

PartCHOI: Part-Aware Guidance for Clothed Human-Object Interaction Generation

Mingwen Shao^{1,2}, Xinyuan Chen^{2*}, Qiao Zhang², Xiang Lv², Lingzhuang Meng², Qinglin Zhan², Chang Liu^{3,2}, and Chao Dong^{4,1}

¹ Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology, Shenzhen, China

² Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China
b24070011@s.upc.edu.cn

³ College of Computer Science, Fudan University, Shanghai, China

⁴ Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Abstract. Existing single-view 3D human generation methods remain limited in modeling clothed human-object interactions (CHOI), due to two key challenges: (1) the lack of reliable guidance for human-object spatial relationships resulting in implausible layouts, and (2) the absence of physical interaction modeling between entities leading to unrealistic contacts. To address these issues, we propose **PartCHOI**, a part-aware method that formulates the CHOI task as generating a human-centered scene with objects treated as parts of the human body. Specifically, to model spatial relationships between objects and the human, we design a Human-Guided Pose Estimator (H-GPE) that efficiently infers object-part poses by separating and localizing them within the coupled interaction scene. Furthermore, we devise an Object-Guided Refinement Module (O-GRM) to perform fine-grained human surface deformation for preventing penetration into objects, thus yielding realistic contact. Benefiting from the above design, PartCHOI generates plausible and detailed interactive scenes while avoiding lengthy layout optimization. Extensive experiments demonstrate our approach surpasses state-of-the-art methods in generation and interaction quality.

Keywords: Digital Humans · Human-Object Interaction · 3D Generation

1 Introduction

Modeling 3D digital humans and their interactions with other entities is of great importance, with applications in virtual reality, education, gaming, 3D

* Corresponding author.

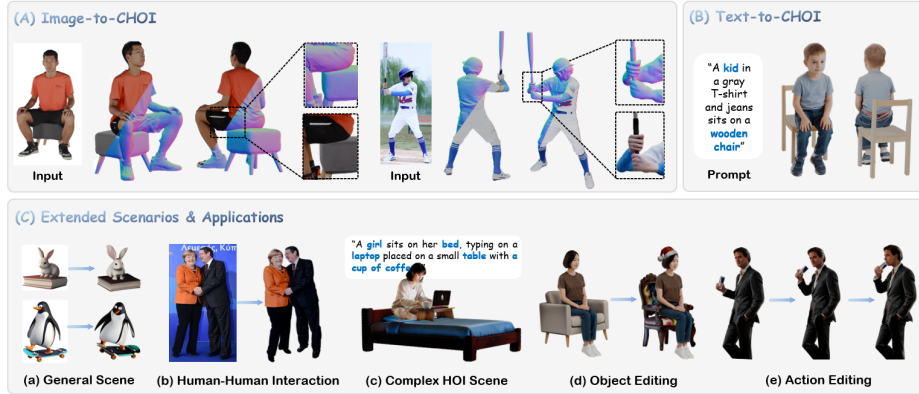


Fig. 1: We propose PartCHOI, a part-aware method for single-view 3D clothed human-object interaction (CHOI) generation. Our approach (A) achieves plausible layouts and natural contacts in image-to-CHOI scenes, (B) generates realistic scenes with complex interaction semantics from text prompts, and (C) supports diverse extended scenarios (a-c) and applications (d-e).

printing, and embodied intelligence [7, 32]. Compared with approaches based on videos [30, 55] or multi-view capture [12], generating 3D humans or inferring interactions from a single view is far more cost-effective, attracting increasing research attention. Existing single-view clothed human generation methods [17, 26, 49] achieve solid results on clean, non-interactive images. However, their capabilities are limited when modeling human-object interactions (HOI), i.e., cases involving handheld items, hats, or chairs that are ubiquitous in in-the-wild scenes. This challenging scenario is referred to as the single-view clothed human-object interaction (CHOI) generation task, which aims to generate high-quality clothed humans and objects while accurately modeling their interactions within the scene.

