2D Affordance to 3D

From the previous section, we obtain a set of 2D HOI images $\{I_d\}_{d=1}^D$. Even though we have multiple views with geometrically consistent object renderings, the inserted humans have diversity without multi-view consistency, and thus it is non-trivial to directly lift 3D humans from these 2D images.

To address this, we first utilize a single view 3D human prediction model to obtain a corresponding 3D human pose and shape of the image, and then estimate the depth of the human with weak auxiliary cue provided by the set of 2D HOI images, generating a 3D HOI sample corresponding to the image.

Single View 3D Human Prediction. To uplift the 2D HOI images into 3D, we first predict the pose and shape of 3D humans from the images. We apply off-the-shelf 3D human prediction model [57] (denoted as $\mathbf{F}_{\text{human}}$) to predict

SMPL-X humans [63] from generated images:

$$\{\boldsymbol{\theta}_d, \boldsymbol{\beta}_d, \mathbf{j}_d, \mathbf{R}_d^h, \mathbf{s}_d^h, x_d, y_d\} = \mathbf{F}_{\text{human}}(I_d) \quad (1)$$

where $\boldsymbol{\theta}_d \in \mathbb{R}^{54 \times 3}$, $\boldsymbol{\beta}_d \in \mathbb{R}^{10}$, $\mathbf{j}_d \in \mathbb{R}^{65 \times 3}$ are predicted SMPL-X pose, shape, and joints (see Supp. Mat. for the increased number of joints), respectively. $\mathbf{R}_d^h \in \text{SO}(3)$ is global rotation, and \mathbf{s}_d^h, x_d, y_d are each scale, and x, y direction offsets representing the parameters of weak perspective camera $\Pi_{\gamma(d)}$ assigned to image I_d . $\gamma(\cdot)$ is the function that maps the index of the 2D HOI image to the index of the weak perspective camera from which it is rendered. We additionally define \mathbf{j}_{d^\dagger} as joints in world coordinate, obtained by transforming inferred joints \mathbf{j}_d with the camera rotation and translation of $\Pi_{\gamma(d)}$. As 3D human lies within the weak perspective camera framework, there remains depth ambiguity to solve.

Depth Optimization using Weak Auxiliary Cue. To address the depth ambiguity of 3D human obtained from reference image I_d , we leverage 2D HOI images generated in advance. Since the human poses in images are geometrically inconsistent, we select the largest subset inliers \mathcal{I} among them which are “semi-consistent” with the reference image by applying RANSAC [15] in terms of joint re-projection error. The joint triangulation is performed on two images, including the reference image. Note that we generate sufficiently large amount of 2D HOI images to ensure the existence of the inlier set. The inliers \mathcal{I} is used to estimate the human depth z_d by optimizing the joint re-projection loss defined below:

$$\mathcal{L}_{\text{re-projection}} = \frac{1}{|\mathcal{I}|} \sum_{I_\delta \in \mathcal{I}} \|\Pi_{\gamma(\delta)}(\mathbf{j}_{d^\dagger} + z_d \mathbf{f}_d) - \Pi_{\gamma(\delta)}(\mathbf{j}_{\delta^\dagger})\|^2 \quad (2)$$

where $\mathbf{f}_d \in \mathbb{R}^3$ is the normalized direction of the camera $\Pi_{\gamma(d)}$. Inspired by Jiang *et al.* [33], we leverage the human segmentation masks to reason about occlusion and promote a initial depth for human. We initialize SMPL-X [63] models along the camera direction and select the model with the best IoU between the rendered mask (regarding occlusion) and the segmented region to be an initial of optimization. This is useful when the optimal object position is in between two human body parts along the camera direction (*e.g.*, a motorcycle body between the legs while riding) as we use a collision loss term during optimization.

The collision loss is used to estimate fine-grained depth, as the collision cue can provide physical plausibility of 3D HOI samples. We adopt collision loss from COAP [53] where the human body is represented as signed distance field [59]. We set object vertices as query points to reduce collision between human and object, namely $\mathcal{L}_{\text{collision}}$. We optimize depth z_d to minimize the total loss:

$$\mathcal{L} = \mathcal{L}_{\text{re-projection}} + \lambda_{\text{collision}} \mathcal{L}_{\text{collision}} \quad (3)$$

Filtering. We filter out potentially low-quality 3D human samples in the cases of: (1) human rendering (regarding occlusion) and predicted human segmentation do not overlap much; (2) number of inliers during RANSAC [15] is small; (3) human and object collides significantly in 3D.

3.3 Learning Comprehensive Affordance

Our main objective is to learn ComA, a new affordance representation which encompasses orientational and spatial affordance, along with contact-based affordance. In order to learn ComA from the previously generated 3D HOI samples, we first apply Poisson disk sampling [50] to uniformly select 3D surface points from the object and uniformly sample mesh vertices from SMPL-X [63] human, constructing ordered point sets for each object and human.

Given the ordered point sets, the density function on the i -th object point and the j -th human point, which we denote as $\mathcal{P}_{ij}(\mathbf{p}, \mathbf{n})$, represents the joint probability of the relative position $\mathbf{p} \in \mathbb{R}^3$ and relative normal orientation $\mathbf{n} \in \mathbb{S}^2$ of the j -th human point with respect to i -th object point within the 3D HOI samples. We obtain \mathbf{p} and \mathbf{n} by canonicalizing the orientation of the human surface normal \mathbf{n}_j^h and relative position $\mathbf{p}^{o \rightarrow h}$, assuming the object surface normal \mathbf{n}_i^o is rotated to face $\hat{\mathbf{n}} = [0, 0, 1]^T$ (see Supp. Mat. for details on canonicalization). The distribution \mathcal{P}_{ij} sums up to 1 for all possible \mathbf{p} and \mathbf{n} :

$$\int_{\mathbb{R}^3} \int_{\mathbb{S}^2} \mathcal{P}_{ij}(\mathbf{p}, \mathbf{n}) d\mathbf{n} d\mathbf{p} = 1 \quad (4)$$

In practice, we set the domain of \mathbf{p} as voxelized grid and domain of \mathbf{n} as equispaced grid on \mathbb{S}^2 , obtained via Fibonacci spirals [17], and compute the discrete probabilities by aggregating Gaussian kernels calculated for \mathbf{n} and \mathbf{p} (we use geodesic metrics for \mathbf{n}). Under the probability distribution $\mathcal{P}_{ij}(\mathbf{p}, \mathbf{n})$, we define three types of functions $f(\mathbf{p}, \mathbf{n})$ capturing different aspects of affordances: (1) Contact-based Affordance, (2) Orientational Affordance, (3) Spatial Affordance. Given each function $f(\mathbf{p}, \mathbf{n})$, the pointwise affordance between i -th object point and j -th human point is determined by computing its expectation:

$$\mathbb{E}_{\mathbf{p}, \mathbf{n} \sim \mathcal{P}_{ij}} [f(\mathbf{p}, \mathbf{n})] \quad (5)$$

Contact-based Affordance. From ComA, we can infer the contact score for human-object point pairs regarding proximity and normal alignment as:

$$f_{\text{contact}}(\mathbf{p}, \mathbf{n}) = \left(\frac{1 - \mathbf{n} \cdot \hat{\mathbf{n}}}{2} \right) e^{-\|\mathbf{p}\|} \quad (6)$$

where $e^{-\|\mathbf{p}\|}$ term encourages high score when distance $\|\mathbf{p}\|$ is close. The $\frac{1-\mathbf{n} \cdot \hat{\mathbf{n}}}{2}$ term, motivated by the fact that physical interactions are conducted via exerted normal force between contact points, encourages high score when the object normal \mathbf{n}_i^o and human normal \mathbf{n}_j^h align in an antiparallel direction. The normal alignment term improves the precision of the contact map, especially when small body parts (*e.g.*, fingertips) interact with an object and allows robust affordance learning even when using noisy HOI datasets, compared to prior work that only uses proximity for contact terms [2, 23, 24, 94, 98] or normal for computing only penetration [20, 90], without considering the alignment of surface normals.

