

TeHOR: Text-Guided 3D Human and Object Reconstruction with Textures

Supplementary Material

In this supplementary material, we present additional technical details and more experimental results that could not be included in the main manuscript due to the lack of pages. The contents are summarized below:

- S1. Evaluation on semantic alignment
- S2. Comparison with HDM
- S3. Impact of contact estimation accuracy
- S4. Impact of Gaussian attributes optimization
- S5. Impact of Gaussians-to-mesh conversion
- S6. Details of text captioning
- S7. Implementation details
- S8. Limitations and future work
- S9. More qualitative results

S1. Evaluation on semantic alignment

In this section, we introduce additional evaluation to compare TeHOR with state-of-the-art reconstruction methods on semantic alignment between 3D reconstructions and text descriptions. Since direct comparison between 3D reconstructions and text is not feasible, we instead evaluate appearance-text alignment metrics on 2D renderings of the reconstructions, following the process shown in Fig. S1. To ensure a fair comparison, we unify the underlying 3D representation across all methods, since existing methods predominantly use mesh-based representations, whereas our framework is Gaussian-based. Specifically, we use the same initial 3D human and object Gaussians, including both shape and texture attributes, for all methods. We then extract each method’s human (θ and β) and object (R , t , and s) pose parameters and apply them to transform the 3D Gaussians. This setup ensures that the only variable in the experiments is the set of 3D pose parameters provided by each method. For evaluation, we render the transformed Gaussians on the 2D background from pre-defined viewpoints 0° , 90° , 180° , and 270° . Each rendered image is then paired with its corresponding text description to compute two image-text alignment metrics: 1) CLIPScore [6] and 2) VQAScore [13]. CLIPScore computes the cosine similarity between the embeddings of the rendered image and the text description. VQAScore utilizes a powerful visual-question-answering (VQA) model to compute the alignment score by converting the text description into a simple query and measuring the generative likelihood of a desired response. Here, we use InstructBLIP-FlanT5-XL [4] as the underlying VQA model to compute VQAScore. Tab. S1 shows that

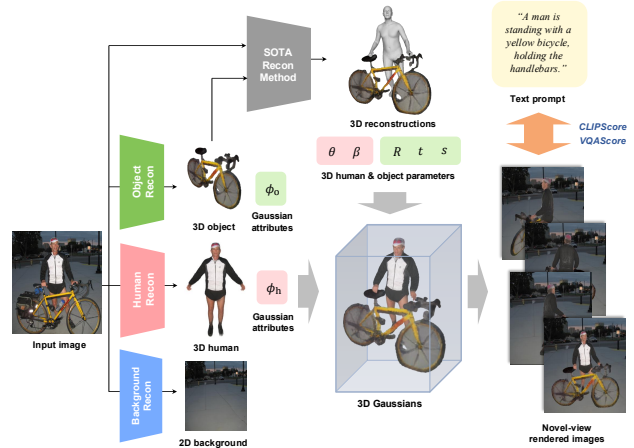


Figure S1. Process of evaluating text alignment for state-of-the-art reconstruction methods.

| Methods | Open3DHOI | |
|-----------------------|----------------------|---------------------|
| | CLIPScore \uparrow | VQAScore \uparrow |
| PHOSA [23] | 0.689 | 0.631 |
| LEMON [22] + PICO [3] | 0.696 | 0.642 |
| InteractVLM [5] | 0.694 | 0.647 |
| HOI-Gaussian [19] | 0.698 | 0.648 |
| TeHOR (Ours) | 0.706 | 0.652 |

Table S1. Quantitative evaluation of appearance-text alignment.

our framework outperforms other state-of-the-art methods in text alignment by effectively capturing the holistic and semantic context of human-object interaction.