There are two closely related directions in solving this task. The first is monocular HOI reconstruction [23, 34], which directly infers interaction parameters from encoded image features. However, these methods typically rely on pre-defined object templates and parameterized human models. Such reliance, coupled with the scarcity of large-scale in-the-wild data, severely limits their generalization, preventing accurate modeling of complex poses and diverse object geometries. In addition, the absence of clothing and texture modeling further constrains their applicability. The second direction is compositional scene generation [4, 19]. These approaches generate individual entities separately and subsequently arrange them within a scene through layout optimization. While such methods can prevent obvious floating or penetration by imposing hard constraints, they remain effective only for simple cases. When extended to complex HOI scenarios, coarse spatial optimization fails to capture the fine-grained details of real interactions. Moreover, without explicit interaction supervision or spatial priors, the process tends to be both slow and unreliable.

In summary, existing methods remain limited for CHOI generation due to two key challenges: (1) the lack of reliable guidance for resolving complex human-object spatial relationships in real-world inputs leads to implausible layouts, and (2) the difficulty in modeling subtle physical interactions results in unnatural contacts, such as object penetration through human clothing.

To address these challenges, we propose PartCHOI, a single-view CHOI generation framework that requires no additional task-specific training. Our key insight is that CHOI can be reformulated as a human-centered scene, in which each interacting object is treated as a **part** of the human body. Based on this perspective, we design a part-aware mutual guidance mechanism that extends part-level 3D modeling [2, 18, 44] from single-object reconstruction to compositional scene generation. Specifically, to leverage the spatial relationships implicitly encoded within the scene, we design a Human-Guided Pose Estimator (H-GPE) for object-part pose estimation. Starting from a coarse voxel reconstruction of the scene, H-GPE uses the human body surface as a boundary to extract object-part voxels and efficiently estimate each object pose based on their voxel distribution. Furthermore, to model physical human-object contact, we devise an Object-Guided Refinement Module (O-GRM), employing a 3D Jacobian deformation field [1] to deform the human mesh. Using already placed object parts as anchors, O-GRM guides the deformation so that the human surface forms natural, close contact with object surfaces without interpenetration. Moreover, we introduce a Pose-Aligned Reconstruction (PAR) stage that aligns the human pose with its voxel structure, improving both reconstruction quality and the effectiveness of subsequent modules. Building on high-quality reconstruction of separate entities, PartCHOI eliminates the reliance on object templates and avoids unreliable layout optimization, yielding realistic and editable CHOI scenes.

Our main contributions are summarized as follows:

- We propose PartCHOI, a novel single-view CHOI generation paradigm that models interactions through a part-aware mutual guidance strategy.
- We deliberately design H-GPE, which enables stable layout inference by reformulating pose estimation as part localization in a coupled scene.
- O-GRM is further devised to prevent human-object penetration by refining contact through smooth surface deformation, producing more natural and realistic interactions.
- Extensive experiments are conducted to assess both appearance quality and interaction fidelity, demonstrating the effectiveness of the proposed method.

2 Related Work

2.1 3D Clothed Human Generation

Existing single-image 3D human generation methods [3, 26, 42, 57] typically employ feedforward networks to directly predict the 3D human representation.

While computationally efficient, these approaches often lack fine-grained realism and geometric detail. In contrast, recent methods such as MagicMan [10] and PSHuman [17] adopt a hybrid paradigm: they first synthesize multi-view images, then refine the geometry and texture through optimization-based reconstruction. This two-stage design offers a favorable trade-off between fidelity and efficiency compared with purely feedforward or fully optimization-based frameworks [9, 14, 52]. However, these methods only handle input images without interactions. When applied to HOI scenarios, their reliance on human-centered datasets and predefined human priors (e.g., SMPL-X [8]) results in incomplete geometric structures and unrealistic contact patterns. To address these limitations, our PartCHOI extends the two-stage generation paradigm to complex CHOI scenes through part-aware mutual guidance, enabling the joint generation of clothed humans and interactive objects.