Orientational Affordance. We aim to capture the pattern and tendency of the orientation of body parts with respect to the object based on the concept

of Shannon entropy [73]. Intuitively, lower entropy means there exists more typical patterns of human body orientations during the interactions. To quantify the orientational tendency for ComA, we use negated normalized Shannon entropy [73] to enforce high value when the human surface normal shows consistent orientational tendency (low variance) with respect to the object surface normal:

$$f_{\text{orientation}}(\mathbf{n}) = 1 + \frac{\log \mathcal{P}_{ij}(\mathbf{n})}{\log n_b} \quad (7)$$

where n_b denotes the number of discretized bins of probability measure. Intuitively, the orientational affordance defined by $\mathbb{E}_{\mathbf{n} \sim \mathcal{P}_{ij}} [f_{\text{orientation}}(\mathbf{n})]$ has a low value when marginalized distribution $\mathcal{P}_{ij}(\mathbf{n})$ is close to the uniform distribution, meaning that there is no dominant orientational pattern. Note that Eq. 7 is agnostic to the proximity term \mathbf{p} , which allows us to model the nonphysical orientational effects even for far-distance points.

Spatial Affordance. Following CHORUS [21], we also consider the 3D spatial occupancy distribution of a human surface point with respect to the object point by simply counting the occurrence:

$$f_{\text{spatial}}(\mathbf{p}) = \mathbb{1}(\mathbf{x} - \mathbf{p}) \quad (8)$$

where $\mathbb{1}(\cdot) : \mathbb{R}^3 \rightarrow \{0, 1\}$ is a binary function that returns 1 if the argument is zero vector, else 0. Thus the spatial affordance $\mathbb{E}_{\mathbf{p} \sim \mathcal{P}_{ij}} [f_{\text{spatial}}(\mathbf{p})]$ outputs scalar occupancy for the spatial vector $\mathbf{x} \in \mathbb{R}^3$. In practice, we implement the function as voxel array, counting 1 to the voxel that contains \mathbf{p} . Note that f_{spatial} is agnostic to the normal direction \mathbf{n} via marginalization. Learning the occupancy informs us about macro positioning of the human relative to the given object.

4 Experiments

In this section, we conduct experiments to demonstrate the efficacy of our representation ComA, and the method for learning it. In Sec. 4.3, we illustrate how ComA extends beyond contact-based affordance to learn distributions for orientational tendency and spatial relation, by visualizing it across various object categories. In Sec. 4.4, we perform quantitative comparisons with other baselines and verify the efficacy of our Adaptive Mask Inpainting. Finally, in Sec. 4.5, we introduce an optimization framework achievable through the knowledge provided by ComA, demonstrating the potential of our new affordance representation.

4.1 Datasets

3D Object Datasets. We obtain 3D object meshes from various 3D object datasets for qualitative evaluation. Specifically, We utilize 20 from BEHAVE [2], 10 from InterCap [30], 2 from ShapeNet [10], and 8 from SAPIEN [87] to learn ComA individually. As textures were not available in InterCap [30], making the diffusion model difficult to inpaint, we use TEXTure [68] to generate textures on the meshes using text prompt automatically generated from ChatGPT [60].

For quantitative evaluation, we utilize BEHAVE [2] as a ground truth dataset to verify the quality of generated 3D HOI samples, assuming that interactions captured in lab-controlled environments exhibit the upper-bound quality of HOI. BEHAVE [2] contains video sequences of HOI scenarios for 20 object categories. Each video frame is annotated with a single 3D human and object, where each human is given as SMPL-H [46, 70] format and object as rotation and translation applied to the canonical mesh. We preprocess the dataset by transferring the SMPL-H annotations into SMPL-X format following Choutas *et al.* [63]. We use the 3D human-object pair in each frame as a sample, compute vertex-wise contact scores in the same way as our method, and use them as ground truth. **Internet Search Datasets.** We also collect 60 free 3D object meshes from Internet source, SketchFab [75] and learn ComA.

4.2 Baselines and Metric

As previous studies focused on contact rather than learning the overall aspect of affordances, we use contact-based affordance derived from ComA to compare with them. However, most of them output 3D contact of a single HOI situation by inferring affordance from a given HOI sample (mostly an image), while we produce an overall affordance distribution for the given 3D object. To address this, we treat the contact inferred by state-of-the-art methods, DECO [83] (for the human) and IAGNet [91] (for the object) from the HOI image as a single sample, and aggregate them across multiple images to construct a distribution for comparison against BEHAVE [2].

Specifically, we aggregate the results of DECO [83] and IAGNet [91] on (1) 2D HOI images we've generated beforehand and (2) BEHAVE [2] test images, as these represent the usage of the BEHAVE [2] object. The aggregated scores are then normalized to form a distribution (summing up to 1), considering only the relative probability of potential contact, as the original scale of the contact scores varies between methods. Subsequently, we compare similarity scores (SIM) [79] and mean absolute error (MAE) [85] between these distributions.

4.3 Qualitative Results

Contact for Various Objects. Similar to other studies, ComA can be derived into contact-based affordance as shown in Fig. 5. As our representation provides contact scores for all pairs of human-object vertices, we assign the contact score of each vertex as the maximum score among the pairs that include it. We can observe the conventional object usage patterns as heatmap, which fits with general human usages. Specifically, in case Fig. 5(b), where significant contact is absent, we can infer that the interaction does not frequently appear by contact.

Orientational Tendency for Various Objects. ComA also contains information on orientational tendencies during the HOI. The last column in Fig. 5 visualizes the negated entropy of the vertex normal direction of human relative to the object as a heatmap. Fig. 5(a), Fig. 5(b), and Fig. 5(c) exhibit relatively high orientational tendency, while the rest show lower. This aligns with the fact that

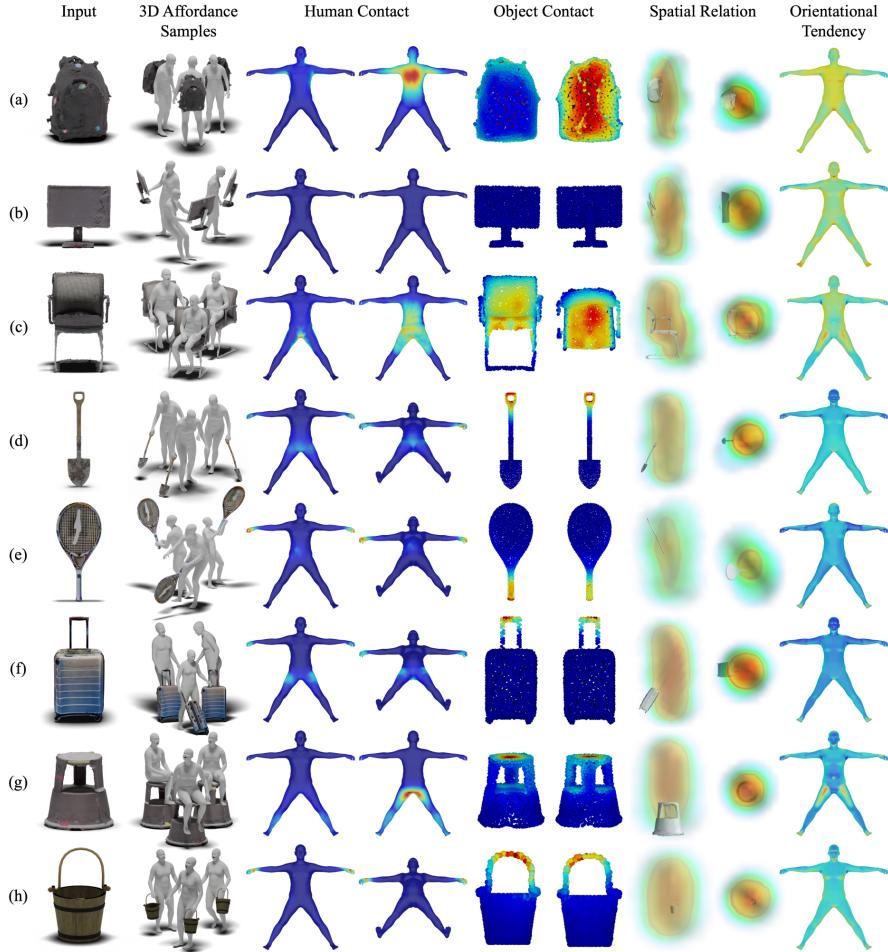


Fig. 5: Qualitative Results. ComA can model distributions of contact, orientation, and spatial relation exhibited during the interaction between humans and novel objects.

the human’s orientation is fixed by the attachment to the back as in Fig. 5(a), and by the restrictions present on the back and side of the object as in Fig. 5(c). In Fig. 5(b), there exists an orientational tendency without any attachments or restrictions, which arises due to long-distance interaction: viewing.