S2. Comparison with HDM

Tab. S2 shows that our framework significantly outperforms HDM [20], a state-of-the-art method that reconstructs 3D human and object shapes as point clouds rather than meshes. Since HDM does not estimate SMPL-X mesh surfaces, the standard root-based alignment described in Sec. 4.2 cannot be directly applied for the evaluation. Instead, we employ the Iterative Closest Point (ICP) algorithm to align the reconstructed human vertices with the ground-truth (GT) vertices. Consequently, HDM, which is a learning-based approach, is particularly vulnerable to unseen object categories compared to those encountered during training (e.g., balls and suitcases). On the other hand,

| | Seen object categories | | | Whole object categories | | |
|---------------------|------------------------|------------------------|--------------|-------------------------|------------------------|--------------|
| | CD _{human} ↓ | CD _{object} ↓ | Collision↓ | CD _{human} ↓ | CD _{object} ↓ | Collision↓ |
| HDM [20] | 4.977 | 16.349 | 0.040 | 6.084 | 30.422 | 0.038 |
| TeHOR (Ours) | 2.511 | 14.905 | 0.035 | 2.582 | 17.938 | 0.047 |

Table S2. Quantitative comparison with HDM [20] on Open3DHOI [19].

| Contact estimation methods | Contact estimation | | | 3D reconstruction | | | |
|----------------------------|------------------------|------------------------|-------------------------|-----------------------|------------------------|----------|------------|
| | Contact _p ↑ | Contact _r ↑ | Contact _{ti} ↑ | CD _{human} ↓ | CD _{object} ↓ | Contact↑ | Collision↓ |
| w/o contact | — | — | — | 5.311 | 19.849 | 0.374 | 0.054 |
| P_{contact} | 0.282 | 0.342 | 0.309 | 4.941 | 16.701 | 0.412 | 0.047 |
| DECO [18] | 0.200 | 0.264 | 0.228 | 5.115 | 17.229 | 0.353 | 0.051 |
| LEMON [22] | 0.426 | 0.225 | 0.295 | 5.084 | 17.060 | 0.389 | 0.050 |
| InteractVLM [5] | 0.422 | 0.458 | 0.439 | 4.988 | 16.009 | 0.408 | 0.052 |

Table S3. Impact of contact estimation accuracy on TeHOR’s reconstruction, evaluated on Open3DHOI [19].



Figure S2. Enhancement of Gaussian via optimization process.

| | Open3DHOI | | | |
|--------------------------------|-----------------------|------------------------|--------------|--------------|
| | CD _{human} ↓ | CD _{object} ↓ | Contact↑ | Collision↓ |
| Before conversion | 5.020 | 16.987 | 0.394 | 0.052 |
| After conversion (Ours) | 4.941 | 16.701 | 0.412 | 0.047 |

Table S4. Impact of Gaussians-to-mesh conversion

since our framework is an optimization-based approach, it generalizes effectively to unseen objects by leveraging the strong prior knowledge of the diffusion network.

S3. Impact of contact estimation accuracy

Tab. S3 illustrates the impact of contact estimation accuracy on the 3D reconstruction performance of our framework. We compare our framework under different contact estimation settings, including specialized contact prediction models such as LEMON [22] and InteractVLM [5]. While these models improve the precision of contact localization on human and object surfaces, their contribution to final 3D reconstruction quality remains marginal. The contact estimation methods primarily focus on accurately predicting the boundaries of the contact region at a fine-grained level. However, regardless of how precise these contact boundaries are, the contact information alone cannot capture the holistic and semantic context of human-object interaction.

This observation suggests that capturing the holistic interaction context is far more important for the joint reconstruction of 3D human and object than precisely delineating fine-grained contact boundaries. Accordingly, the core

strength of our framework is determined by holistic contact reasoning supported by text-guided optimization rather than by accurate contact prediction. This validates our use of the contact text prompt P_{contact} as a lightweight yet effective alternative to external contact prediction models.

S4. Impact of Gaussian attributes optimization

Fig. S2 demonstrates the importance of optimizing the 3D human and object Gaussian attributes (ϕ_h and ϕ_o) within our framework. Since the initial Gaussian attributes can occasionally be incomplete due to occlusions in the input image, we further refine them through the appearance loss L_{appr} . This optimization process enhances the visual plausibility and overall coherence of the reconstructed human-object interactions. As there is no existing 3D HOI dataset that provides both geometry and texture annotations, quantitative evaluation of this optimization remains challenging; thus, we primarily present qualitative results.

S5. Impact of Gaussians-to-mesh conversion

Tab. S4 shows that our Gaussians-to-mesh conversion process, detailed in Sec. 3.4, is a crucial step for accurate mesh reconstruction. Direct conversion of 3D Gaussians to mesh surfaces often produces inconsistencies near contact regions. Accordingly, we use the conversion procedure that enforces geometric consistency between Gaussian-defined contact regions and the corresponding mesh vertices. As a result, it improves overall geometric accuracy and yields substantial gains in the contact evaluation score (Contact).

