A. Implementation Details

All our experiments are conducted on a single Nvidia A6000 GPU. Given the user input of image scene and text prompts, we used ChatGPT-4o to conduct all the VLM parsing, including the screenwriting stage and VLM-based CoT contact reasoning. For each interactive action in the parsing results, we will generate the 3D-aware keyframes. Before pose generation, we first need to use Trellis [8] to reconstruct the canonicalized target object, which costs around 1 minute. Once reconstructed, the object can be reused in all related interaction generation. We use the Orientate-Anything [7] to canonicalize the object. Using the canonical view rendering results, the diffusion inpainting model [1] uses 50 steps to inpaint the human on it, which costs around 15 seconds. For 3D human lifting, we used the AdamW optimizer with $lr = 1e^{-3}$ for 1.5k steps, and it only costs less than 2 minutes. To place the 3D human from object space to scene space, we need to estimate the reconstructed object pose. Then, we sample the candidate poses based on the initial guess and use the general visual features extracted by DINOv2 [4] to measure the similarity between the input object image and the candidate views' observation. After matching the reconstructed 3D object based on the input object image, we need to transform the object from the object space to world space. Since the reconstructed 3D object from Trellis has both mesh representation and 3D Gaussian representation, we refine the 6D pose of the object to place it into the scene, aligning it with the original scene image based on RGB constraints via 3D Gaussian rendering and depth constraints via mesh rasterization. The ground truth depth of the input image is estimated from MoGe [6].

B. Non-contact HSI Case

In some cases, the character is not in contact with the object, such as "standing next to the chair"; consequently, we will not receive any contact cues from VLM CoT reasoning. To impose spatial constraints on interactions, we reformulate the interaction loss $\mathcal{L}_{hoi}$ to constrain the minimal distance d_{HO} between the human mesh and the object mesh, measured in the normalized human mesh space rather than world space:

$$\mathcal{L}_{hoi} = \begin{cases} 0, & d_{HO} \leq \delta \\ d_{HO} - \delta, & d_{HO} > \delta \end{cases} \tag{1}$$

where $\delta = 10cm$.

C. Unnatural Avatar Appearance

One of the limitations of our work is that the avatar often looks "plastic" in the generated video, due to the independent modeling between the 3D Gaussian Avatar and the 3D Gaussian scene. Although these lighting differences between the scene and the avatar are not the problem we focused on, we now include a harmonization step [2] that significantly improves appearance (Fig. 1). We also believe that this problem can eliminate artifacts in most existing composition image harmonization methods through post-processing. Sometimes, the human scale in the generated video is implausible which is the limitation of the image inpainting model.

Figure 1. Harmonization (green) significantly improves our appearance.

D. Human Body Surface Partition Definition

To encourage ChatGPT outputs faithful and executable results, we divide the surface of the human body into 15 parts based on SMPL-X [5] template, *i.e.*, "head", "left upper arm", "right upper arm", "left forearm", "right forearm", "left hand", "right hand", "back", "buttocks", "left thigh", "right thigh", "left calf", "right calf", "left foot", and "right foot".

E. 2D Human Insertion Baseline

We add a 2D baseline based on Flex[1] with the pose generated from GenHSI. Flex changes the contents in the scene (Fig. 2b) and is time-consuming (30s vs 0.1s for 3DGS) to create the keyframes, which is inefficient for long video generation via keyframe interpolation. Note that GenHSI only uses 3D to constrain plausible affordances and rendering, and it will not decrease the controllability of pose generation. Different with feed-forward inpainting solution, our inpainting solution can detect human inpainting results during the denoising loop, effectively serving as an automatic verifier of the inpainting process.

(a) Scene Image (b) 2D HSI Image

F. Qualitative Results of 3D human Scene Interaction with incomplete 3D Scene

We also test previous methods in our challenging case, *i.e.*, the single-view image of the scene is the only accessible