Spatial Relation for Various Objects. The relative position of the human during HOI can be also derived from ComA. Fig. 5 visualizes the distribution of the human’s full body positions relative to the object. Specifically, as shown in Fig. 5(d), Fig. 5(e), and Fig. 5(f), we find that the spatial relations for the object with multiple utilities and geometries (*e.g.*, in the case of shovel, handle for holding and blade for digging) align with general human usage, demonstrating the plausibility of the results. Especially, in the case of tennis racket shown in Fig. 5(e), we can observe a relatively weak tendency in spatial relation, as the racket can move almost freely relative to human during the interaction.

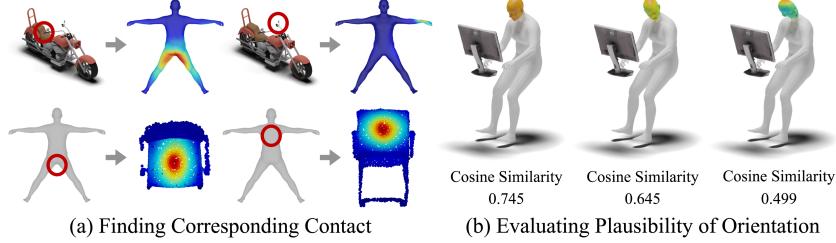


Fig. 6: Pointwise Affordance Knowledge. ComA can provide pointwise affordance knowledge such as corresponding contacts or plausible orientation of body parts.

Methods	SIM _{Human} (%)↑	SIM _{Object} (%)↑	MAE _{Human} (×100)↓	MAE _{Object} (×100) ↓
Ours	52.82	70.55	0.126	0.0288
IAGNet [91] on Our 2D HOI Images	-	66.34	-	0.0329
IAGNet [91] on BEHAVE Test Images	-	63.86	-	0.0353
DECO [83] on Our 2D HOI Images	21.66	-	0.246	-
DECO [83] on BEHAVE Test Images	17.39	-	0.259	-

Table 1: Quantitative Results on BEHAVE. We report SIM, MAE to evaluate the similarity of contact distribution (normalized contact scores) on BEHAVE sequences.

Pointwise Affordance Knowledge. Different from previous studies, ComA models the distribution of relative orientation and proximity for all human-object vertex pairs, allowing it to provide pointwise affordance knowledge such as corresponding contacts and plausible orientation direction of specific body parts. As shown in Fig. 6(a), we can infer how each human parts and object parts contact. For the case of motorcycle, we can see that saddle and handle each contacts with human’s hips and hands; while for the case of chair, seat and backrest each contacts with human’s hips and back. Also, we can utilize ComA to infer the probable direction of the head while using the monitor by computing the cosine similarity between the head direction and the vertex normal with highest orientational tendency. As illustrated in Fig. 6(b), we can see that the similarity maximizes when gazing towards the screen; which is a plausible head direction when interacting with the monitor.

Ablation on Adaptive Mask Inpainting. As shown in Fig. 4, without Adaptive Mask Inpainting, the original geometry of the object changes (left) or gets replaced by hallucinated objects (right), resulting in false affordances unfaithful to the original object. Consequently, the difference between the object in 2D and the original one existing in 3D causes the uplifted 3D HOI samples to fail to represent the right usage. This shows the importance of Adaptive Mask Inpainting as it prevents 2D inpainting diffusion model from generating false affordances.

4.4 Quantitative Results

Comparison with Baselines. We evaluate the similarity of contact distribution between BEHAVE [2] and each of DECO [83], IAGNet [91], and ours. We sample 2048 points from the 3D object mesh surface following IAGNet [91] and only consider the contact underneath the head for DECO [83], as DECO [83] uses SMPL [46] format, which only has vertex correspondence with SMPL-X [63] in

Inpainting Methods	$\text{RMSE}_{\text{Background}} \downarrow$	$\text{mIoU} \uparrow$	$\text{mIoU}_{\text{Occlusion Aware}} \uparrow$
Vanilla	42.23	0.535	0.631
Adaptive Mask	15.33	0.706	0.815

Table 2: Effects of Adaptive Mask Inpainting We evaluate the performance of background preservation in vanilla inpainting and our Adaptive Mask Inpainting.

non-head regions. As shown in Tab. 1, the contact distribution from IAGNet [91] and DECO [83] using our 2D HOI images outperforms the one modeled from BEHAVE test images. This means that our generated HOI images represent the real-world affordance as well as BEHAVE [2] test images, demonstrating the efficacy of leveraging pre-trained diffusion model for generating affordance. Additionally, the contact distribution from our method outperforms the one modeled using IAGNet [91] and DECO [83] with our generated HOI images. Since we extract 3D contacts from the same image sets, we argue that our pipeline is better than other models trained on annotations for uplifting 2D affordance to 3D.

Efficacy of Adaptive Mask Inpainting. To verify the effectiveness of our Adaptive Mask Inpainting, we compare the amount of background preservation for our generated 2D HOI images with the vanilla inpainting diffusion model. We sample at most 2 inpainted images with a single human detected, per each object renderings for all categories in BEHAVE [2]; resulting in 691 images for vanilla inpainting and 692 images for Adaptive Mask Inpainting. We compute 3 metrics for these images; (1) $\text{RMSE}_{\text{Background}}$: pixel error between images excluding the predicted human region; (2) mIoU: average IoU between object predictions in each inpainted image and original object image; (3) $\text{mIoU}_{\text{Occlusion Aware}}$: same metric as (2) but excluding the predicted human region during computation. To ensure fairness, we use open vocabulary segmentation model [37, 45] to predict detailed human/object segmentation. As shown in Tab. 2, we demonstrate that our adaptive mask algorithm better preserves the background, not only pixelwise ($\text{RMSE}_{\text{Background}}$), but also semantic-region-wise (mIoU, $\text{mIoU}_{\text{Occlusion Aware}}$) than the vanilla inpainting baseline. This means that the generated images from Adaptive Mask Inpainting are more likely to preserve the original object.

4.5 Application

As ComA is a newly proposed affordance representation, we demonstrate its potential by showcasing an application. One possible task is to reconstruct HOI by optimizing the human pose and position to interact with an object. As shown in Fig. 7(a), it is nearly impossible to achieve the task with traditional methods which relies on contact affordance only. In contrast, ComA provides both relative orientation and contact during the HOI, making it possible to generate plausible 3D HOI samples from T-shaped humans. For the objects with multiple modes of affordances, we apply mode seeking algorithm [12] on 3D HOI samples with their vertex normals and learn ComA separately; for example, swing in Fig. 7(a).

However, ComA learned on a specific object cannot be directly applied to other objects with different geometries. To address this, we transfer ComA through the process of (1) object canonicalization and (2) finding vertex cor-

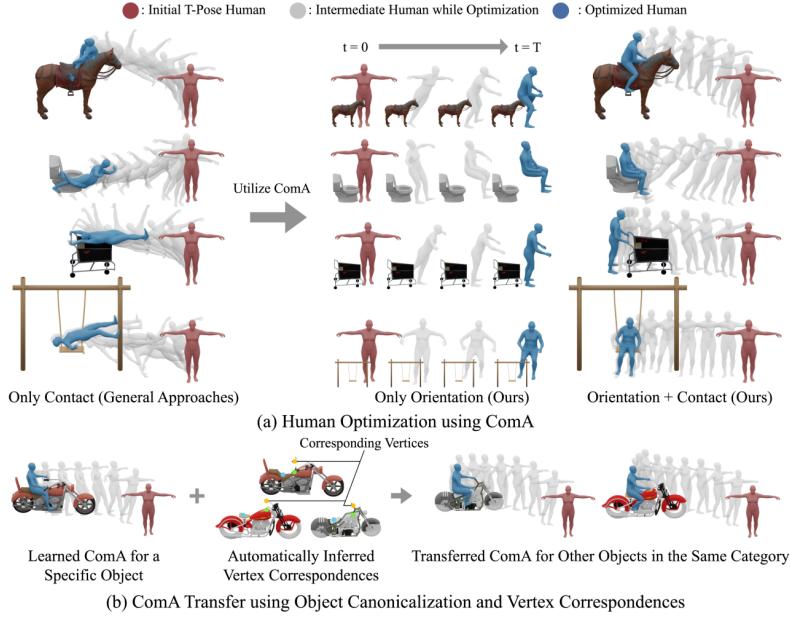


Fig. 7: Application. ComA can reconstruct HOI through optimization, which can be scaled to other objects in same category by transferring such knowledge between them.

respondences. Fig. 7(b) provides an example of transferring ComA learned from a specific motorcycle to other motorcycles with different geometries. We use objects from ShapeNet [10] which are already canonicalized and apply CPNet [47] to find vertex correspondences automatically for transferring the vertices with contact. As object canonicalization and vertex correspondence can each transfer macro orientation information and local contact information, we can transfer ComA and scale the human optimization task to other objects in the same category without any additional learning.

