

# SUPPLEMENTARY MATERIAL

**Anonymous authors**

Paper under double-blind review

## 1 IMPLETION DETAILS

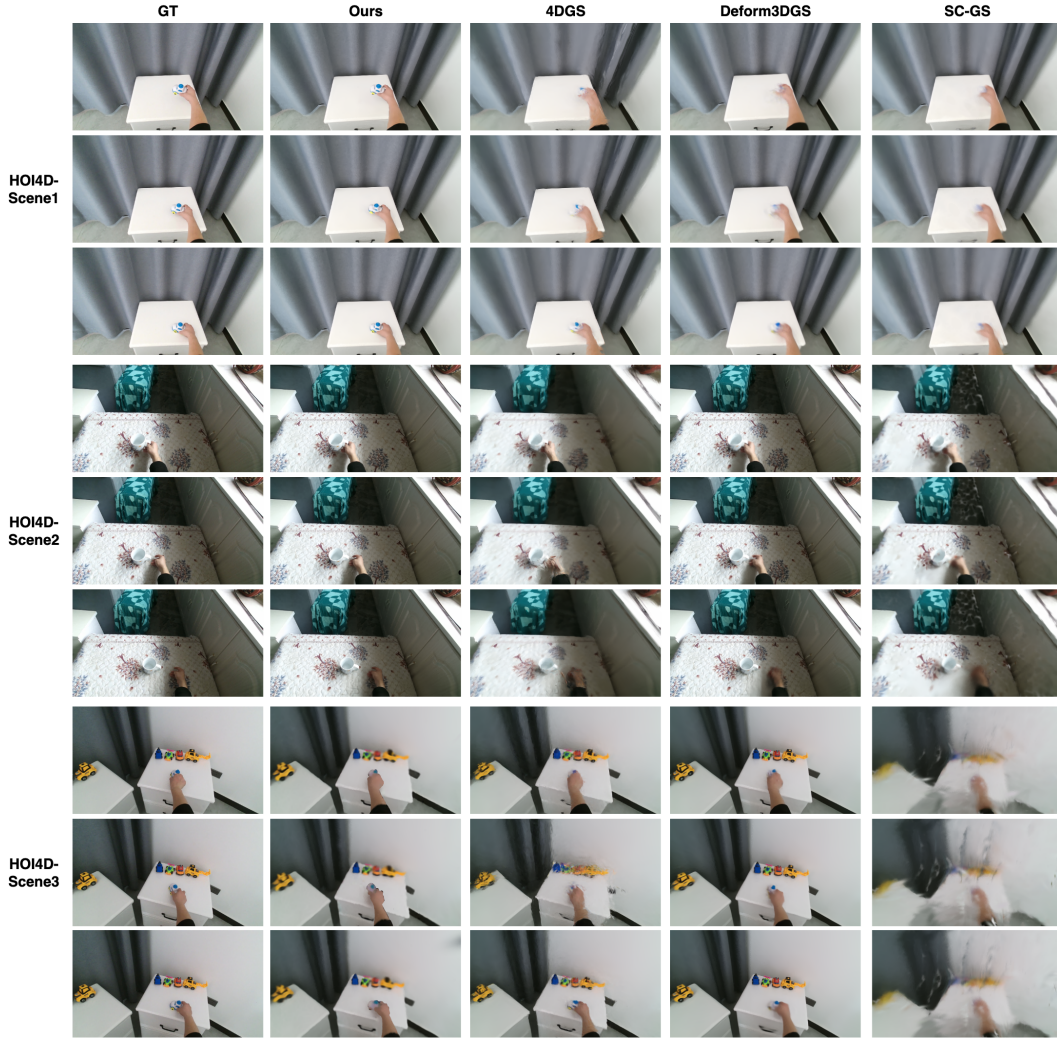


Fig. 1. More Qualitative comparison of our method and the baseline methods on the HOI4D dataset.