S6. Details of text captioning

Fig. S3 and Fig. S4 illustrate the two captioning instructions used to generate the holistic text prompt P_{holistic} and contact text prompt P_{contact} , with the GPT-4 [1] vision-language model (VLM). First, we generate the holistic prompt P_{holistic} , which describes the interaction between the

```

### TASK ###
Your goal is to provide a detailed description of the
given image, which depicts the interaction between a
person and an object.

- Focus only on the person whose body center is closest
to the image center.
- Identify the object most directly interacted with and
state the action.
- Output must be one sentence, no explanations, labels,
or reasoning.
- Additionally, explicitly output the single object that
is most directly interacted with.

### OUTPUT FORMAT ###
Output: {{interacting object}}, {{description}}

### OUTPUT EXAMPLE ###
Example 1 - Output: soccer ball, A woman is playing
soccer on a grassy field, dribbling the ball.
Example 2 - Output: small box, A man is seated on a
small box with legs crossed.
Example 3 - Output: chair, A woman is moving a chair
with one hand.

```

Figure S3. Captioning instruction for the VLM [1] to acquire holistic text prompt P_{holistic} .

person closest to the image center and the object most directly involved with that person. Then, we generate the contact prompt P_{contact} by providing both the input image and the holistic description P_{holistic} as inference cues, enabling it to infer which human body parts are in direct physical contact with the object. This two-stage captioning strategy encourages each stage to specialize in a distinct role: inferring global interaction semantics and localized contact information, respectively.

Fig. S5 highlights the strong capability of the text captioning process. As shown in the examples, it successfully captures key contextual cues essential for reasoning about human-object interactions, including the human’s action (e.g., sitting, riding, and performing) and the surrounding environment (e.g., pathway, mid-air, and grassy field). Even when the same object appears in different interaction scenarios, the VLM provides accurate and semantically appropriate descriptions. This demonstrates the richness of holistic contextual information, in contrast to contact cues that convey only local geometric proximity. Such comprehensive interaction cues play a crucial role in guiding our reconstruction framework toward more accurate and globally coherent 3D human and object reconstructions.

S7. Implementation details

We explain the implementation details of two stages: the reconstruction stage (Sec. 3.2) and HOI optimization stage (Sec. 3.3), below. PyTorch [15] is used for implementation.

S7.1. Reconstruction stage

Human reconstruction. When using SmartEraser [9], the object regions to be removed are inpainted using classifier-

```

### TASK ###
Your goal is to list the body parts of the person in the
given image that are in direct physical contact
with the object.

- Choose ONLY from this pre-defined list (multi-select
allowed): head, hips, ...
- The interacting object and reference description are
provided as follows: "{object}", "{description}".
- Focus only on the person whose body center is closest
to the image center.
- Identify the object most directly interacted with and
state the action.
- LEFT/RIGHT must be relative to the person (egocentric)
, not the viewer/camera.
- If no clear physical contact is visible, output none.

### OUTPUT FORMAT ###
Output: {{comma-separated list}}

### OUTPUT EXAMPLE ###
Example 1 - Output: left hand, right hand
Example 2 - Output: right foot
Example 3 - Output: none

```

Figure S4. Captioning instruction for the VLM [1] to acquire contact text prompt P_{contact} .

free guidance with a guidance scale of 1.5 in its generative diffusion network. To segment human region from the inpainted image, we use Grounded-SAM [12, 14] with a text prompt corresponding to the object category name obtained from the text captioning (Sec. S6). From the segmented human image, LHM operates on a canonical set of 40,000 Gaussian anchors uniformly sampled over the SMPL-X surface. For each anchor, LHM predicts the Gaussian attributes ϕ_h , including canonical offsets, opacity, scale, and appearance features, through a single feed-forward inference.

Object reconstruction. When using SmartEraser [9], we adopt the same settings as in the human reconstruction stage. To segment object region from the inpainted image, we use Grounded-SAM [12, 14] with the text prompt “human”. From the segmented object image, InstantMesh [21] first synthesizes six multi-view images using Zero123++ [17] with 75 diffusion steps, and then reconstructs a textured mesh through its triplane-based reconstruction network. The resulting textured mesh is subsequently converted into 3D object Gaussians ϕ_o , where the Gaussian centroids are placed at the mesh vertex positions and their initial appearance features are assigned from the mesh vertex colors. The Gaussian attributes are further optimized to match the 2D images rendered from the reconstructed textured mesh at 360 uniformly sampled viewpoints, following the optimization procedure of 3DGS [10].