2.2 HOI Reconstruction

Research on HOI reconstruction from text [6], multi-view [48], and sparse-view [33] inputs has recently been extended to unconstrained in-the-wild scenarios. Despite this progress, most single-view HOI methods [23, 34, 41, 50], which hold greater potential for practical applications, still depend on predefined object templates and HOI detection, producing only pose parameters for the human and objects. Recent works, such as HDM [40] and Open3DHOI [36], attempt to generalize HOI reconstruction to in-the-wild settings by removing template dependency. However, they still fail to explicitly model the appearance of clothed humans, and their interaction quality remains limited by training data. In contrast, our PartCHOI achieves realistic CHOI generation by performing layout inference and interaction refinement through part-aware strategies, without requiring HOI-specific training.

2.3 Compositional Scene Generation

General single-view image-to-3D generation methods [31, 37–39] typically treat the entire input image as a single object, which limits the editability of the generated 3D models and constrains their ability to handle multi-object scenes. To address these, recent works such as ComboVerse [4], Flash Sculptor [13], CAST [45] and Depr [54], disentangle individual objects via segmentation and occlusion recovery, and construct composite scenes by optimizing layouts. However, these methods are often slow and prone to local minima, leading to unstable results. In contrast, rather than interpreting the input as a complex multi-entity scene, our PartCHOI treats it as a human-centered structure composed of multiple parts, enabling efficient and reliable interaction modeling and generation.

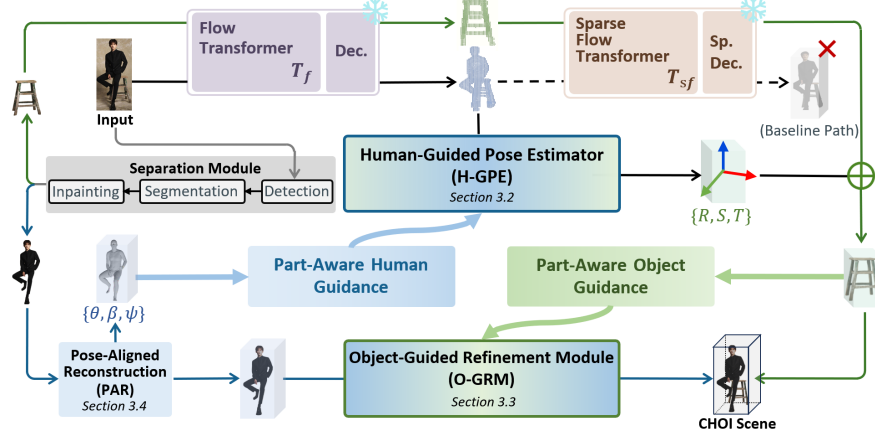


Fig. 2: Overview of PartCHOI. After separating the human and object images, the pipeline consists of three main branches: (1) the **object generation branch** generates object models; (2) the **pose estimation branch** infers object poses from scene voxels under part-aware human guidance (H-GPE); (3) the **human generation branch** reconstructs a pose-aligned human model (PAR) and further performs surface refinement under part-aware object guidance (O-GRM). Finally, the human and object models are integrated to form the final CHOI scene.

3 Method

3.1 Overview

Problem formulation. We begin by formulating the single-view CHOI generation task. Given a single input image I (for text input, I can be generated via a text-to-image method [24]), the goal is to generate a 3D scene $S = \{M_h, \{O_i\}_{i=1}^N\}$ that contains both the human mesh M_h and the interacting objects $\{O_i\}_{i=1}^N$, preserving accurate spatial layouts and interaction details. Interacting objects include backpacks, handheld items, seats, and other entities that exhibit explicit physical contact with the human. The human body is parameterized by SMPL-X [8], with parameters $\Theta_{\mathcal{H}} = \{\theta, \beta, \psi\}$. Each object pose is parameterized by $\Theta_{\mathcal{O}} = \{R, S, T\}$, representing rotation, scale, and translation.