**6D Representation of Rotation.** While previous works Yang et al. (2024); Huang et al. (2024) typically output quaternions directly from the implicit field, this approach can lead to numerical instability, particularly when handling large angular rotations in Hand-Object Interaction (HOI) scenes. To address this, we adopt an intermediate 6D representation  $\Delta \mathbf{R}_{6D} \in \mathbb{R}^6$  for rotations. This 6D representation avoids the ambiguities associated with quaternions and eliminates the need for direct conversion between quaternions and rotation matrices. By converting the 6D representation to a rotation matrix  $\mathbf{R}$ , we naturally satisfy the unit-norm constraints of quaternions, thereby accelerating

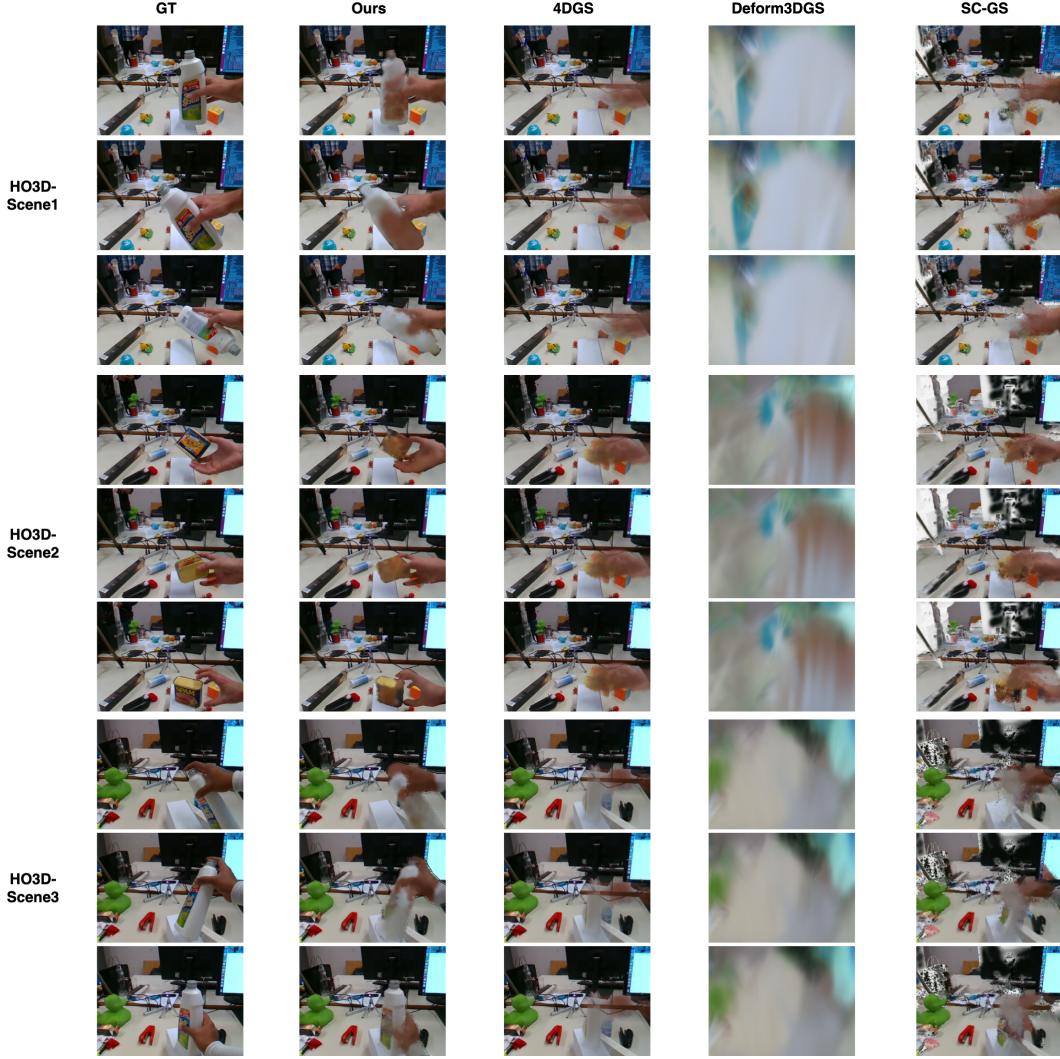


Fig. 2. More Qualitative comparison of our method and the baseline methods on the HO3D dataset.



