InteractAnything: Zero-shot Human Object Interaction Synthesis via LLM Feedback and Object Affordance Parsing

Supplementary Material

A. Details of Our Method

A.1. LLM-guided Initialization Prompts

To better use the Large Language Model (LLM) to initialize the relative relations of humans and given objects, inspired by Chain-of-Thought [6], we design a set of fill-in-the-blank prompts for querying GPT, and below are the prompt templates.

```
System: Assume you are a human-object
interaction estimator. Given <HOI prompt>,
let's do it step by step.
```

Step 1: Body Part

```
User: You need first estimate body part;
LLM: One of <body_part>;
```

Step 2: Object State

```
User: Return the object state;
LLM: One of <obj_state>;
```

Step 3: RST initialization

```
User: Return relative direction,
orientation, and scale of object.
LLM:
One of <rel_dir>, <rel_ori>, <scale>.
```

The `body_part` includes semantic labels of SMPL-H like `hands`, `back`, and `feet`, and `obj_state` includes `dynamic` and `static` to represent if object rotation state can be optimized or not. Both `rel_dir` and `rel_ori` include `[up, bottom, left, right, front, back]` to indicate global position to human and local orientation to human. The `[large, middle, small, very small]` is chosen as the scale of the object initialization. Then we transfer these choices to pre-defined values to initialize the object rotation, translation, and scale.

A.2. Visualization of Object Affordance Parsing.

We visualize the contact probability for each 3D vertex of open-set objects, as shown in Figure S.1. The objects are downsampled to approximately 1000 points, and 2D probability functions are computed for each inpainting image. These contact scores are then projected back onto the 3D vertices, under the assumption that the expectation of all points is 1. The color gradient from blue to red represents the contact probability, with red indicating a higher likelihood of interaction.

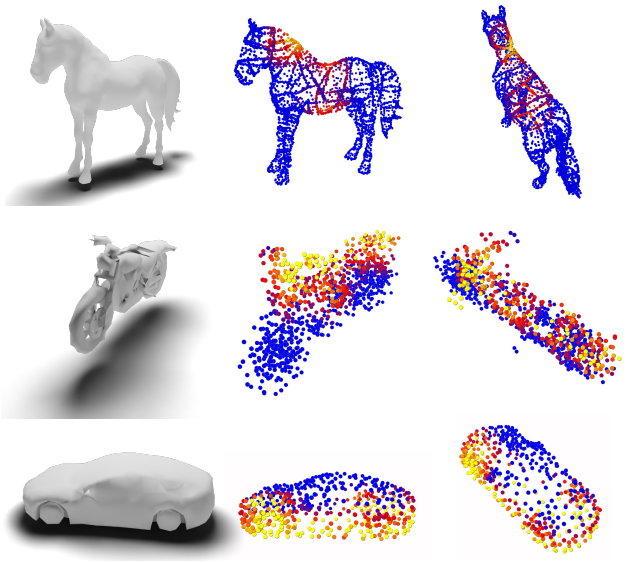

Figure S.1. **Visualization of the open-set object affordance parsing results.** The object meshes are listed in the left column, and the affordance score maps of "rides", "sits", and "lifts" interactions are visualized on downsampled objects in the right two columns. We normalize the score to [0, 1] and use red to represent contact probability.

B. Implementation Details

We re-implement parts of the baseline methods [3, 4] following third-party implementations [2], using 8000 sampling timesteps. For the DreamFusion* setting, we follow [7] and employ only MVDream [5] for diffusion guidance. During the optimization process, we project the object mesh from various views to ensure consistent object geometry and render results. Rendering images are directly used as comparison outputs for these methods. For HOI mesh comparisons, we use the pose estimation results and object mesh from the second stage of DreamHOI [7].

In our method, the object translation and rotation are initialized based on feedback from Large Language Models (LLMs), considering the human pose. We compute the 3D bounding box of the human mesh and apply an expansion factor of 1.2 to avoid collisions. During object parsing, to improve efficiency, we downsample the object points D to 1024 and compute contact scores. For human pose synthesis, we set the overall SDS optimization step to 10000 and use the last 6000 timesteps for the spatial constraint loss. During optimization, we adopt the intersection regularizer