Preliminaries. TRELLIS [39] is a native 3D generative model. Its core design introduces two rectified-flow Transformers: a flow Transformer T_f generates a sparse voxel structure S_v , while a sparse flow Transformer T_{sf} assigns unified structural latents (SLAT) to active voxels, which are then decoded by a Sparse Decoder (Sp.Dec.) into multiple 3D representations, such as meshes, 3D Gaussian Splatting (3DGS) [12], and NeRFs [22]. Formally, the SLAT z is represented as:

$$z = \{(z_i, p_i)\}_{i=1}^L, \quad z_i \in \mathbb{R}^C, \quad p_i \in \{0, 1, \dots, N-1\}^3, \quad (1)$$

where p_i denotes the spatial index of an active voxel on the 3D grid, and z_i is its associated local latent feature. N is the spatial resolution of the voxel

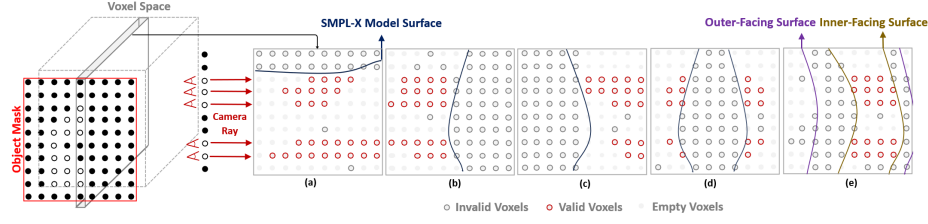


Fig. 3: Introduction to the voxel filtering strategy of H-GPE. Taking a z -axis slice of the voxel space as an example, each voxel’s position relative to the human’s exterior or interior is first determined by casting camera rays from front-view object-relevant pixels. Next, the occlusion relationships between objects and the human body are used to identify which non-empty voxels belong to object parts.

space. Intuitively, p_i captures the coarse spatial structure, while z_i enriches each voxel with geometry and appearance details. Through the generated models are relatively coarse, we observe that TRELLIS can (1) produce complete and spatially coherent interactive scenes, and (2) reasonably recover complex human poses. Therefore, we adopt it as a spatial and pose guidance backbone.

Pipeline. Based on the above formulation, we design PartCHOI, as illustrated in Fig. 2. We first introduce a *Separation Module* to decompose I into a human image I^h and a set of object images $\{I_i^o\}$. This module consists of three stages: leveraging a vision–language model (VLM) [24] to predict object categories and occlusion relations; employing a detection–segmentation model Grounded-SAM [28] to localize and separate human and object regions based on text prompts; and applying image inpainting techniques [21, 29] to each region to remove occlusions, producing clean separated images.

Subsequently, the generation pipeline proceeds as follows:

- (1) The object models are generated using the two-stage TRELLIS pipeline.
- (2) In the Pose-Aligned Reconstruction (PAR) stage, the optimized human parameters Θ'_h are obtained through iterative alignment and used as the reference pose for reconstructing the voxel-aligned clothed human mesh M_h .
- (3) Θ'_h and the voxelized scene S_v inferred from I are jointly fed into the Human-Guided Pose Estimator (H-GPE) to estimate the object poses Θ_o .
- (4) The object parts are placed into the human-centered scene according to Θ_o . M_h is further refined by the Object-Guided Refinement Module (O-GRM) to produce a part-aware human mesh, resulting in the final scene S .

The following sections detail the proposed H-GPE, O-GRM, and PAR.

3.2 Human-Guided Pose Estimator

To achieve efficient and robust pose estimation, H-GPE directly infers object spatial cues from the voxelized scene S_v . The activated voxels in S_v arise from both the human and object parts, implicitly encoding their spatial relationships.

Therefore, instead of adopting learning-based approaches [44] or relying on unreliable priors such as depth estimation, we devise a training-free filtering strategy that extracts object-specific voxel subsets from S_v for part-level localization. Concretely, we obtain a SMPL-X human body mesh M_b from Θ'_H , rescale it and center it within the voxel grid. The mesh surface then serves as geometric guidance: the activated voxels within the surface belong to the human body, while those outside are associated with potential object parts.