Fig. 3. **Failure case on complex interaction and trajectory.** When the scene has more complex interactions, such as holding an object and interacting with another object, tracking such a scene for a long time will cause Gaussian constraints to be difficult and affect the Gaussian of the two objects.

convergence and reducing computational complexity:

$$\mathbf{R}_i^t = r_{6D \rightarrow \text{quaternion}} \left( \sum_{k=1}^K w_i^k \Delta \mathbf{R}_{6D} \right) \otimes \mathbf{R}_i, \quad (1)$$

$\mathbf{R}_i \in \mathbb{R}^4$  denotes the rotation of the  $i$ -th canonical Gaussian, and  $\mathbf{R}_i^t \in \mathbb{R}^4$  represents the rotation of the  $i$ -th Gaussian at timestamp  $t$ . The function  $r_{6D \rightarrow \text{quaternion}}(\cdot)$  transforms Kocabas et al. (2024) the 6D representation into a quaternion, and  $\otimes$  denotes quaternion multiplication.

**Details of HO3D Datasets.** We use the HO3D V3 dataset Hampali et al. (2020) for experiments. The dataset provides images at 640x480 resolution and segmentation masks at 320x240 or 160x120



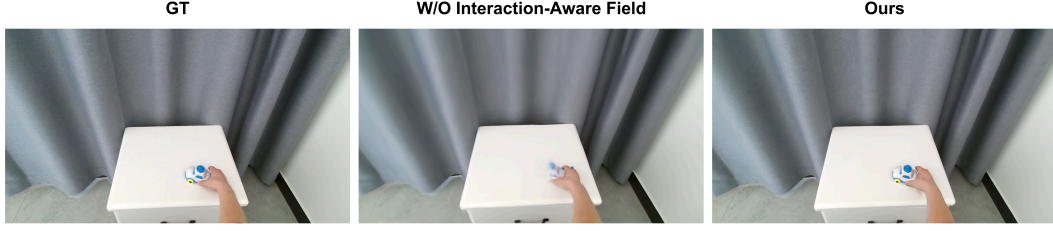


Fig. 4. A qualitative ablation on our interaction-aware design.

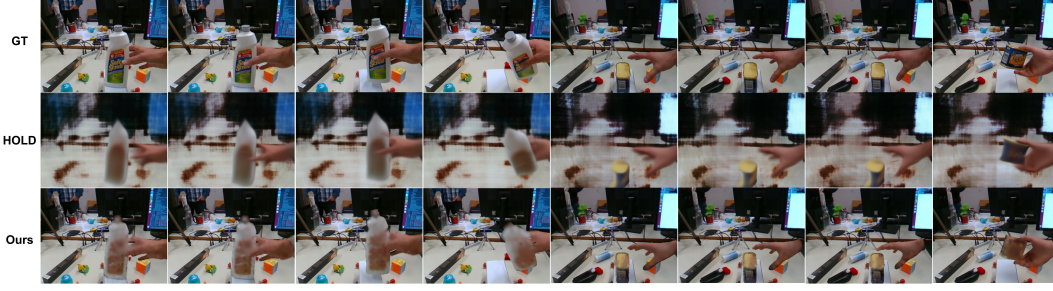


Fig. 5. **Qualitative comparison of our approach and HOLD Fan et al. (2024) on HO3D dataset.** Compared to HOLD, our approach achieves better rendering quality and more complete HOI scene preservation

resolution. To ensure consistency, we process both images and segmentation masks to  $320 \times 240$  resolution. This setting is uniformly applied to all baseline methods. We select 4 sub-datasets (ABF14, GPMF14, GSF14, SB14) from the HO3D v3 dataset.

**Details of HOI4D Datasets.** We evaluate our approach and all baseline methods on four sub-datasets from the HOI4D dataset Liu et al. (2022): ZY20210800001/H1/C1/N19/S100/s02/T1 (transl 1), ZY20210800001/H1/C2/N32/S92/s05/T1 (transl 2), ZY20210800001/H1/C1/N19/S100/s02/T2 (r&t 1), and ZY20210800001/H1/C1/N44/S99/s02/T2 (r&t 2).