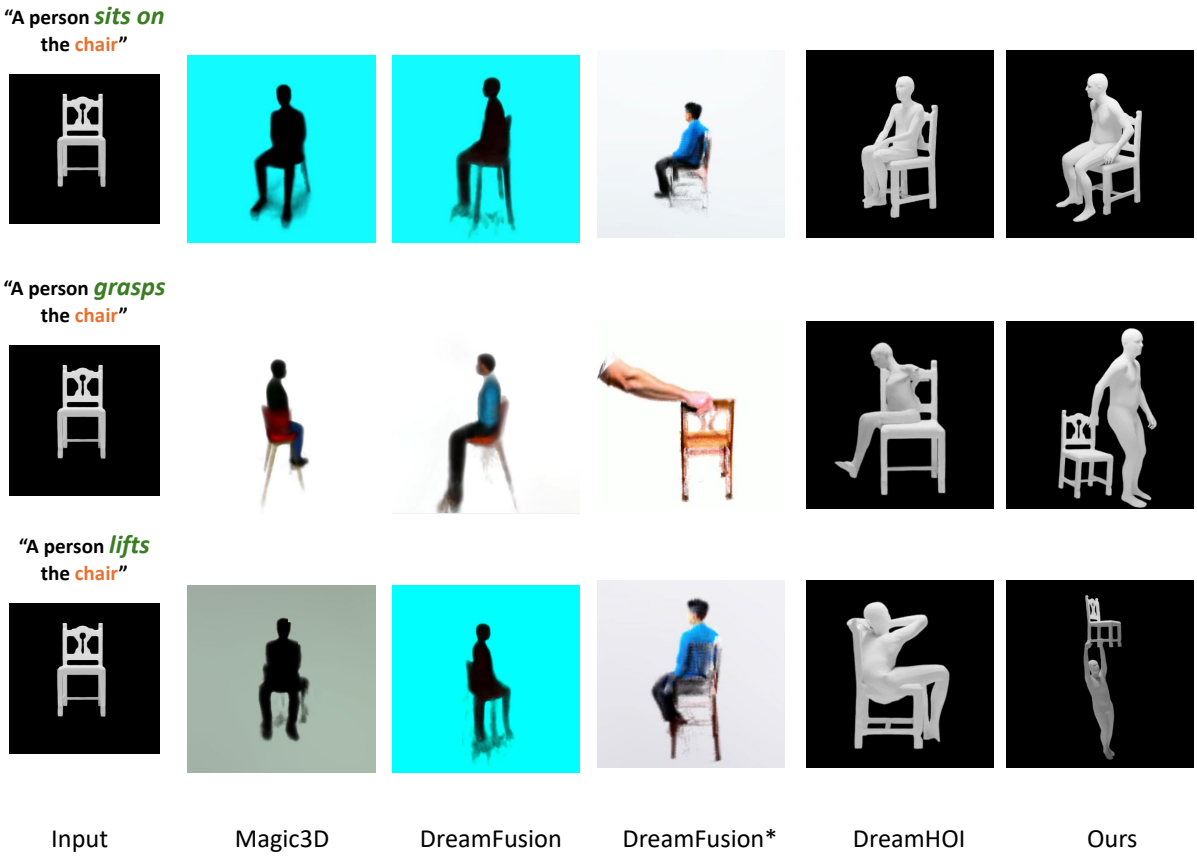

Figure S.2. **Qualitative comparison results on the same object with different interactions.**

from DreamHOI [7] on 2D object masks to discourage the model from generating body parts or other objects within the object mesh. For HOI optimization, the object rotation matrix is initialized using axis angles along the [x, y, z] axes. Object rotation gradients are then optimized along the x and z axes based on object state feedback. When the body parts feedback includes hands or palms, we apply a force closure loss on interacted meshes to synthesize grasping poses. The learning rate is set to 0.01, and the loss weights for ϕ_i, ϕ_n, ϕ_s, and ϕ_p are set to 10.0, 1.0, 5.0, and 10.0, respectively, by default.

C. Experiments

C.1. Ablation on Adaptive Mask Inpainting

In this section, we evaluate the impact of the adaptive mask inpainting component by comparing different masking strategies: using only the full-body mask (w/o BP), using only body-part masks (w/o FB), and our full adaptive mask inpainting approach. As shown in Figure S.3, the results illustrate the importance of each masking type in accurately parsing object affordances based on the provided text prompt. The ideal object affordance parsing should identify and extract reasonable contact regions that align with the described interaction, in this case, for "A person grasps the chair": (1) When dealing with w/o BP (Body-Part Mask Only) setting, using only the body-part mask yields localization ability of specific regions involved in the interaction, but it lacks the global context provided by the

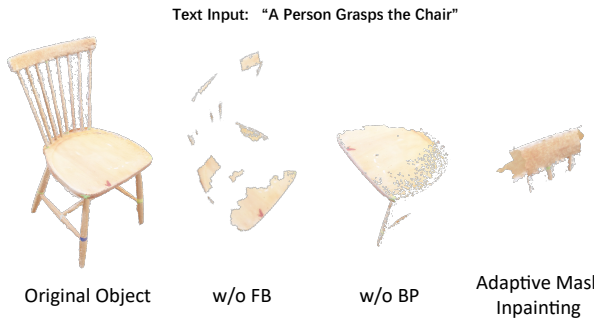

Figure S.3. **Ablation study on the different inpainting mask settings.** FB indicates Full-Body adaptive mask, and BP is Body-Part mask.