S7.2. HOI optimization stage

We use the Adam [11] optimizer with an exponentially decaying learning rate. The initial learning rate is set to 1×10^{-2} for the object pose parameters (R , t , and s), 1×10^{-4} for the human pose parameters (θ and β), and



Figure S5. **Text captioning results on Open3DHOI [19].** Our text captioning produces accurate and rich text descriptions for a wide range of interaction scenarios.

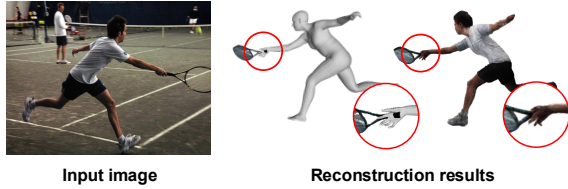


Figure S6. **Failure case of reconstructing local details.**

1×10^{-4} for the human and object Gaussian attributes (ϕ_h and ϕ_o). We run the optimization for $N = 200$ steps on a single NVIDIA RTX 8000 GPU. Under this setting, the average optimization time per sample is 134 seconds.

Appearance rendering. During optimization, we render the 3D human Gaussians Φ_h and object Gaussians Φ_o using a spherical coordinate system (r, v, ψ) , where r denotes the distance to the spherical origin, v the elevation angle, and ψ the azimuth angle. We uniformly sample viewpoints with $r \in [1.0, 2.5]$, $v \in [-30^\circ, 30^\circ]$, and $\psi \in [-180^\circ, 180^\circ]$. Since human-object interaction primarily involves the upper body, such as the head and hands, we additionally use zoomed-in camera views focused on this region. For these upper-body views, we set the spherical origin to the 3D position of the SMPL-X spine keypoint and sample $r \in [0.7, 1.5]$, $v \in [-30^\circ, 30^\circ]$, and $\psi \in [-180^\circ, 180^\circ]$.

Appearance loss. We compute the appearance loss $\mathcal{L}_{\text{appr}}$ of Eq. (2) using StableDiffusion-v2.1 [16] and apply classifier-

free guidance [7] with a guidance scale of 15.0 for noise estimation. The noise levels are defined at randomly sampled timesteps within $[0.02, 0.98]$. To ensure stable optimization, we clip loss gradients to a maximum norm of 1.0.

S8. Limitations and future work

Reconstruction of local details. While our framework captures holistic human-object interactions effectively, it may overlook fine-grained local details such as small accessories or subtle surface deformations, as shown in Fig. S6. This limitation occurs because the appearance loss of our framework primarily offers global guidance and lacks fine-grained supervision that specifically addresses local regions. A promising future direction is to design localized, text-driven supervision that specializes in local regions to further enhance fine-detail reconstruction.

Video as input. Our framework aims to jointly reconstruct 3D human and object from a single image. When extending the method to video input, additional considerations become essential, such as maintaining temporal consistency across frames and ensuring consistent geometry and texture over time. With the recent emergence of text-to-video generative models [2, 8], future work could leverage these advances by using text descriptions as a key guidance, enabling more stable and temporally coherent 3D HOI reconstruction.

S9. More qualitative results

We provide additional qualitative comparison results of our TeHOR in Figs. S7 to S10. These examples further demonstrate the effectiveness of our method in reconstructing realistic and semantically coherent human–object interactions. Please note that the left-side results of TeHOR are mesh-based renderings, while the right-side results are Gaussian-based renderings. Due to the inherent characteristics of 3D Gaussian representations, Gaussian renderings can appear slightly larger and exhibit blurred boundaries.