## 2 DETAILS OF LOSS FUNCTIONS.

**Mask Loss.** We optimize the hand field and object field using hand-object interaction region masks. The mask loss is defined separately for each field as:

$$\begin{aligned}\mathcal{L}_{\text{mask}}^{\text{H}} &= \|\hat{H} - H\|_2, \\ \mathcal{L}_{\text{mask}}^{\text{O}} &= \|\hat{O} - O\|_2,\end{aligned}\tag{2}$$

where  $\hat{H}$  is the rendered hand image,  $H$  is the ground truth hand region from segmentation masks,  $\hat{O}$  is the rendered object image, and  $O$  is the ground truth object region from segmentation masks.

**Alpha Loss.** The predicted alpha mask  $\alpha_{\text{motion}}$  is computed as the sum of the alpha values from the rendered object ( $\alpha_{\text{obj}}$ ) and the rendered hand ( $\alpha_{\text{hand}}$ ):

$$\alpha_{\text{motion}} = \alpha_{\text{obj}} + \alpha_{\text{hand}}.\tag{3}$$

To ensure that the alpha values remain within the valid range  $[0, 1]$ , the predicted alpha mask is clamped:

$$\alpha_{\text{motion}} = \text{clamp}(\alpha_{\text{motion}}, 0.0, 1.0).\tag{4}$$

The alpha loss  $\mathcal{L}_{\text{mask}}$  is computed as the L2 distance (mean squared error) between the predicted alpha mask  $\alpha_{\text{motion}}$  and the ground truth alpha mask  $\alpha_{\text{gt}}$ :

$$\mathcal{L}_{\text{alpha}} = \|\alpha_{\text{motion}} - \alpha_{\text{gt}}\|_2^2,\tag{5}$$

where:

- $\alpha_{\text{motion}}$  is the predicted alpha mask,
- $\alpha_{\text{gt}}$  is the ground truth alpha mask,

| Method                        | HOI4D           |                 |                    |                 |                 |                    |                 |                 |                    |                 |                 |                    |
|-------------------------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|
|                               | transl 1        |                 |                    | transl 2        |                 |                    | r&t 1           |                 |                    | r&t 2           |                 |                    |
|                               | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ |
| 4DGS Wu et al. (2024)         | 26.8711         | 0.8811          | 0.4444             | 22.8575         | 0.7186          | 0.4876             | 24.7754         | 0.8642          | 0.4126             | 22.5872         | 0.8297          | 0.3594             |
| Deform3DGS Yang et al. (2024) | 26.6744         | 0.9120          | 0.3597             | 25.9850         | 0.8191          | 0.2191             | 25.5696         | 0.9145          | 0.2814             | 21.5617         | 0.8563          | 0.2233             |
| SC-GS Huang et al. (2024)     | 29.8171         | 0.9253          | 0.4204             | 23.3511         | 0.7485          | 0.4956             | 17.4740         | 0.7456          | 0.4800             | 17.1735         | 0.6822          | 0.4815             |
| Ours                          | 32.9579         | 0.9490          | 0.3574             | 27.6724         | 0.9049          | 0.2181             | 25.7351         | 0.8945          | 0.3928             | 22.5911         | 0.8186          | 0.3438             |

