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 Kwon[2], and Hanbyul Joo[1]

[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 8 weak perspective cameras around the object with 45°azimuth intervals. For *dynamic* type objects, we install 4 weak perspective cameras with 90° azimuth intervals. The elevation is set constant within [0°, 30°] 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 ignoring the relative scale with respect to object.

The strategy starts by rendering the inpainting masks while rendering the object, using the same camera parameters. For each camera, we place an upright

[*] Indicates equal contribution

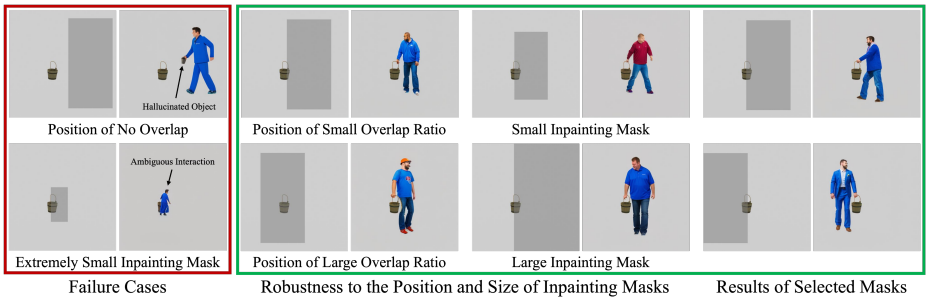

Fig. S.1: Experiments for Mask Selection. While our *Adaptive Mask Inpainting* method is quite robust to initial masks as shown in the right (green box), our inpainting mask selection strategy helps avoid generating failure cases shown in the left (red box).

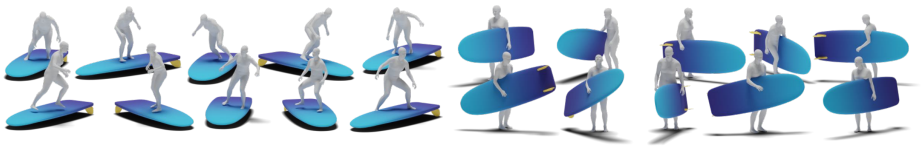

Fig. S.2: Diversity of 3D HOI Samples and Interaction Type. We initially create multiple HOI prompts and inpainting masks, which allows us to generate diverse 3D HOI samples with different interaction types by selecting the prompt and masks.

window perpendicular to the xy plane, 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 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 [14], 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 [21]), we use ChatGPT [13] 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 [18, 19]) 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 [22] 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.*, $t : T \to 0$), the quality of the predicted denoised image $\hat{x}_0$ improves over progress of the 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 segmentation region using off-the-shelf segmentation model [8]. 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

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=$T, ..., 1$ **do**
  $\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t; c, m_t, I_{\text{orig}}, t))$
  if $t \in$ ProvokeSchedule **then**
    $s = \mathbf{S}(\mathbf{D}(\hat{x}_0))$
    if $s \neq \varnothing$ **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 $\mathbf{D}(x_0)$

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

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

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 ($t \in$ ProvokeSchedule) or not ($t \notin$ ProvokeSchedule) during timestep t. Specifically, we use the following schedule:

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

A.5 Lifting 2D Affordance to 3D

Number of Joints Used. We use Hand4Whole [11] 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 [15] format; which consists of 21 body joints, $15 + 15$ hand joints, 1 jaw joint, and $1 + 1$ eye joints. We add 21 OpenPose [2] 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 $1 + 54 + 21 - 11 = 67$ 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 which shows maximum n_{target} 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 4^{th} 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 and the predicted human segmentation mask.

Optimization Settings. For optimization, we set $\lambda_{\text{collision}} = 400$, and use Adam [6] 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 [3] 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:

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

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

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

The canonicaliation 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 1000 points from human mesh and object mesh using Poisson Disk Sampling [10]. For human mesh, we find the closest vertex of SMPL-X [15] 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 [4]. 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 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