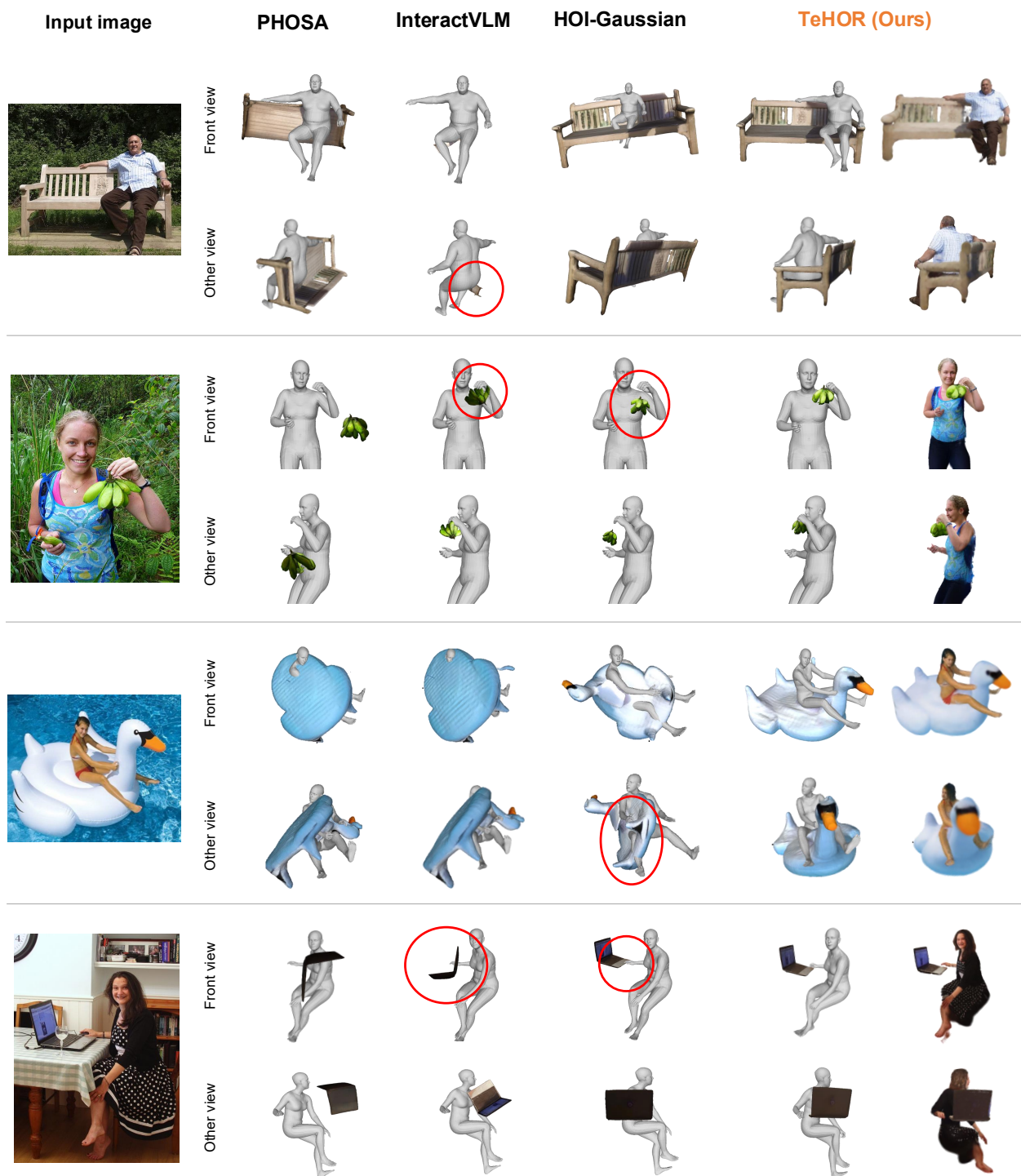


Figure S7. More qualitative comparison of 3D human and object reconstruction with PHOSA [23], InteractVLM [5], and HOI-Gaussian [19] on Open3DHOI [19]. We highlight their representative failure cases with red circles.