As shown in Fig. 3, for each object part O_i , camera rays are projected from the foreground pixels of its mask (segmented from I_i^o) and passed through the voxel grid. Based on occlusion relations inferred by the VLM, voxel assignments fall into five cases:

- (a) **No occlusion:** voxels along the ray are assigned to O_i .
- (b) **Human occludes object:** only voxels after the last intersection with the human surface are assigned to O_i .
- (c) **Object occludes human:** only voxels before the first intersection.
- (d) **Human enclosed by object:** voxels both before the first and after the last intersection, corresponding to wearable objects (e.g., hats, hula hoops).
- (e) **Object enclosed by human:** voxels between inner-facing surfaces (arms, hands), modeling held objects.

Following these rules, we obtain the voxel subset $\mathcal{V}_i$ corresponding to each object O_i . To estimate the object pose parameters $\Theta_{O,i}$, we align the voxels of the decoded voxel set $\mathcal{V}'_i$ from I_i^o with $\mathcal{V}_i$. This occlusion-free alignment can be addressed via a closed-form Kabsch alignment [15]:

$$\Theta_{O,i}^* = \arg \min_{\{R,S,T\}} \|\mathcal{V}_i - (SR\mathcal{V}'_i + T)\|_2^2. \quad (2)$$

Given scale S calculated from the segmentation mask area, this strategy requires only a single Singular Value Decomposition (SVD) to compute rotation R and translation T . Specifically, we first normalize both voxel sets $\mathcal{V}_i$ and $\mathcal{V}'_i$ by subtracting their centroids $\bar{v}$ and $\bar{v}'$. Then, a covariance matrix $\mathbf{H} = (\mathcal{V}'_i - \bar{v}')^\top (\mathcal{V}_i - \bar{v})$ is constructed to capture their spatial correspondence. Performing SVD on $\mathbf{H}$, i.e., $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, yields the optimal R and T in closed form:

$$R = \mathbf{V}\mathbf{U}^\top, \quad T = \bar{v} - S \cdot R\bar{v}'. \quad (3)$$

This formulation avoids iterative optimization and ensures stable alignment under partial visibility, enabling efficient and reliable estimation of object poses.

Discussion. Since SMPL-X mesh does not represent clothing, some voxels outside the mesh may correspond to garments rather than objects, which could seemingly introduce pose estimation errors. However, empirical observations show that clothing usually conforms closely to the body surface upon contact with objects, and even loose garments deform toward the body under pressure. Therefore, using the SMPL-X model as the interaction boundary helps position reconstructed objects naturally close to the body instead of floating, while also promoting realistic clothing deformation in subsequent processing.

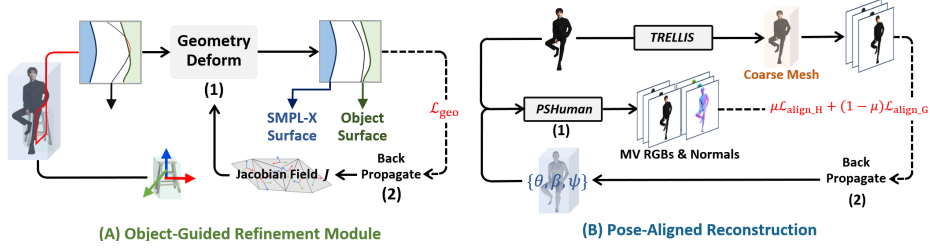


Fig. 4: Overview of O-GRM and PAR. (A) Refining human-object interactions using the Jacobian deformation field guided by object parts. (B) Progressively align human poses with the structure predicted by the generative model, along with multi-view predicted RGB images and normal maps.