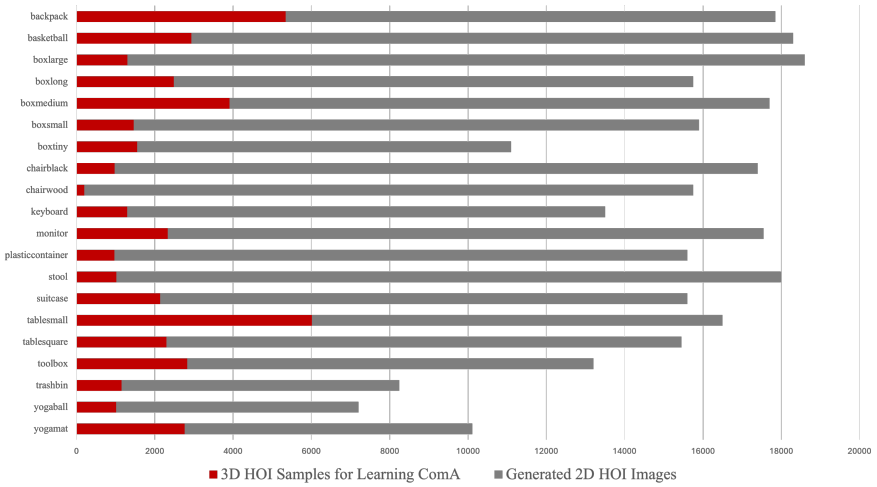

Fig. S.3: Statistics on Generated Samples. We report the number of generated 2D HOI images (gray) and 3D HOI samples (red) for BEHAVE objects, which we use in both qualitative and quantitative evaluation.

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 [5]. Since Intercap [5] does not provide texture for the 3D objects, we generate the texture for the objects using TEXTure [17] where the stylization prompts are generated via ChatGPT [13] 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 on the BEHAVE [1] dataset objects, which were used for both qualitative and quantitative evaluation. We generate images with varying mask regions, text