5 Conclusion

In this paper, we point out the importance of learning non-contact patterns in HOIs and propose a novel affordance representation called ComA which models both contact and non-contact patterns. To learn ComA, we utilize a pre-trained 2D diffusion model which already possesses prior knowledge of affordance. To leverage the affordance knowledge, we design a Rendering-Inpainting-Uplifting pipeline to generate 3D HOI samples and learn ComA from the samples. The results show that ComA effectively models the common usage of objects in terms of contact, orientation, and spatial relations. Within the pipeline, we propose *Adaptive Mask Inpainting*, which allows insertion of human without damaging the original image, where we evaluate its efficacy both qualitatively and quantitatively. We expect ComA to be used as a prior knowledge for various downstream tasks, where we demonstrate its potential through human optimization framework and the possibility of transfer to various objects in the same category.

Acknowledgements

This work was supported by Naver Webtoon. The work of SNU members was also supported by NRF grant funded by the Korean government (MSIT) (No. 2022R1A2C2092724 and No. RS-2023-00218601), and IITP grant funded by the Korean government (MSIT) [RS-2021-II211343, AI Graduate School Program (SNU)]. H. Joo is the corresponding author.

References

1. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: CVPR (2023)
2. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: CVPR (2022)
3. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: CVPR (2019)
4. Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: Contactgrasp: Functional multi-finger grasp synthesis from contact. In: IROS (2019)
5. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose. In: ECCV (2020)
6. Cai, M., Kitani, K.M., Sato, Y.: Understanding hand-object manipulation with grasp types and object attributes. In: RSS (2016)
7. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Real-time multi-person 2d pose estimation using part affinity fields. In: IEEE TPAMI (2019)
8. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV (2020)
9. Chai, L., Zhu, J.Y., Shechtman, E., Isola, P., Zhang, R.: Ensembling with deep generative views. In: CVPR (2021)
10. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 (2015)
11. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In: ICCV (2019)
12. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. In: IEEE TPAMI (2002)
13. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: CVPR (2021)
14. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: CVPR (2020)
15. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In: CACM (1981)
16. Gao, W., Tedrake, R.: kpam-sc: Generalizable manipulation planning using key-point affordance and shape completion. In: ICRA (2021)
17. Garg, M., Garg, P., Vohra, R.: Advanced fibonacci sequence with golden ratio. In: IJSER (2014)

18. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin (1979)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
20. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: CVPR (2021)
21. Han, S., Joo, H.: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In: ICCV (2023)
22. Hao, Z., Mallya, A., Belongie, S., Liu, M.Y.: Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In: ICCV (2021)
23. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: ICCV (2019)
24. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: CVPR (2021)
25. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? In: arXiv:2210.07574 (2022)
26. Hermans, T., Rehg, J.M., Bobick, A.: Affordance prediction via learned object attributes. In: ICRA (2011)
27. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: arXiv:2208.01626 (2022)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
29. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: CVPR (2022)
30. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view rgbd images. In: IJCV (2024)
31. Jahanian, A., Puig, X., Tian, Y., Isola, P.: Generative models as a data source for multiview representation learning. In: arXiv:2106.05258 (2021)
32. Jian, J., Liu, X., Li, M., Hu, R., Liu, J.: Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In: ICCV (2023)
33. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020)
34. Jiang, Y., Jiang, S., Sun, G., Su, Z., Guo, K., Wu, M., Yu, J., Xu, L.: Neuralholofusion: Neural volumetric rendering under human-object interactions. In: CVPR (2022)
35. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: arXiv:1412.6980 (2014)
37. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: ICCV (2023)
38. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: CVPR (2020)
39. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A.A., Singh, K.K.: Putting people in their place: Affordance-aware human insertion into scenes. In: CVPR (2023)

40. Lee, Y.J., Grauman, K.: Predicting important objects for egocentric video summarization. In: IJCV (2015)
41. Levine, S., Shah, D.: Learning robotic navigation from experience: principles, methods and recent results. In: Philos. Trans. R. Soc. B (2022)
42. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: CVPR (2023)
43. Li, L., Dai, A.: GenZI: Zero-shot 3D human-scene interaction generation. In: CVPR (2024)
44. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3d motion and forces of person-object interactions from monocular video. In: CVPR (2019)
45. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: arXiv:2303.05499 (2023)
46. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: Proc. ACM SIGGRAPH Asia (2015)
47. Lou, Y., You, Y., Li, C., Cheng, Z., Li, L., Ma, L., Wang, W., Lu, C.: Human correspondence consensus for 3d object semantic understanding. In: ECCV (2020)
48. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)
49. Mao, C., Cha, A., Gupta, A., Wang, H., Yang, J., Vondrick, C.: Generative interventions for causal learning. In: CVPR (2021)
50. McCool, M., Fiume, E.: Hierarchical poisson disk sampling distributions. In: Graphics Interface (1992)
51. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Finding an unsupervised image segmenter in each of your deep generative models. In: arXiv:2105.08127 (2021)
52. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)
53. Mihajlovic, M., Saito, S., Bansal, A., Zollhoefer, M., Tang, S.: COAP: Compositional articulated occupancy of people. In: CVPR (2022)
54. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: CVPR (2019)
55. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR (2023)
56. Monszpart, A., Guerrero, P., Ceylan, D., Yumer, E., Mitra, N.J.: imapper: interaction-guided scene mapping from monocular videos. In: ACM TOG (2019)
57. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: CVPRW (2022)
58. Nagarajan, T., Grauman, K.: Learning affordance landscapes for interaction exploration in 3d environments. In: NeurIPS (2020)
59. Oleynikova, H., Millane, A., Taylor, Z., Galceran, E., Nieto, J., Siegwart, R.: Signed distance fields: A natural representation for both mapping and planning. In: RSS Workshop (2016)
60. OpenAI: Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/> (2023)
61. OpenAI: Gpt-4v(ision) system card. <https://openai.com/research/gpt-4v-system-card> (2023)

62. Pan, X., Dai, B., Liu, Z., Loy, C.C., Luo, P.: Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In: arXiv:2011.00844 (2020)
63. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
64. Peebles, W., Zhu, J.Y., Zhang, R., Torralba, A., Efros, A.A., Shechtman, E.: Gan-supervised dense visual alignment. In: CVPR (2022)
65. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In: CVPR (2023)
66. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. In: arXiv:2204.06125 (2022)
67. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video. In: ECCV (2020)
68. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. In: Proc. ACM SIGGRAPH (2023)
69. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
70. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. In: Proc. ACM SIGGRAPH Asia (2017)
71. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. In: ACM TOG (2016)
72. SG_161222: Realistic vision v5.1. <https://civitai.com/models/4201?modelVersionId=130090> (2023)
73. Shannon, C.E.: A mathematical theory of communication. In: ACM SIGMOBILE (2001)
74. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: NeurIPS (2021)
75. Sketchfab: Sketchfab. <https://sketchfab.com/> (2023)
76. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
77. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: arXiv:2010.02502 (2020)
78. Sun, G., Chen, X., Chen, Y., Pang, A., Lin, P., Jiang, Y., Xu, L., Yu, J., Wang, J.: Neural free-viewpoint performance rendering under complex human-object interactions. In: ACM MM (2021)
79. Swain, M.J., Ballard, D.H.: Color indexing. In: IJCV (1991)
80. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: ECCV (2020)
81. Tanaka, F.H.K.d.S., Aranha, C.: Data augmentation using gans. In: arXiv:1904.09135 (2019)
82. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: arXiv:2302.07944 (2023)
83. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: DECO: Dense estimation of 3D human-scene contact in the wild. In: ICCV (2023)
84. Tritrong, N., Rewatbowornwong, P., Suwajanakorn, S.: Repurposing gans for one-shot semantic part segmentation. In: CVPR (2021)
85. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. In: Climate Research (2005)

86. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. In: NeurIPS (2023)
87. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., Yi, L., Chang, A.X., Guibas, L.J., Su, H.: SAPIEN: A simulated part-based interactive environment. In: CVPR (2020)
88. Xu, J., Zhu, S., Guo, H., Wu, S.: Automated labeling for robotic autonomous navigation through multi-sensory semi-supervised learning on big data. In: IEEE TBD (2021)
89. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023)
90. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: ICCV (2021)
91. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. In: ICCV (2023)
92. Yang, Y., Zhai, W., Luo, H., Cao, Y., Zha, Z.J.: Lemon: Learning 3d human-object interaction relation from 2d images. In: CVPR (2024)
93. Ye, Y., Li, X., Gupta, A., Mello, S.D., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: CVPR (2023)
94. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-aware object placement for visual environment reconstruction. In: CVPR (2022)
95. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: ECCV (2020)
96. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: Place: Proximity learning of articulation and contact in 3d environments. In: 3DV (2020)
97. Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. In: ECCV (2022)
98. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: CVPR (2020)
99. Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., Fidler, S.: Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In: ICLR (2021)
100. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: CVPR (2021)
101. Zhou, K., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Toch: Spatio-temporal object correspondence to hand for motion refinement. In: ECCV (2022)

Supplementary Materials for Beyond the Contact: Discovering Comprehensive Affordance for 3D Objects from Pre-trained 2D Diffusion Models

Hyeonwoo Kim^{1*}, Sookwan Han^{1*}, Patrick Kwon², and Hanbyul Joo¹

¹ Seoul National University

² Naver Webtoon AI

A Implementation Details

We provide further details on the implementations of our pipeline (Sec. 3.1~3.2) and the formulation of ComA (Sec. 3.3) in our main paper.

A.1 Rendering Object from Multi-Viewpoints

For *static* type objects, we install weak perspective cameras around the object with 45° azimuth intervals, creating 8 distinct views. For *dynamic* type objects, we install 4 weak perspective cameras with 90° azimuth intervals. The elevation is set constant within $[0^\circ, 30^\circ]$ range, where the values differ by category. For dynamic objects, we perturb the object with random rotations and translations. Specifically, we uniformly sample the rotation in the form of euler angle, where yaw, pitch, and roll are uniformly sampled from a predefined range. Random translations are sampled in a similar way, where each component of 3D displacement is sampled from a predefined range. Note that we set weak perspective camera scale and additional z -direction displacement of camera as hyperparameters. We repeat the rendering procedures 10 times with different perturbations, resulting in 40 distinct views.

A.2 Inpainting Mask Selection

Via thorough experiments, while the inpainting pipeline is quite robust to initial masks, we found that initial mask selection is beneficial for avoiding failure cases as shown in Fig. S.1. Specifically, we build strategies for selecting appropriate positions and sizes of masks to prevent generating: (1) hallucinated objects, which typically occurs when initial mask do not overlap with the rendered object; (2) ambiguous interactions, when inpainting masks are too small compared to the object, generating humans that ignore the relative scale with respect to the object.

*Indicates equal contribution

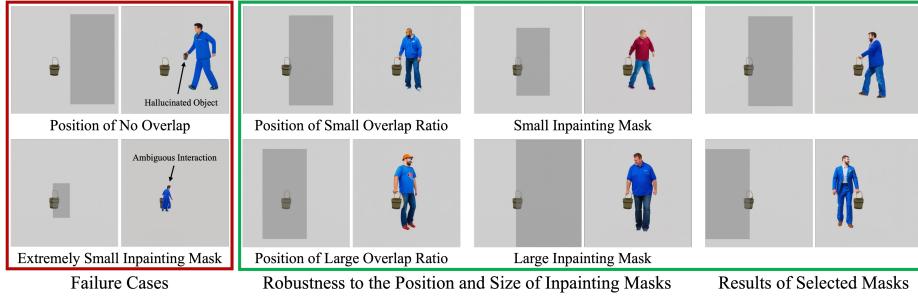


Fig. S.1: Experiments for Mask Selection. While our *Adaptive Mask Inpainting* method is quite robust to initial masks as shown in the right (green box), our inpainting mask selection strategy helps avoid generating failure cases shown in the left (red box).



Text Prompt: 1 person rides the surfboard **Text Prompt:** 1 person carries the surfboard

Fig. S.2: Diversity of 3D HOI Samples and Interaction Type. We initially create multiple HOI prompts and inpainting masks, which allows us to generate diverse 3D HOI samples with different interaction types by selecting the prompt and masks.

The strategy starts by rendering the inpainting masks while rendering the object, using the same camera parameters. For each camera, we place an upright window perpendicular to the xy plane in the world coordinate, also perpendicular to the xy projection of the camera's front vector. For each mask, the center of the intersection with $z = 0$ plane lies within the xy projection of the 3D object. Note that the strides of the upright windows with respect to x, y direction are given as hyperparameters, along with the height and width of the window. We render these upright windows using assigned cameras to obtain 2D rectangular masks, consequently used as initial inpainting masks that occlude the original object. To reduce the number of unnecessary masks (*e.g.*, masks that do not cover the object but mostly the background), we only retain the masks if the Intersection over Union (IoU) between the mask and the original object lies within the predefined range, also given as hyperparameters.

A.3 Prompt Generation

We design a generalizable pipeline to generate human-object interaction prompts even when the category of the object is unknown. We utilize GPT4v [61], where we input the rendered object image and the following query template:

Generate at most 3 simple subject-verb-object prompt where subject's word is exactly '1 person' and object's image is given. You should use

diverse and general word but no pronoun for subject. Generated prompt must align with common sense. Verb must be simple as possible, and should depict physical interaction between subject and object. Also, only the interaction with given object is allowed, and no other objects should be introduced in the prompt.

For 3D objects of which category is already known (generally for objects obtained from SketchFab [75]), we use ChatGPT [60] to generate prompts for human-object interaction using the following query template:

Generate at most 3 simple subject-verb-object prompt where subject’s word is exactly ‘1 person’ and object’s word is exactly ‘{category}’. You should use diverse and general word but no pronoun for subject. Generated prompt align with common sense. Verb must be simple as possible, and should depict physical interaction between subject and object. Also, only the interaction with given object is allowed, and no other objects should be introduced in the prompt.

We do not augment the prompt except “*full body*” at the end, where we empirically find this augmentation useful when generating the whole human body instead of a zoom-in shot of body parts (*e.g.*, face, hand). The generated prompts usually describe different types of HOI, which allows our pipeline to generate diverse 3D HOI samples by altering the input prompt or varying inpainting masks in multiview renderings, as shown in Fig. S.2.

A.4 Adaptive Mask Inpainting

The full pipeline of *Adaptive Mask Inpainting* algorithm is described in Algorithm. S.1. We use a publicly available inpainting diffusion model (RealisticVision [69, 72]) in our implementation, although we note that our adaptive mask algorithm can be applied to any inpainting diffusion model. We use classifier-free guidance scale of 11.0, and apply DDIM [77] scheduler of $T = 50$ timesteps with denoising strength set between $0.9 \sim 1.0$.