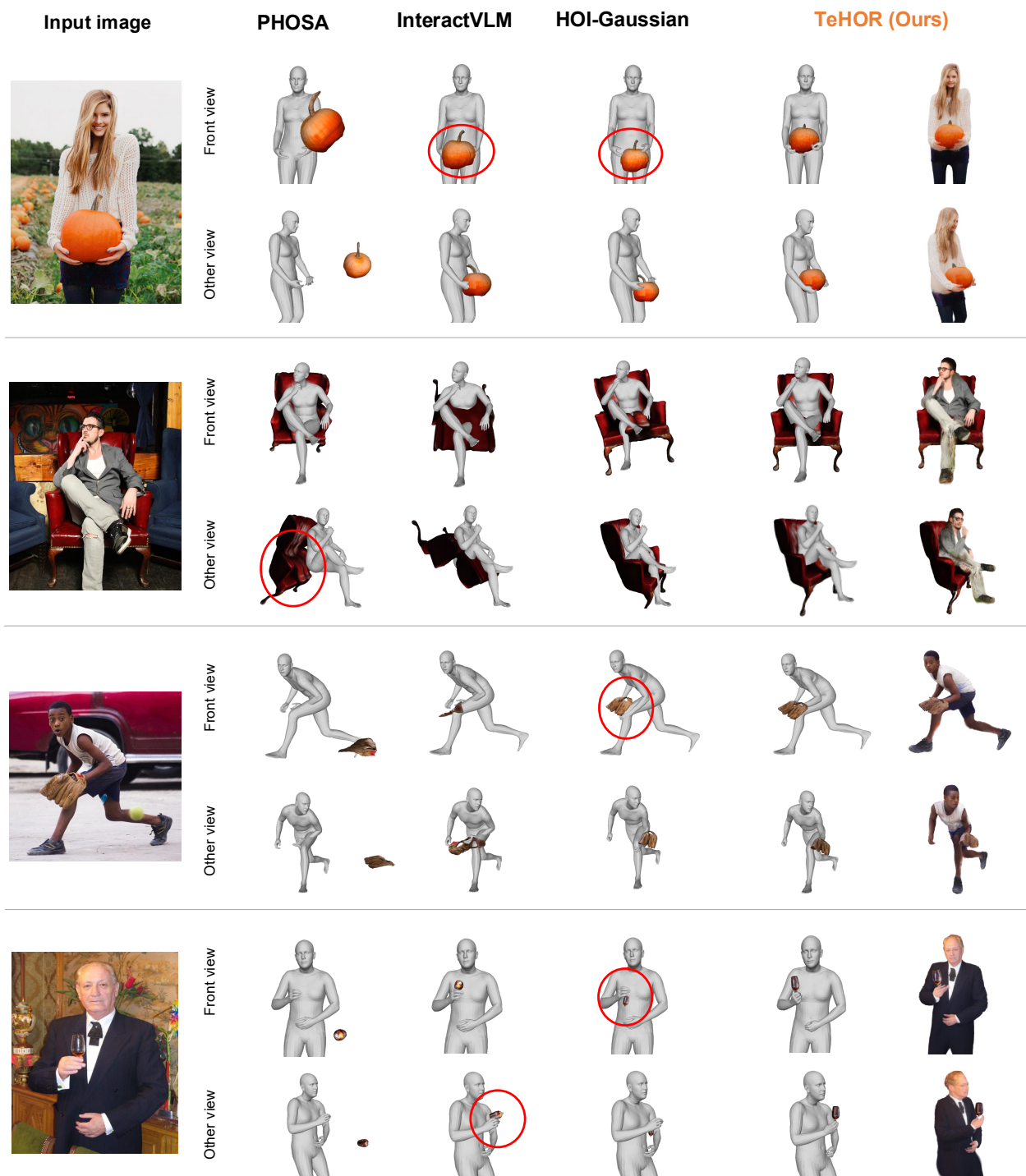


Figure S8. More qualitative comparison of 3D human and object reconstruction with PHOSA [23], InteractVLM [5], and HOI-Gaussian [19] on Open3DHOI [19]. We highlight their representative failure cases with red circles.