3.3 Object-Guided Refinement Module

Although H-GPE achieves coarse-grained pose estimation, fine-grained penetration may still occur. To address this, O-GRM refines unreasonable contact details between the clothed human mesh $M_h = (V, T)$ and the object parts, where V denotes the vertices and T the triangular faces. Existing methods typically avoid penetration by adjusting poses, yet fail to model contact pressure and may compromise layout stability. In contrast, we apply surface deformations guided by a Jacobian field J [1, 46, 47], which enables localized adjustments while preserving global structural coherence. J is parameterized as a set of per-face Jacobians $J_t \in \mathbb{R}^{3 \times 3}$, with each J_t defined as:

$$J_t = \nabla_t \langle V_p \rangle, \quad (4)$$

where ∇_t denotes the gradient operator of triangle t . The deformed vertices V' are then obtained by solving the following Poisson equation:

$$V' = \arg \min_{V'} \|LV' - \nabla^\top \mathcal{A}J\|_2^2, \quad (5)$$

where $L \in \mathbb{R}^{|V| \times |V|}$ is the Laplacian operator, ∇ stacks the per-face gradient operators, and $\mathcal{A} \in \mathbb{R}^{3|T| \times 3|T|}$ is the mass matrix. This formulation preserves geometric plausibility while allowing smooth local deformations. Building on this, each optimization iteration consists of two steps, as shown in Fig. 4(A): (1) applying mesh surface deformations guided by J ; and (2) updating J via backpropagation using the following loss:

$$\mathcal{L}_{\text{geo}} = \lambda_1 \mathcal{L}_{\text{int}} + \lambda_2 \|J\|_1 + \lambda_3 \|LV'\|_2^2, \quad (6)$$

where $\mathcal{L}_{\text{int}}$ denotes the part-aware interaction loss. The key idea is to constrain each triangle to remain within the gap between the SMPL-X surface and the object surface, without penetrating either. Formally, the loss is defined as:

$$\mathcal{L}_{\text{int}} = \sum_{t \in \{t_r\}} [\max(0, -s_o(c_t)) + \max(0, -s_h(c_t))], \quad (7)$$

where $s_o(\cdot)$ and $s_h(\cdot)$ denote the pre-computed signed distance field (SDF) [25] values to the object surface and the Θ'_H surface, respectively. c_t is the centroid of triangle t . $\{t_r\}$ is the set of faces covered by the interaction region $\mathcal{R}$, which is precomputed by projecting the object parts onto the human body surface. In addition, the second term of Eq. 6 penalizes excessive local deformations, and the third term enforces global geometric smoothness. The hyperparameters λ_1 , λ_2 , and λ_3 are specified in the supplementary material.

Through this iterative surface refinement process, O-GRM (1) prevents the human mesh from intruding into objects, and (2) encourages clothing to conform naturally to the human body under object-induced pressure. Notably, the TRELLIS pose prior is designed for single-object generation and thus tends to produce tightly coupled structures, which suppress gaps but may cause penetrations. Therefore, O-GRM resolves only penetrations, without causing excessive deformation from forced contact.

3.4 Pose-Aligned Reconstruction

During the PAR stage, we align the poses Θ_H with the human voxels predicted by TRELLIS to ensure the validity of the voxel-based H-GPE. However, since the texture of the human model is derived from PSHuman [17]-generated multi-view images, naively aligning Θ_H with the TRELLIS mesh would introduce structure–texture misalignment. To address this, PAR performs alignment in a progressive manner. As illustrated in Fig. 4(B), each optimization iteration consists of two steps: (1) generating multi-view images from the PSHuman generator G_h ; and (2) updating Θ'_H via backpropagation using the following loss:

$$\begin{aligned}\mathcal{L}_{\text{align}} &= \mu \mathcal{L}_{\text{align-H}} + (1 - \mu) \mathcal{L}_{\text{align-G}}, \\ \mathcal{L}_{\text{align-H}} &= \lambda_n \mathcal{L}_n + \lambda_s \mathcal{L}_s, \quad \mathcal{L}_{\text{align-G}} = \lambda_s \mathcal{L}_s + \lambda_j \mathcal{L}_j,\end{aligned}\tag{8}$$

where $\mathcal{L}_n$, $\mathcal{L}_s$, and $\mathcal{L}_j$ denote the L_1 losses between the rendered outputs and the reference normal maps, silhouettes, and 2D keypoints [10], respectively. The references for $\mathcal{L}_{\text{align-H}}$ are obtained from G_h , whereas those for $\mathcal{L}_{\text{align-G}}$ are rendered from the TRELLIS-generated mesh. During optimization, the weight μ is linearly increased from 0 to 1, while λ_n , λ_s , and λ_j are detailed in the Supp.