As the sequence of denoised image latents $\{x_t\}_{t=T}^0$ progresses (*i.e.*, $t : T \rightarrow 0$), the quality of the predicted denoised image \hat{x}_0 improves over the progress of timestep, thus the low-level structure of the target prompt (*i.e.*, human) becomes more apparent (as shown in Fig. 4). This allows us to ground the inpainting region for the next timestep (m_{t-1}) around that low-level structure by predicting the region using off-the-shelf segmentation model [38]. Note that we use the initial inpainting mask (m_{default}) if the structure is not detected. We dilate the predicted human mask to tolerate the imperfectness of the generated structure

Algorithm S.1 Adaptive Mask Inpainting

Latent Diffusion Model: ϵ_θ
 Latent VAE Decoder: \mathbf{D}
 Segmentation Model: \mathbf{S}
 DDIMSchedule: $\{\alpha_t\}_{t=1}^T$
 Dilation Schedule: $\{n_t\}_{t=1}^T$, Dilation Kernel: k
 Inputs: Prompt (c), Initial Mask (m_{default}), Image (I_{orig})
 Initialize Noise Latent: $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 Initialize Adaptive Mask: $m_T \leftarrow m_{\text{default}}$
for $t=T, \dots, 1$ **do**
 $\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t; c, m_t, I_{\text{orig}}, t))$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$
 if $t \in \text{ProvokeSchedule}$ **then**
 $s = \mathbf{S}(\mathbf{D}(\hat{x}_0))$
 if $s \neq \emptyset$ **then**
 $m_{t-1} = \text{Dilate}(s; n_t, k)$
 else
 $m_{t-1} = m_{\text{default}}$
 end if
 else
 $m_{t-1} = m_t$
 end if
 $x_{t-1} = \text{DDIMStep}(x_t, \hat{x}_0, t)$
end for
return $\mathbf{D}(x_0)$

during early steps, using 3×3 kernel k with n_t times repeat:

$$k = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad n_t = \begin{cases} 20 & 50 \geq t > 45 \\ 10 & 45 \geq t > 40 \\ 5 & 40 \geq t > 35 \\ 4 & 35 \geq t > 30 \\ 3 & 30 \geq t > 25 \\ 2 & 25 \geq t > 20 \\ 1 & 20 \geq t > 15 \\ 0 & 15 \geq t > 0 \end{cases} \quad (\text{S.1})$$

We also employ ‘‘Provoke Schedule’’ (refer to Algorithm. S.1) for faster generation speed. The provoke scheduler determines whether to skip the mask adaptation step ($t \in \text{ProvokeSchedule}$) or not ($t \notin \text{ProvokeSchedule}$) during timestep t . Specifically, we use the following schedule:

$$\text{ProvokeSchedule} = \{t \mid (40 \geq t \geq 2 \text{ and } t \text{ is even}) \text{ or } t = 45\} \quad (\text{S.2})$$

A.5 Lifting 2D Affordance to 3D

Number of Joints Used. We use Hand4Whole [57] to predict 3D humans from images ($\mathbf{F}_{\text{human}}$ in Eq. 1). The model returns 1 global rotation (for pelvis joint) and 54 human joint rotations following the SMPL-X [63] format; which consists of 21 body joints, 15 + 15 hand joints, 1 jaw joint, and 1 + 1 eye joints. We add 21 OpenPose [7] joints (“nose”, “right eye”, “left eye”, “right ear”, “left ear”, “left big toe”, “left small toe”, “left heel”, “right big toe”, “right small toe”, “right heel”, “left thumb”, “left index”, “left middle”, “left ring”, “left pinky”, “right thumb”, “right index”, “right middle”, “right ring”, “right pinky”) and exclude 11 original joints (“spine1”, “spine2”, “spine3”, “left foot”, “right foot”, “left collar”, “right collar”, “head”, “jaw”, “left eye smplhf”, “right eye smplhf”), resulting in $1 + 54 + 21 - 11 = 65$ joints.

Finding Inlier Set. We utilize a semi-consistent largest inlier set obtained from the generated 2D HOI image set $\{I_d\}_{d=1}^D$ to uplift the given generated 2D HOI image I_{ref} to 3D. To find the inlier set, we first choose target image set $\mathcal{I}_{\text{target}}$ from $\{I_d\}_{d=1}^D$, consisting of images which is generated from different views and shows high consistency with reference image I_{ref} . Specifically, we first triangulate human joints for every pairs of $\{I_{\text{ref}}, I_{\text{other}}\}$, where $I_{\text{other}} \in \{I_d\}_{d=1}^D$ is the image generated from different view with I_{ref} . We choose the best $N_{\text{triangulation}}$ images which the triangulated 3D human joints shows less re-projection error on reference image than threshold $\tau_{\text{triangulation}}$, constructing $\mathcal{I}_{\text{target}}$.

For every image I_{target} in the target image set $\mathcal{I}_{\text{target}}$ and the corresponding 3D human joints obtained via triangulation of I_{ref} and I_{target} , we find the number of inliers images which shows less re-projection error than τ_{ransac} , denoted as n_{target} . We use the inlier set with maximum n_{target} samples for the further depth optimization. In practice, we set $\tau_{\text{triangulation}} = 100$, $\tau_{\text{ransac}} = 100$, $N_{\text{triangulation}} = 400$, and use mean squared joint error on pixel space for all re-projection error.

Initializing Depth. We initialize 7 human candidates equispaced along the orthographic ray, where we place 4th candidate (center candidate) to the position that minimizes the average distance between the pelvis joint and all object vertices. The distance between human candidates is proportional to the width of the human along the orthographic camera ray, where we set the multiplier as 0.3. We initialize the depth using the human candidate with maximum IoU between the rendered human mask (regarding occlusion) and the predicted human segmentation mask.

Depth Optimization Settings. For depth optimization, we set $\lambda_{\text{collision}} = 400$, and use Adam [36] optimizer with learning rate 1×10^{-2} for 200 iteration to optimize \mathcal{L} .

Filtering. We filter out the 3D human samples if (1) the IoU between the human rendering and predicted human segmentation is below 0.3 or over 0.8, (2) number of inliers after RANSAC [15] is below τ_{inlier} (which varies from $1 \sim 50$, based on the given 3D object), or (3) the intersection volume over human volume is higher than 0.01.

A.6 Learning Comprehensive Affordance

Canonicalization from $\mathbf{n}_j^h, \mathbf{p}^{o \rightarrow h}$ to \mathbf{n}, \mathbf{p} . We address 2 types of 3D object (including 3D human): (1) the object is assumed rigid, meaning that current object can be obtained via applying rigid transformation $\mathbf{T}^{\text{original} \rightarrow \text{current}} \in \text{SE}(3)$ on the original object; or (2) the object is non-rigid (*e.g.*, 3D human), meaning that no original object exist, and also the rigid transformation. We provide the canonicalization procedure that addresses both cases.

Given the human surface normal \mathbf{n}_j^h and relative position $\mathbf{p}^{o \rightarrow h}$, we rotate them the same amount when rotating the object surface normal \mathbf{n}_i^o to face $\hat{\mathbf{n}} = [0, 0, 1]^T$. Specifically, we canonicalize the human normal \mathbf{n}_j^h to \mathbf{n} following:

$$\mathbf{n} = (\mathbf{n}_i^o \cdot \hat{\mathbf{n}})\mathbf{n}_j^h + (\mathbf{n}_j^h \cdot \mathbf{n}_i^o)\hat{\mathbf{n}} - (\mathbf{n}_j^h \cdot \hat{\mathbf{n}})\mathbf{n}_i^o + \left[\frac{\mathbf{n}_j^h \cdot (\mathbf{n}_i^o \times \hat{\mathbf{n}})}{1 + \mathbf{n}_i^o \cdot \hat{\mathbf{n}}} \right] (\mathbf{n}_i^o \times \hat{\mathbf{n}}) \quad (\text{S.3})$$

and similarly, we canonicalize the relative position $\mathbf{p}^{o \rightarrow h}$ to \mathbf{p} following:

$$\mathbf{p} = (\mathbf{n}_i^o \cdot \hat{\mathbf{n}})\mathbf{p}^{o \rightarrow h} + (\mathbf{p}^{o \rightarrow h} \cdot \mathbf{n}_i^o)\hat{\mathbf{n}} - (\mathbf{p}^{o \rightarrow h} \cdot \hat{\mathbf{n}})\mathbf{n}_i^o + \left[\frac{\mathbf{p}^{o \rightarrow h} \cdot (\mathbf{n}_i^o \times \hat{\mathbf{n}})}{1 + \mathbf{n}_i^o \cdot \hat{\mathbf{n}}} \right] (\mathbf{n}_i^o \times \hat{\mathbf{n}}) \quad (\text{S.4})$$

The canonicalization procedure described in Eq. S.3 and Eq. S.4 preserves the length of the vector ($\|\mathbf{n}\| = \|\mathbf{n}_j^h\|$, $\|\mathbf{p}^{o \rightarrow h}\| = \|\mathbf{p}\|$) and preserves the orientation with respect to the object normal ($\mathbf{n}_i^o \cdot \mathbf{n}_j^h = \hat{\mathbf{n}} \cdot \mathbf{n}$, $\mathbf{n}_i^o \cdot \mathbf{p}^{o \rightarrow h} = \hat{\mathbf{n}} \cdot \mathbf{p}$). Also, the procedure above describes the movement of human normal \mathbf{n}_j^h and relative position $\mathbf{p}^{o \rightarrow h}$ following the object normal, when the object normal is taking the “shortest path” to $\hat{\mathbf{n}}$ along the sphere surface \mathbb{S}^2 . Note that for the case of rigid object, we transform the current object (and corresponding human) back to the original state using $(\mathbf{T}^{\text{original} \rightarrow \text{current}})^{-1}$ before applying Eq. S.3 and Eq. S.4. For non-rigid objects (*e.g.*, human mesh), we directly apply Eq. S.3 and Eq. S.4.