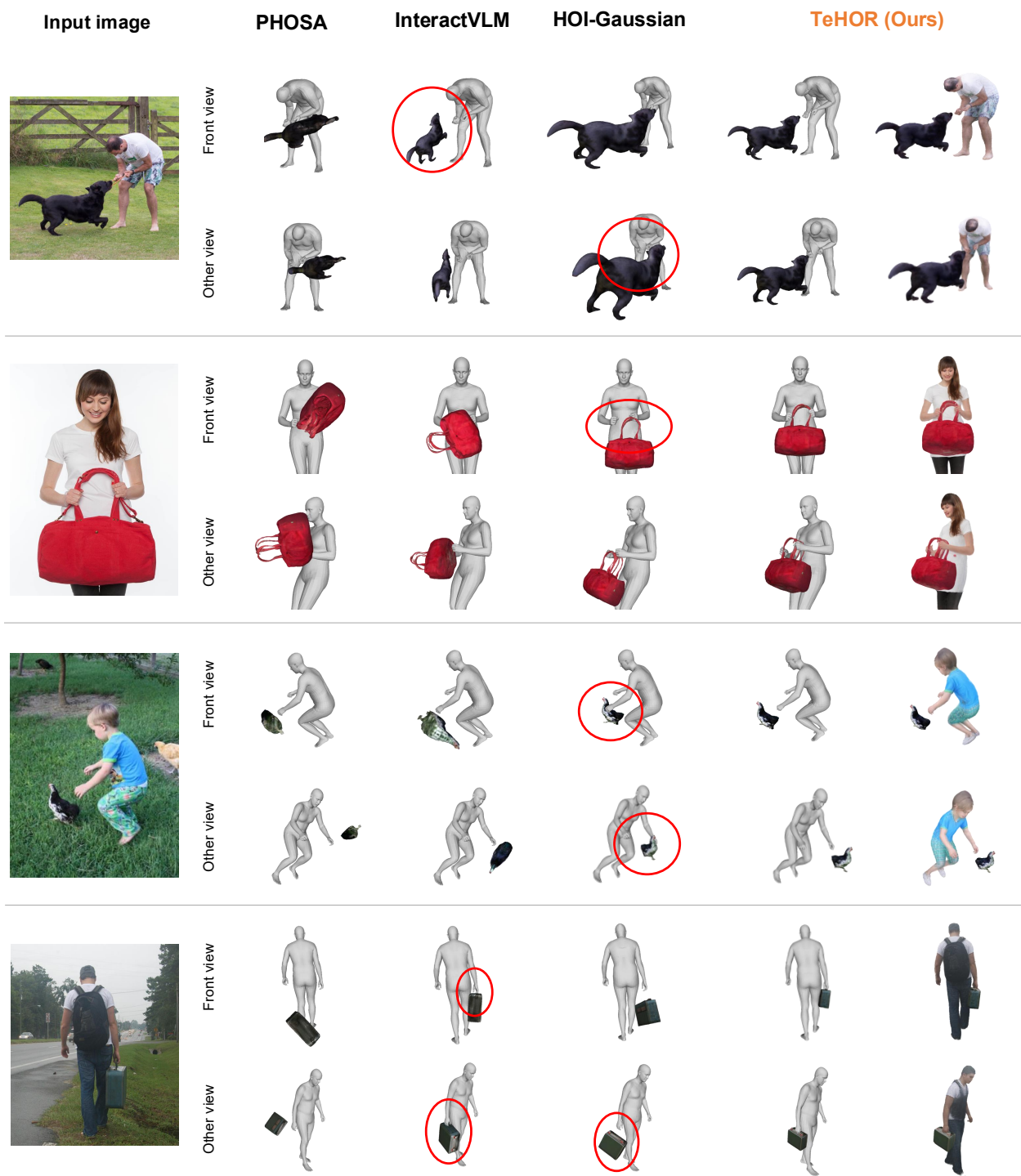


Figure S9. More qualitative comparison of 3D human and object reconstruction with PHOSA [23], InteractVLM [5], and HOI-Gaussian [19] on Open3DHOI [19]. We highlight their representative failure cases with red circles.