[1] https://huggingface.co/ostris/Flex.2-preview

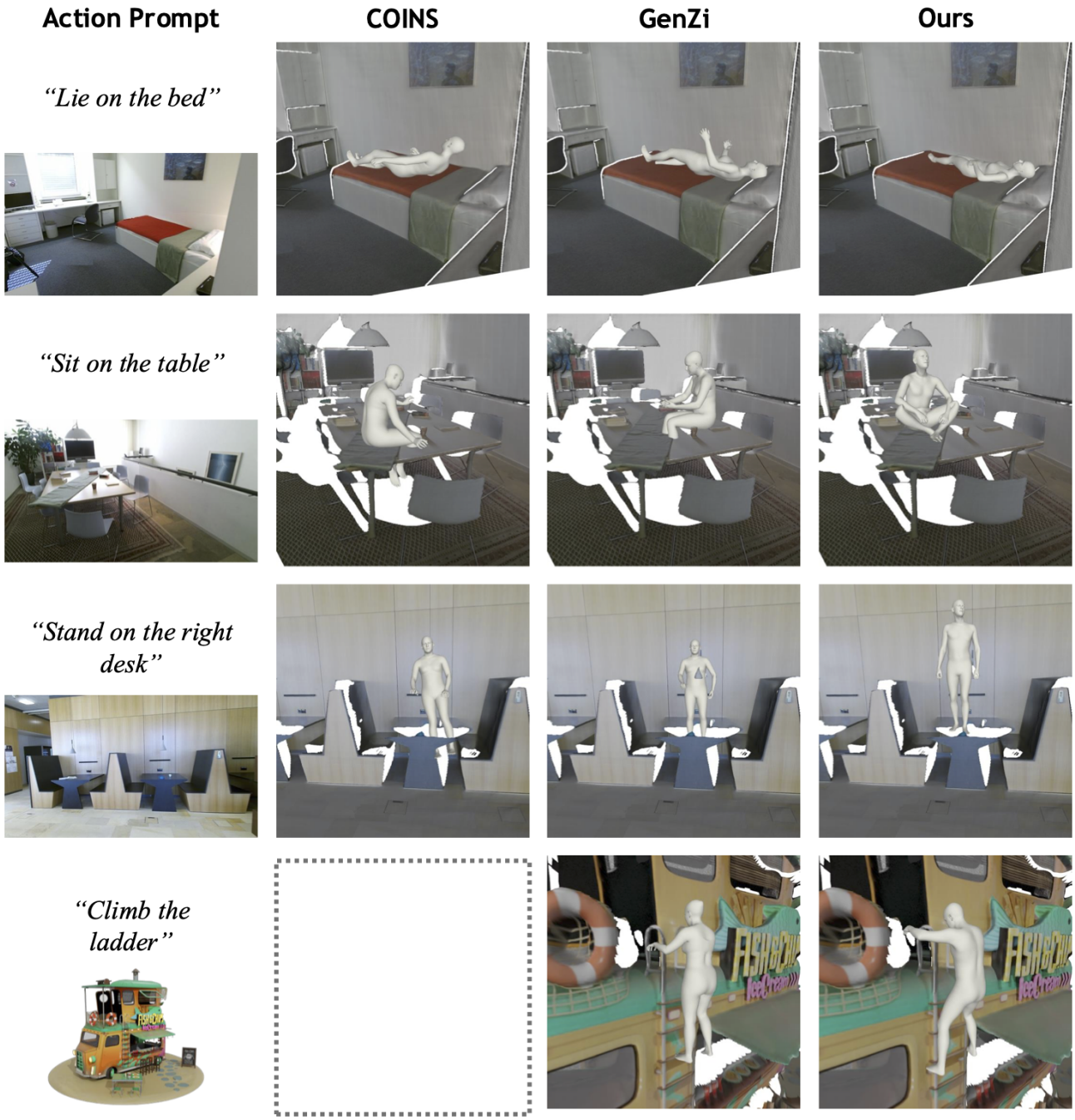

Figure 3. **3D Human-Object Interactions** GenHSI performs improved human object interactions even when we don't have access to accurate scene geometry. Our work also produces more plausible poses for lying, sitting, and standing. Prior works like GenZI have inconsistent multiview inpainting resulting in diverse but uncomfortable human poses as seen in lying down and sitting on table.

input about the environment, shown as Fig. 3. We still use MoGe [6] to reconstruct the coarse scene geometry. COINS [9] will use this incomplete scene reconstruction as its point net inputs. GenZI [3] will render the multi-view image based on the MoGe [6] output. Neither will execute the penetration terms in human joint pose optimization, since they rely on accurate scene geometry. The results show that our method is more robust based on object-centric human inpainting than GenZI [3], especially when the scenes are not well structured and well painted. Also, since we reconstruct the object, even when there is no visible affordance in the input single view, we can still generate plausible human interaction results.

References

[1] Realistic vision inpainting. https://huggingface.co/Uminosachi/realisticVisionV51_v51VAE-inpainting. 1

[2] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W.H. Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[3] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20465–20474, 2024. 2

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[5] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[6] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *ArXiv*, abs/2410.19115, 2024. 1, 2

[7] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024. 1

[8] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 1

[9] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022. 2