Additional Details. We sample $750 \sim 2048$ points from human mesh and object mesh using Poisson disk sampling [50]. For human mesh, we find the closest vertex of SMPL-X [63] and save the vertex indices as the human mesh is not rigid and geometry may alter when the pose differs. We set the domain of \mathbf{p} as $30 \times 30 \times 30$ voxelgrid with each voxel the size of 0.04, and the domain of \mathbf{n} as an equispaced spherical grid with 250 points, where the domain points are obtained via Fibonacci spirals [17]. To save memory during quantitative evaluation (which only compares contact scores with previous approaches), we only accumulate $e^{-\|\mathbf{p}\|}$ instead of using full voxelgrid since f_{contact} from Eq. 6

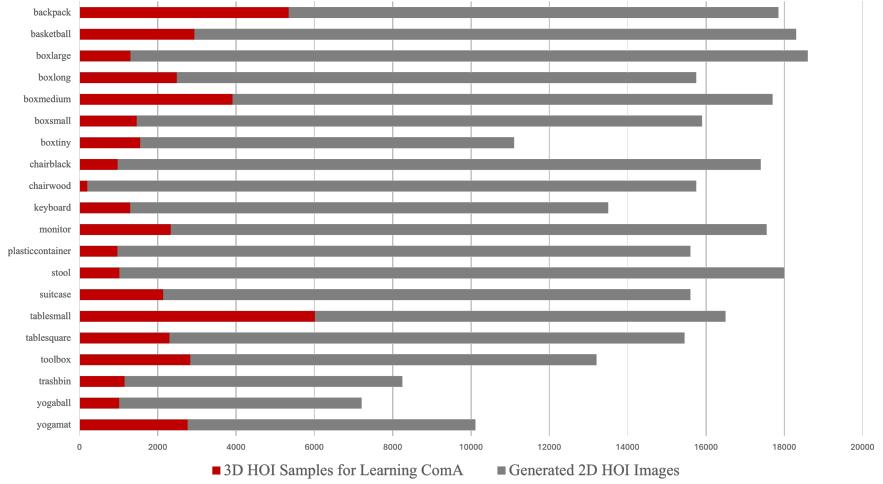


Fig. S.3: Statistics on Generated Samples. We report the number of generated 2D HOI images (gray) and 3D HOI samples (red) for BEHAVE objects, which we use in both qualitative and quantitative evaluation.

only requires relative distance $\|\mathbf{p}\|$ to compute. We fit the Gaussian kernel with $\sigma = 0.2$ for the domain of \mathbf{n} , and $\sigma = 0$ (quant) / $\sigma = 0.1$ (qual) for the domain of \mathbf{p} . Note that the Gaussian kernel for \mathbf{n} is computed using geodesic metrics, where we assume the radius of spherical grid as 1. Finally, we set $n_b = 10^6$ when computing $f_{\text{orientation}}$ from Eq. 7.

B Additional Details on Experiments

Preprocessing Intercap [30]. Since Intercap [30] did not provide texture for the 3D objects at the time of submission, we generated the texture for the objects using TEXTure [68] where the stylization prompts are generated via ChatGPT [60] using the following query:

Give a simple appearance description of an object of given categories as a form of “a {category}, {appearance description}”.

Method for Aggregating Contact Maps. When aggregating N samples to compute \mathcal{P}_{ij} 's for all pairs of i -th object point and j -th human point, we also count the number of times when $\|\mathbf{p}\| < d_{\text{thres}}$, which we denote as N_{sig}^{ij} . To create a holistic contact map, we aggregate the contact values derived from \mathcal{P}_{ij} only if $N_{\text{sig}}^{ij}/N > \tau_{\text{sig}}$, assuming such ij pairs show significant contact. When aggregating the contact values, we simply choose the maximum value between ij pairs. We set $d_{\text{thres}} = 0.1$, $\tau_{\text{sig}} = 0.05$ in our implementation.

Statistics of 2D HOI Images and 3D HOI Samples We report statistics on the generated 2D HOI images and 3D HOI samples. Fig. S.3 presents statistics

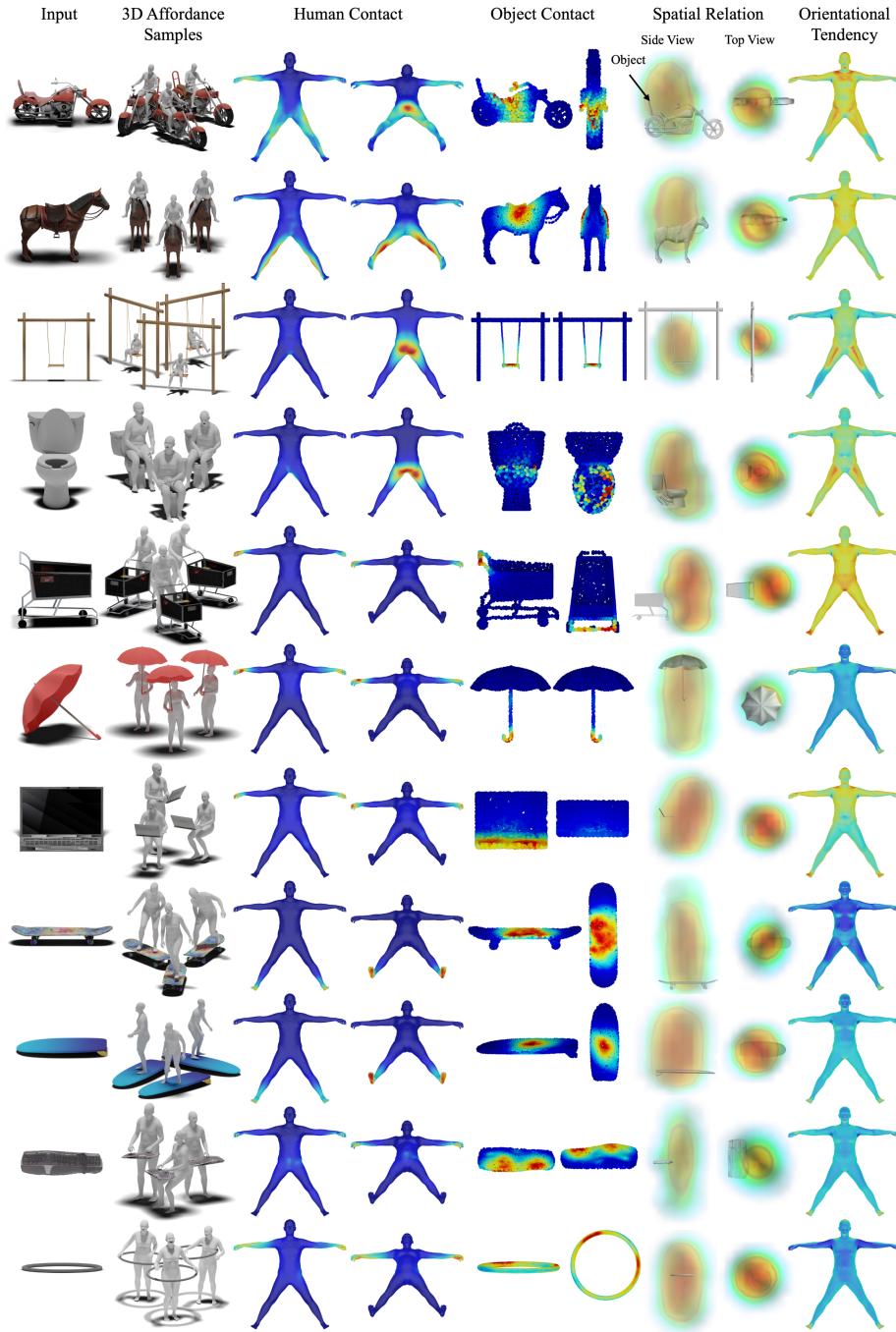


Fig. S.4: Additional Qualitative Results. Our method can be applied to various 3D objects obtained from diverse sources.

on the BEHAVE [2] dataset objects, which were used for both qualitative and quantitative evaluation. We generate images with varying mask regions, text prompts, and random seeds, resulting in a minimum of 7200, a maximum of 18600, and an average of 14965.15 images. After filtering out malicious data, we finally obtain a minimum of 198, a maximum of 6011, and an average of 2198.5 3D HOI samples for learning ComA. The overall acceptance ratio is 14.69%, meaning that we need approximately 7 images to obtain a single 3D HOI sample.