| Method                        | HO3D            |                 |                    |                 |                 |                    |                 |                 |                    |                 |                 |                    |
|-------------------------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|-----------------|-----------------|--------------------|
|                               | ABF14           |                 |                    | GPMF14          |                 |                    | GSF14           |                 |                    | SB14            |                 |                    |
|                               | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ |
| 4DGS Wu et al. (2024)         | 18.8732         | 0.7898          | 0.2585             | 20.0059         | 0.8225          | 0.2457             | 19.7866         | 0.8440          | 0.2272             | 19.0762         | 0.8150          | 0.2506             |
| Deform3DGS Yang et al. (2024) | 8.1700          | 0.3273          | 0.6551             | 10.4354         | 0.3351          | 0.6670             | 9.2258          | 0.3420          | 0.6532             | 10.8954         | 0.4178          | 0.6207             |
| SC-GS Huang et al. (2024)     | 22.1154         | 0.8415          | 0.2059             | 19.9291         | 0.8591          | 0.1999             | 20.0910         | 0.7619          | 0.3269             | 19.3326         | 0.7341          | 0.3147             |
| HOLD Fan et al. (2024)        | 17.6263         | 0.8505          | 0.2841             | 18.2775         | 0.8195          | 0.3145             | 18.1284         | 0.8499          | 0.2630             | 18.0960         | 0.8532          | 0.2191             |
| Ours                          | 26.0459         | 0.8814          | 0.1563             | 24.3312         | 0.9009          | 0.1360             | 24.0997         | 0.8960          | 0.1501             | 24.2757         | 0.8942          | 0.1799             |

Tab. 1. **Comparison with SOTA dynamic Gaussian Splatting methods on HOI4D and HO3D.** We evaluate our method and three other SOTA baselines on the HOI4D and HO3D datasets.

- $\|\cdot\|_2^2$  denotes the squared L2 norm.

The alpha loss function is designed to measure the discrepancy between the predicted alpha mask and the ground truth alpha mask. This loss is used to ensure that the rendered transparency values (alpha) of both the object and the hand align with the ground truth mask.

**Momentum Loss.** Momentum loss is a constraint term used to enforce the conservation of momentum in physical systems. It ensures that the predicted motion of interacting entities (e.g., hands and objects) adheres to the principle of momentum conservation, which states that the total momentum of a closed system remains constant in the absence of external forces. Mathematically, the momentum loss  $\mathcal{L}_{\text{momentum}}$  is defined as:

$$\mathcal{L}_{\text{momentum}} = |m_h \bar{\mathbf{a}}_h + m_o \bar{\mathbf{a}}_o|, \quad (6)$$

where:

- $m_h$  and  $m_o$  are the masses of the hand and object, respectively,
- $\bar{\mathbf{a}}_h$  and  $\bar{\mathbf{a}}_o$  are the mean accelerations of the hand and object, respectively.

This loss penalizes deviations from momentum conservation, ensuring that the predicted interactions are physically plausible.

**Penetration Loss.** Penetration loss is a geometric constraint term used to prevent unrealistic intersections between objects in a 3D space. It is particularly important in scenarios involving human-object interaction (HOI), where the hand or body should not penetrate the object. The penetration loss  $\mathcal{L}_{\text{penetrate}}$  is computed as:

$$\mathcal{L}_{\text{penetrate}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\mathbf{p}_i \in \text{BBox}), \quad (7)$$

where:

- $\mathbf{p}_i = (x_i, y_i, z_i)$  is the  $i$ -th point of the hand,
- BBox is the bounding box of the object, defined as  $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}] \times [z_{\min}, z_{\max}]$ ,
- $\mathbb{I}(\cdot)$  is an indicator function that returns 1 if the point lies inside the bounding box and 0 otherwise.

This loss penalizes points that penetrate the object's bounding box, ensuring that the predicted hand and object positions are geometrically consistent.



## 2.1 OPTIMIZATION DETAILS.

**Warm-Up.** During the warm-up stage, we optimize hand and object Gaussians using object loss  $\mathcal{L}_{\text{rot}}^{\text{O}}$ , hand loss  $\mathcal{L}_{\text{trans}}^{\text{H}}$ , and render loss  $\mathcal{L}_{\text{render}}$ , formulated as:

$$\begin{aligned}\mathcal{L}_{\text{H}} &= \mathcal{L}_{\text{render}}^{\text{H}} + 0.001 \cdot \mathcal{L}_{\text{trans}}^{\text{H}} + 0.001 \cdot \mathcal{L}_{\text{mask}}^{\text{H}}, \\ \mathcal{L}_{\text{O}} &= \mathcal{L}_{\text{render}}^{\text{O}} + 0.001 \cdot \mathcal{L}_{\text{rot}}^{\text{O}} + 0.001 \cdot \mathcal{L}_{\text{mask}}^{\text{O}}.\end{aligned}\quad (8)$$

Here,  $\mathcal{L}_{\text{render}}$  combines  $\mathcal{L}_1$  and SSIM metrics, following the baseline approach Huang et al. (2024); Kerbl et al. (2023).  $\mathcal{L}_{\text{render}}^{\text{H}}$  and  $\mathcal{L}_{\text{render}}^{\text{O}}$  apply  $\mathcal{L}_{\text{render}}$  to the rendered hand and object images, respectively. These losses establish an appropriate initial distribution of hand-object Gaussians and provide a foundation for collaborative reconstruction.