In the early iterations, PAR updates Θ_H to progressively align the poses of images generated by G_h with those output by TRELLIS. As alignment improves, the supervision smoothly transitions to multi-view guidance from G_h to maintain appearance–pose consistency. Overall, the PAR stage not only enhances the reliability of pose estimation, but also mitigates structural inaccuracies in the reconstructed human model M_h caused by pose errors via optimizing Θ_H .

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate the performance of PartCHOI, we utilize three distinct data sources. (1) Internet-sourced in-the-wild HOI images are



Fig. 5: Appearance qualitative comparisons of single-image 3D generation results. The first two inputs are captured from in-the-wild scenes, while the latter two are taken from the DNA-Rendering [5] and THuman-Sit [51] datasets. All input images exhibit complex human poses and diverse human-object interactions.

collected exclusively for subjective evaluation, covering diverse interaction scenarios and object categories. (2) We select handheld-object interaction instances from *DNA-Rendering dataset* [5] to evaluate fine-grained interaction generation. (3) Moreover, we randomly select a subset from the *THuman-Sit dataset* [51], focusing on challenging human-seat interaction scenarios. Data from both datasets are used for quantitative and perceptual evaluations.

Evaluation Metrics. We adopt standard metrics for multi-view 3D reconstruction, including PSNR, SSIM [35], LPIPS [53], and CLIP Score [27]. Each generated scene is rendered from four viewpoints (0° , 90° , 180° , and 270°) to compute the above metrics. To further assess perceptual quality, we conduct a *user study* evaluating three aspects of multi-view rendering: appearance fidelity (A), layout consistency (L), and interaction realism (I). Each invited respondent is asked to select the best method for each of the three indicators. We additionally employ GPT-5 [24] as an automated evaluator to perform holistic judgment across all generated scenes and report the percentage of cases it selects as the best (G-S).

Moreover, to evaluate interpenetration in interactions, we first construct a human SDF over a volumetric grid that encloses the object. We then uniformly sample points within the object volume and evaluate the SDF value at each sampled location. The **Penetration Ratio (PR)** is defined as the ratio of points with negative SDF values to the total number of points.

Implementation Details. All experiments are implemented in PyTorch and executed on a single NVIDIA RTX A6000 (48GB) GPU. PAR and O-GRM are

Table 1: Quantitative and perceptual comparison with existing SOTA single-image 3D reconstruction methods. The first three methods are general-purpose, and the last four are human-focused. Among them, PSHuman and TRELLIS are baselines for generating humans and objects, respectively. Best results are in **bold**; second-best are underlined.

Method	User Study (%)			G-S (%)	CLIP	Reconstruction Quality		
	A	L	I			PSNR↑	SSIM↑	LPIPS↓
Wonder3D [20]	0	4.625	2.25	0	0.2637	15.52	0.6833	0.1956
Unique3D [38]	0	5.25	1.875	0	0.2695	15.93	0.7015	0.1716
Ouroboros3D [37]	0	1.5	0.75	0	0.2813	16.35	0.6933	0.1699
TRELLIS [39]	0	<u>33.25</u>	<u>19.375</u>	2.5	<u>0.2894</u>	17.28	0.7188	0.1513
SiTH [11]	1.25	2.25	1.875	0	0.2691	18.09	0.8219	0.1231
IDOL [57]	0.25	0	0.25	0	0.2564	17.73	0.7968	0.1584
MultiGO [49]	6.375	2.25	2.125	1.25	0.2758	18.38	0.8006	0.1330
PSHuman [17]	<u>23.125</u>	4.625	9.625	<u>3.75</u>	0.2766	<u>19.47</u>	<u>0.8420</u>	<u>0.1104</u>
PartCHOI (Ours)	69	46.25	61.875	92.5	0.2962	20.04	0.8532	0.0967