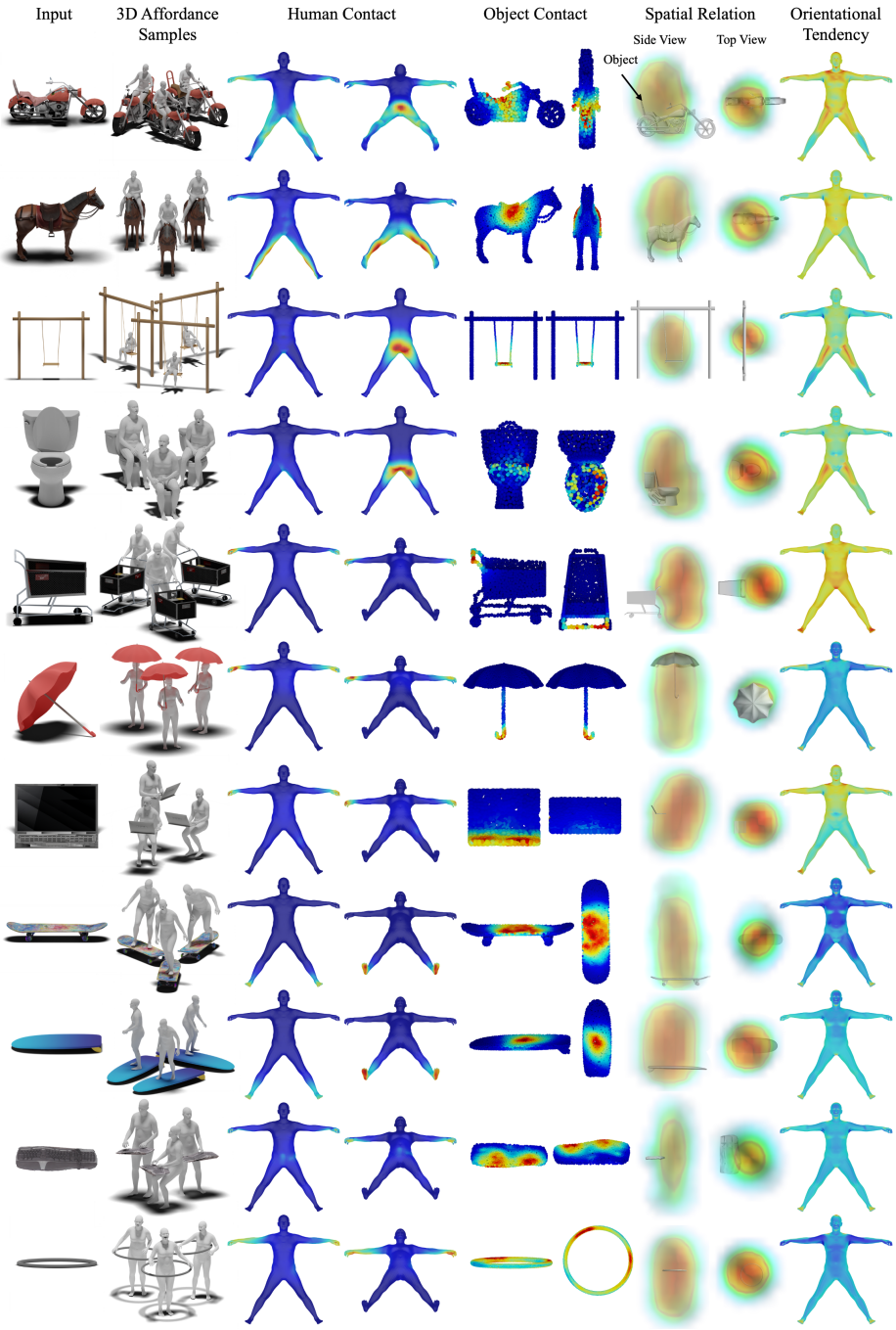

Fig. S.4: Additional Qualitative Results. Our method can be applied to various 3D objects obtained from diverse sources.

prompts, and seeds in multiview renderings, 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 [1, 5].

Details on Application. We use SMPL-X [15] model to optimize global orientation, translation, body pose, and hand pose. For the body pose, we optimize pose embedding following VPoser [15] 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:

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

In practice, we use $\lambda_1 = 1 \times 10^{-6}$, $\lambda_2 = 3.17 \times 10^4$, $\lambda_3 = 1 \times 10^{12}$, and $\lambda_4 = 2.6 \times 10^{11}$.

C Limitations & Future Works

Spatial Bias in Inpainting Diffusion Models. Our method utilizes inpainting diffusion models [18, 19] 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 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 [14] to misidentify the object in the image, resulting in incorrect HOI prompts that describe implausible situations for the given object.

Limits and Potentials of Adaptive Mask Inpainting. We propose an *Adaptive Mask Inpainting* algorithm to preserve the original object during inpainting.

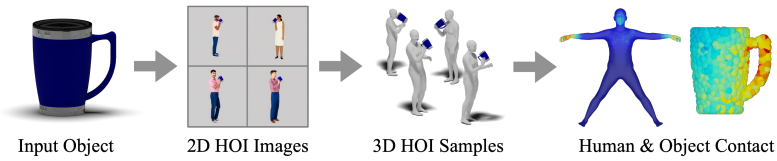

Fig. S.5: Results of ComA for Small Object. While our pipeline captures plausible affordance even for small objects (such as cup), the results show low granularity compared to big objects.

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 [7, 9].

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 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 [12], DMTet [20]), as such representations model the continuous surface as function.

Low Granularity for Small Objects. Although our ComA 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 information. Although our method for deriving contact is well-founded, there are opportunities for enhancement. Specifically, incorporating pressure modeling could extend ComA's applicability to deformable objects. Additionally, the concept of orientational af-

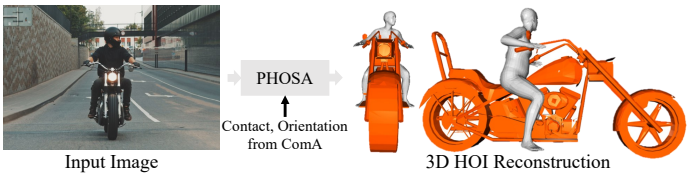

Fig. S.6: Single Image HOI Reconstruction for Any Object using ComA. We can directly replace the contact heuristics in PHOSA [23] and apply additional orientation loss from ComA, which allows PHOSA [23] to scale to unseen objects.

fordance 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, while chair 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 quantifying 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 [16]); (4) Action recognition from human-object interaction sequence; (5) Application for robotics, especially for humanoids.

References

1. 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)
2. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE TPAMI (2019)
3. 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)
4. Garg, M., Garg, P., Vohra, R.: Advanced fibonacci sequence with golden ratio. In: IJSER (2014)

5. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. In: IJCV (2024)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: arXiv:1412.6980 (2014)
7. 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)
8. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: CVPR (2020)
9. 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)
10. McCool, M., Fiume, E.: Hierarchical poisson disk sampling distributions. In: Graphics Interface (1992)
11. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: CVPRW (2022)
12. 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)
13. OpenAI: Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt/` (2023)
14. OpenAI: Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card` (2023)
15. 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)
16. 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)
17. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. In: Proc. ACM SIGGRAPH (2023)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
19. SG_161222: Realistic vision v5.1. `https://civitai.com/models/4201?modelVersionId=130090` (2023)
20. 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)
21. Sketchfab: Sketchfab. `https://sketchfab.com/` (2023)
22. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: arXiv:2010.02502 (2020)
23. 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)