**Collaboration Reconstruction.** In this stage, we use render loss  $\mathcal{L}_{\text{render}}$ , penetration loss  $\mathcal{L}_{\text{penetrate}}$ , momentum loss  $\mathcal{L}_{\text{momentum}}$ , alpha loss  $\mathcal{L}_{\text{alpha}}$ , and interaction loss  $\mathcal{L}_{\text{interaction}}$ . The total loss  $\mathcal{L}_{\text{total}}$  is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{render}}^{\text{HOI}} + 0.1 \cdot \mathcal{L}_{\text{alpha}} \\ &\quad + 0.01 \cdot (\mathcal{L}_{\text{interaction}} + \mathcal{L}_{\text{penetrate}} + \mathcal{L}_{\text{momentum}}),\end{aligned}\quad (9)$$

where  $\mathcal{L}_{\text{render}}^{\text{HOI}}$  computes  $\mathcal{L}_{\text{render}}$  Kerbl et al. (2023); Huang et al. (2024) using the entire rendered hand-object interaction (HOI) image and the ground truth image. These losses ensure plausible occlusion relationships, smooth edge transitions, geometric fidelity, and lighting coherence.

## 2.2 EXPERIMENTAL SETUP

We evaluate our method on the HOI4D Liu et al. (2022) and HO3D Hampali et al. (2020) datasets using one NVIDIA RTX 3090. Our method achieves optimal results in approximately 1 hour 20 minutes and 21k iterations. As a reference, SC-GS Huang et al. (2024) achieves 29.8171 PSNR in approximately 1 hour over 11k iterations on the HOI4D transl 1 scene.

## 2.3 HOI4D DATASET ANALYSIS

As show in the Tab. 1, We presents more results about our method and other SOTA dynamic Gaussian Splatting methods Wu et al. (2024); Yang et al. (2024); Huang et al. (2024) on the HOI4D dataset Liu et al. (2022).

**4DGS Performance.** 4DGS shows moderate performance on HOI4D. In **Translation 1**, it achieves a PSNR of 26.8711 and SSIM of 0.8811, but its LPIPS (0.4444) is higher than Deform3DGS and Ours, indicating lower perceptual quality. In **Translation 2**, performance drops significantly (PSNR: 22.8575, SSIM: 0.7186), highlighting limitations in complex dynamic scenes. For **r&t 1** and **r&t 2**, 4DGS is outperformed by Ours in both PSNR and SSIM, suggesting room for improvement in rotational modeling.

**Deform3DGS Performance.** Deform3DGS excels in perceptual quality. In **Translation 1**, it achieves SSIM of 0.912 and LPIPS of 0.3597. In **Translation 2**, it achieves the best LPIPS (0.2191) but lower PSNR (25.985) and SSIM (0.8191) compared to Ours. For **r&t 1** and **r&t 2**, Deform3DGS remains competitive but is outperformed by Ours in PSNR and SSIM.

**SC-GS Performance.** SC-GS performs well in simple tasks but struggles in complex scenes. In **Translation 1**, it achieves the highest PSNR (29.8171) and SSIM (0.9253), but its LPIPS (0.4204) is higher than Deform3DGS and Ours. In **Translation 2**, performance drops significantly (PSNR: 23.3511, SSIM: 0.7485). For **r&t 1** and **r&t 2**, SC-GS performs poorly, indicating weak rotational modeling.