full-body mask. Consequently, this approach also generates incomplete affordance parsing, with the mesh displaying isolated and fragmented regions that do not capture the entirety of the intended grasp action. (2) When only the full-body mask is used (w/o BP), the output lacks precision, failing to identify specific body parts required for interaction, leading to incorrect regions in the object. This results in poorly defined contact points and misaligned geometry, as evidenced by the significant missing details on the chair. (3) For adaptive mask inpainting, by combining both full-body and body-part masks adaptively, our method achieves a balance between local detail and global context. This approach allows for precise affordance parsing that successfully identifies the regions of the chair relevant to the grasping action, resulting in a more complete and accurate extraction of the contact regions on the mesh.

These results underscore the importance of adaptive mask inpainting in our framework, enabling effective and accurate affordance parsing that aligns with human object interaction cues described in the text. The combined use of full-body and body-part masks ensures that both general and detailed aspects of the interaction are captured, leading to high-quality, interaction-specific object affordance.

C.2. More Qualitative Results

Qualitative Comparisons on Novel Objects. As shown in Figure S.4, our method demonstrates superior performance on novel open-set objects compared to the baseline methods: **Firstly**, our approach consistently synthesizes plausible human-object interactions with accurate spatial alignment and realistic poses, as illustrated in each row, representing diverse scenarios such as *"lifting a backpack," "cradling a baby doll," "riding a motorcycle,"* and *"hugging a humanoid robot"*. **Secondly**, our method could capture nuanced poses that accurately reflect the intended interaction described by the input text. Unlike baseline methods such as Magic3D [3] or DreamFusion [4], which struggle to generate coherent human-object configurations in cases like *"A person rides the motorcycle"* or *"A person hugs the humanoid robot,"* our approach successfully adapts to the object's unique geometry. This highlights the effectiveness of our method in handling previously unseen or complex object meshes. **Thirdly**, the contact regions between the human model and the object in our results are more accurate compared to other methods. For instance, in the *"A person lifts the car"* scenario, our method ensures proper alignment of hands and maintains a realistic lifting pose, while competing methods produce less convincing interactions or unrealistic distortions. The use of public object meshes and custom-designed objects (as seen in the bottom five rows of Figure S.4) demonstrates our ability to generalize effectively to novel object types. From common objects like a backpack (from the BEHAVE dataset [1]) to complex and unconventional objects like humanoid robots, our method delivers consistently high-quality results.

Overall, the comparisons highlight the robustness and versatility of our method in synthesizing high-quality human-object interactions, particularly in challenging novel open-set object scenarios. This demonstrates the efficacy of incorporating SMPL-based priors, optimized spatial-aware SDS loss, and effective multi-view alignment in our pipeline.

Qualitative performance on Novel Interactions. Figure S.2 showcases the qualitative comparison results across different methods for handling various interactions with the same object, specifically a chair. The interactions include *"sitting on the chair," "grasping the chair,"* and *"lifting the chair."* Our method consistently delivers more accurate and realistic human-object interactions compared to the baselines, as illustrated as follows: **(1)** For the interaction *"A person sits on the chair,"* our method produces a natural sitting pose with precise alignment between the human model and the chair. In contrast, Magic3D and DreamFusion fail to establish realistic contact or exhibit incorrect body proportions, while DreamFusion* and DreamHOI produce poses that are less aligned or lack contextual interaction fidelity. **(2)** For the interaction *"A person grasps the chair,"* our method effectively captures the arm and hand positioning required for the grasping action, maintaining the integrity of both the human and the chair geometries. Other methods either produce distorted hand positions, as seen in Magic3D and DreamFusion, or fail to properly depict the interaction context, as evidenced by the unrealistic grasping angles in DreamHOI. DreamFusion* shows a slight improvement in hand positioning but struggles to represent a convincing interaction. **(3)** For the interaction *"A person lifts the chair,"* our approach uniquely succeeds in synthesizing a physically plausible lifting pose, where the human model exhibits proper posture and contact with the chair. Other methods encounter significant challenges: Magic3D and DreamFusion generate implausible or incomplete lifting poses, while DreamFusion* and DreamHOI show issues with body distortion or fail to convey the lifting action convincingly.