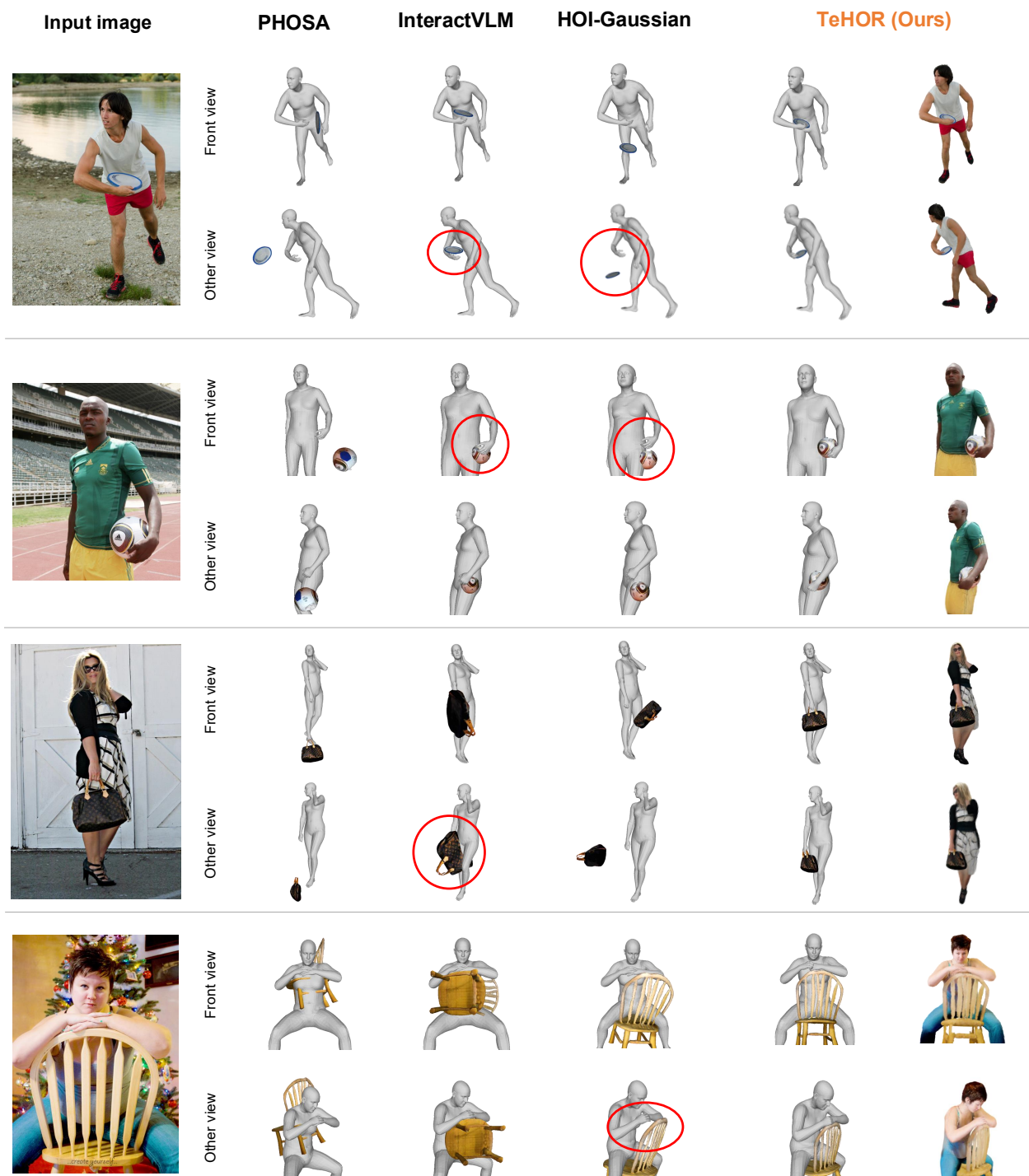


Figure S10. More qualitative comparison of 3D human and object reconstruction with PHOSA [23], InteractVLM [5], and HOI-Gaussian [19] on Open3DHOI [19]. We highlight their representative failure cases with red circles.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [2](#), [3](#)
- [2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. [4](#)
- [3] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *CVPR*, 2025. [1](#)
- [4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstantBLIP: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023. [1](#)
- [5] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *CVPR*, 2025. [1](#), [2](#), [6](#), [7](#), [8](#), [9](#)
- [6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021. [1](#)
- [7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS workshop*, 2021. [4](#)
- [8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. [4](#)
- [9] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. SmartEraser: Remove anything from images using masked-region guidance. In *CVPR*, 2025. [3](#)
- [10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. [3](#)
- [11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [3](#)
- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. [3](#)
- [13] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024. [1](#)
- [14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. [3](#)
- [15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. [3](#)
- [16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. [4](#)
- [17] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. [3](#)
- [18] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *ICCV*, 2023. [2](#)
- [19] Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions. In *CVPR*, 2025. [1](#), [2](#), [4](#), [6](#), [7](#), [8](#), [9](#)
- [20] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *CVPR*, 2024. [1](#), [2](#)
- [21] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. [3](#)
- [22] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. LEMON: Learning 3D human-object interaction relation from 2D images. In *CVPR*, 2024. [1](#), [2](#)
- [23] Jason Y Zhang, Sam PePOSE, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. [1](#), [6](#), [7](#), [8](#), [9](#)