**Ours Performance.** Ours consistently outperforms all methods on HOI4D. In **Translation 1**, it achieves PSNR of 32.9579 and SSIM of 0.949, with the lowest LPIPS values across most tasks. In **Translation 2**, its LPIPS (0.2181) is significantly lower than others. Ours excels in both translation and rotation tasks, particularly in complex HOI scenarios. Meanwhile, to demonstrate the importance of our interaction-aware field, we conducted a relative ablation study by eliminating both the field parameters and their associated training scheme from our framework. The qualitative and quantitative comparisons on HOI4D-Scene 1 (Table 2 and Figure 4) clearly show the degradation in performance when our interaction-aware modeling is absent.

| Method                | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ |
|-----------------------|-----------------|-----------------|--------------------|
| w/o interaction-aware | 28.7685         | 0.9145          | 0.3997             |

Tab. 2. An ablation on the interaction-aware design.

## 2.4 HO3D DATASET ANALYSIS

As show in the Tab. 1, We presents more results about our method and other SOTA dynamic Gaussian Splatting methods Wu et al. (2024); Yang et al. (2024); Huang et al. (2024); Fan et al. (2024) on the HO3D dataset Hampali et al. (2020). Furthermore, we include comparisons with HOLD Fan et al. (2024) on the HO3D dataset, a method specifically designed for hand-object interaction scene reconstruction.

**4DGS Performance.** 4DGS shows moderate performance on HO3D but struggles in complex scenes. In **ABF14**, it achieves PSNR of 18.8732 and SSIM of 0.7898, but its LPIPS (0.2585) is higher than Ours. In **SB14**, it performs poorly (PSNR: 19.0762, SSIM: 0.8150), indicating limitations in complex dynamic scenes.

**Deform3DGS Performance.** Deform3DGS performs poorly on HO3D. In **ABF14**, its PSNR is 8.1700 and SSIM is 0.3273, with the highest LPIPS values across tasks. This suggests significant challenges in handling large pose annotation errors.

**SC-GS Performance.** SC-GS shows mixed performance on HO3D. In **ABF14**, it achieves PSNR of 22.1154 and SSIM of 0.8415, but its LPIPS (0.2059) is higher than Ours. In **GSF14** and **SB14**, it performs poorly, indicating limitations in complex dynamic scenes.

**HOLD Performance.** HOLD demonstrates competitive perceptual metrics (SSIM  $\geq 0.8195$ , LPIPS  $\leq 0.3145$ ) but suffers from lower PSNR values ( $\leq 18.28$  dB), indicating a trade-off between structural preservation and pixel-level accuracy.

**Ours Performance.** Ours consistently outperforms all methods on HO3D. In **GPMF14**, it achieves PSNR of 24.3312 and SSIM of 0.9009, with the lowest LPIPS values across tasks. Ours excels in both translation and co-existing translation&rotation tasks, demonstrating superior performance in complex HOI scenarios.

## 2.5 SUMMARY

On HOI4D, 4DGS shows moderate performance but struggles in complex dynamic scenes. Deform3DGS excels in perceptual quality but underperforms in complex tasks. SC-GS performs well in simple tasks but struggles in complex scenes. Ours consistently outperforms all methods, achieving the highest PSNR, SSIM, and lowest LPIPS values.

On HO3D, 4DGS shows moderate performance but struggles in complex scenes. Deform3DGS performs poorly across all tasks, highlighting limitations in handling large pose errors. SC-GS shows mixed performance, excelling in simple tasks but struggling in complex scenes. Ours consistently outperforms all methods, demonstrating superior performance in both translation and co-existing translation&rotation tasks.

## 3 FAILURE CASES

As shown in Fig. 3, We observe that point-based rendering methods struggle with complex edge occlusion, interaction actions, and long-term tracking when holding and interacting with multiple objects, as Gaussians from both objects produce multiple occlusions. While our method effectively constrains hand-object interactions, it cannot fully constrain interactions between objects, leading to Gaussian overlap and color distortion. This limitation may require additional spatiotemporal data on object interactions to establish stronger constraints and optimization methods.

## REFERENCES

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects

from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 494–504, 2024.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3196–3206, 2020.

Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Scgs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4220–4230, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 505–515, 2024.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20331–20341, 2024.