Overall, our method show superior generalization and adaptability to varying human-object interactions with the same object, effectively overcoming the challenges faced by baseline methods. This highlights the robustness and accuracy of our approach in synthesizing diverse interactions while maintaining geometric and contextual consistency.

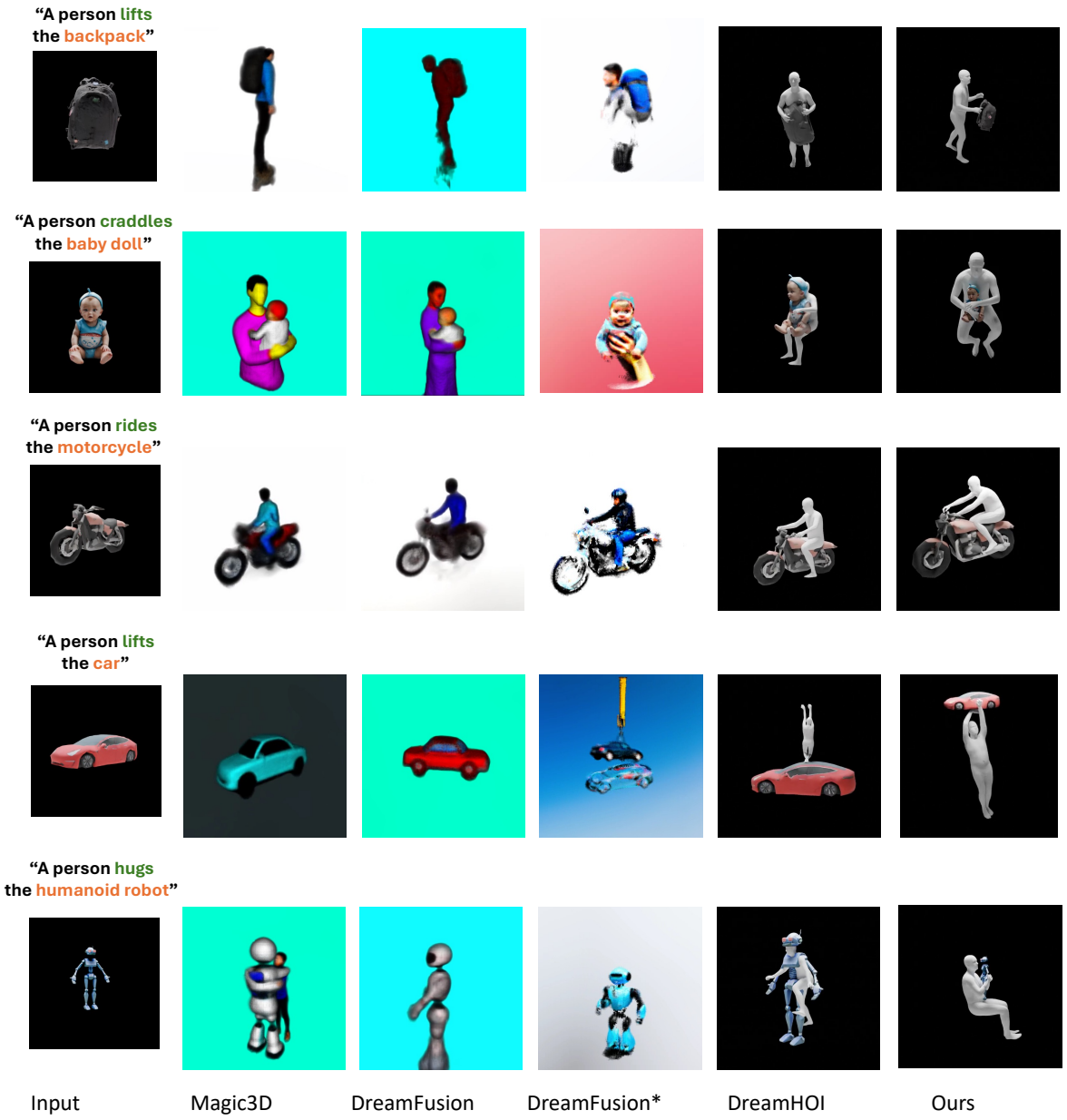

Figure S.4. **Qualitative comparison results with baselines on novel open-set objects.** $*$ indicates we re-implement this method by embedding object mesh into the diffusion process. We use novel object meshes generated by public models or designing websites in the bottom 5 lines, and the backpack mesh comes from the BEHAVE dataset [1].

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[2] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 1

[3] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[4] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 3

[5] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022. 1

[7] Thomas Hanwen Zhu, Ruining Li, and Tomas Jakab. Dreamhoi: Subject-driven generation of 3d human-object interactions with diffusion priors. *arXiv preprint arXiv:2409.08278*, 2024. 1, 2