Additional Qualitative Results. We report additional qualitative results in Fig. S.4. As we utilize the affordance knowledge inherent in pre-trained 2D diffusion models, we are able to learn ComA for uncommon categories (*e.g.*, horse, swing, toilet, cart) which are not typically addressed in traditional 3D HOI datasets [2, 30].

Details on Application. We use SMPL-X [63] model to optimize global orientation, translation, body pose, and hand pose. For the body pose, we optimize pose embedding following VPoser [63] to leverage pose prior loss $\mathcal{L}_{\text{pprior}}$ and angle prior loss $\mathcal{L}_{\text{aprior}}$ which helps generating plausible pose. We define orientation loss $\mathcal{L}_{\text{orientation}}$ as average normalized cosine similarity between maximum probability direction (while object is fixed) obtained from ComA and human vertex normal for all human vertices. We also define contact loss $\mathcal{L}_{\text{contact}}$ as a chamfer distance between each contact points of human and object obtained from ComA. The total loss is defined as below:

$$\mathcal{L}_{\text{opt}} = \lambda_1 \mathcal{L}_{\text{pprior}} + \lambda_2 \mathcal{L}_{\text{aprior}} + \lambda_3 \mathcal{L}_{\text{orientation}} + \lambda_4 \mathcal{L}_{\text{contact}} \quad (\text{S.5})$$

In practice, we use $\lambda_1 = 1 \times 10^{-6}$, $\lambda_2 = 3.17 \times 10^4$, $\lambda_3 = 1 \times 10^{12}$, and $\lambda_4 = 2.6 \times 10^{11}$, and optimize \mathcal{L}_{opt} for 2000 steps using Adam [36].

C Limitations & Future Works

Spatial Bias in Inpainting Diffusion Models. Our method utilizes inpainting diffusion models [69, 72] to insert humans into object images; however, the diffusion model may possess spatial biases, which may alter the inpainting results as the properties of inpainting mask (*e.g.*, center location, aspect ratio, resolution) differs. For example, diffusion model may not be able to generate humans if the objects that usually interact with hands (*e.g.*, sports ball) are rendered on the bottom side of the image. Future research can further improve by mitigating the spatial bias in diffusion models, or the mask selection procedure to reduce the number of unnecessary generations.

Incorrect HOI Prompt Generation. During the prompt generation step, there is a chance for the vision-language model [61] to misidentify the object in the image, resulting in incorrect HOI prompts that describe implausible situations for the given object.

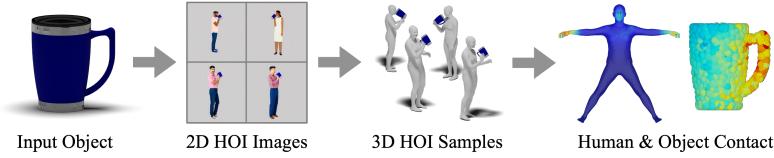


Fig. S.5: Results of ComA for Small Object. While our pipeline captures plausible affordance even for small objects (such as cup), the results show low granularity compared to big objects.

Limits and Potentials of Adaptive Mask Inpainting. We propose an *Adaptive Mask Inpainting* algorithm to preserve the original object during inpainting. However, the method depends on a segmentation model to adapt the inpainting mask, and the errors during segmentation may affect the inpainting results. For example, if the segmentation model predicts part of the object as human due to various reasons (*e.g.*, the texture of the object is similar to the texture of the generated human), the algorithm may not work well as the following inpainting mask also occludes the object. One potential approach for improvement is to use better segmentation models, such as Grounded SAM [37, 45].

While we use adaptive mask inpainting only for the human insertion task, the algorithm can be applied to any categories the segmentation model allows, opening possibilities such as *open-vocabulary object insertion into scene image*.

Bias due to Filtering. Employing heavy filtering at the end of the pipeline may result in bias. For example, filtering out humans with high collision may cause the remaining samples to “slide out” the object, especially if the object is complex and is highly likely to collide even when given the plausible posture (*e.g.*, motorcycle). One alternative is applying soft filtering (*i.e.*, applying confidence weights instead of removing images with hard thresholds).

Large Memory Consumption. ComA returns distributions for each pair of human and object points, which leads to large memory consumption when the resolution of human and object mesh is high; forcing us to downsample the human and object mesh. The limited resolution may cause the representation to lack details, especially when modeling interactions with dexterous objects. It is worth exploring the use of implicit 3D representation for human and object surfaces (*e.g.*, SDF [59], DMTet [74]), as such representations model the continuous surface as function.

Low Granularity for Small Objects. Although our pipeline captures affordance even for small objects (*e.g.*, cup interacting with hand and mouth, as shown in Fig. S.5), the lack of granularity compared to big objects is an existing challenge to solve.

Modelling Hand-Object Interactions. Diffusion models often struggle to produce high-quality images of hands. Future research could benefit from using diffusion models trained specifically on hand images to improve hand generation and employing close-view cameras focused on hands for better modeling.

Possible Improvements in ComA. We introduce ComA as a new representation for affordances and use it to deduce contact and non-contact information. Although our method for deriving contact is well-founded, there are opportuni-

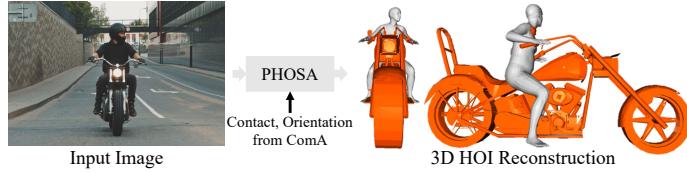


Fig. S.6: Single Image HOI Reconstruction for Any Object using ComA.
We can directly replace the contact heuristics in PHOSA [95] and apply additional orientation loss from ComA, which allows PHOSA [95] to scale to unseen objects.

ties for enhancement. Specifically, incorporating pressure modeling could extend ComA’s applicability to deformable objects. Additionally, the concept of orientational affordance needs refinement. The current approach effectively measures orientation preference using a negated entropy term, but it fails to identify the underlying reasons for this preference. For instance, in the case of chair, while human feet often orient towards the ground, attributing this tendency to the object itself overlooks the influence of gravity. A valuable future research direction would involve distinguishing and analyzing the reasons behind orientational tendencies.

Expanding to Non-Watertight Mesh. ComA can be easily extracted from non-rigid mesh (*e.g.*, human mesh), as long as the mesh provides a closed surface and surface normal can be defined. One possible future direction is to improve ComA to be easily extractable from any 3D surfaces (mesh or other surface representations), including non-watertight mesh, and sharp meshes where defining surface normal is non-trivial.

Evaluation Metrics. There’s potential to explore more complex metrics to accurately assess the effectiveness of our method, particularly when evaluating the volumetric quality of 3D humans and objects to support the use of 2D-to-3D conversion methods.

Potential Applications. Our new method and ComA provides multitudes of possible applications, as demonstrated in Sec. 4.5. We list potential downstream applications: (1) Large-scale 3D affordance dataset generation; (2) Single image HOI reconstruction for any 3D object (as shown in Fig. S.6); (3) Object recognition from 3D human posture (similar to Object Pop-up [65]); (4) Action recognition from human-object interaction sequence; (5) Application for robotics, especially for humanoids.