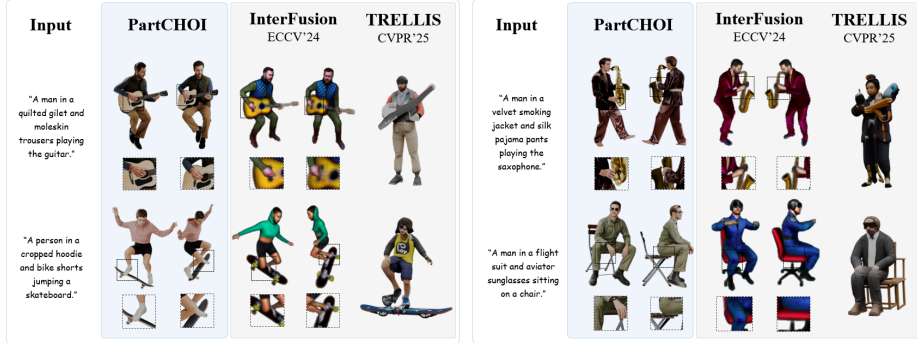


Fig. 6: Comparison of text-to-3D generation results on clothed human-object interaction scenes.

optimized for 700 and 100 iterations (experimentally determined), respectively, using the Adam optimizer [16]. For additional details, please refer to the Supp.

4.2 Comparisons

Baselines. We compare our method against two categories of approaches. The first includes state-of-the-art (SOTA) single-image or text-driven digital human generation methods, namely SiTH [11], IDOL [57], MultiGO [49], PSHuman [17], and InterFusion [6]. The second category comprises general text- or image-to-3D methods, including Wonder3D [20], Unique3D [38], Ouroboros3D [37] and TRELLIS [39].

Image-to-3D Comparison. As shown in Fig. 5, we compare the proposed method with SOTA approaches on multi-view rendering from a single input image. Among them, TRELLIS is capable of modeling interactions but lacks fine

Table 2: Runtime comparison between PartCHOI and optimization-based methods.

	LayoutDreamer [56]	Flash Sculptor [13]	H-GPE	O-GRM	PAR	Full Model
Time	65min	6min	7s	25s	1.5min	3min

details, while PSHuman excels in realism but struggles with modeling interactions in certain regions. Other methods exhibit significant shortcomings in terms of interactive relationships, geometric structures, or textural details. In contrast, PartCHOI not only generates highly realistic and view-consistent appearance, but also accurately recovers the spatial relationships and interaction details between entities, benefiting from H-GPE’s reliable object-part pose estimation and O-GRM’s fine-grained contact refinement. Furthermore, incorporating PAR enables the reconstructed human models to achieve more accurate and natural postures than those produced by PSHuman. Additionally, existing methods typically produce only coarse coupled models, whereas our approach separates the human and objects within the scene, better supporting downstream tasks. Quantitative results in Table 1 further confirm that our method outperforms existing approaches across both subjective and objective metrics, thanks to part-aware modeling of human-object interactions.

Text-to-3D Comparison. As shown in Fig. 6, we evaluate text-driven CHOI generation. Existing HOI generation methods capture spatial relationships to some extent but often lack fine-grained details. General text-to-3D approaches struggle to accurately model complex descriptions involving multiple entities and interaction semantics. By integrating the high-quality text-to-image generation capabilities of VLMs with the robust generation framework of PartCHOI, our method achieves higher visual fidelity and more accurate interaction details.

4.3 Efficiency Analysis

Tab. 2 shows the runtime analysis of our pipeline. In addition to the core modules, the separation stage takes approximately 10–40 seconds, and all other operations require no more than 30 seconds each. Overall, the total runtime is 160–180 seconds, which is substantially more efficient than iterative approaches. While real-time feedforward methods [26, 57] are faster, their performance degrades due to limited training diversity, hindering their ability to generalize to complex CHOI scenarios. The VRAM requirement of the entire pipeline is primarily determined by the human generation model. PartCHOI can switch between serial and parallel modes to accommodate different VRAM limits.

4.4 Ablation Study

Quantitative Ablation Analysis. As shown in Table 3, the results show that H-GPE serves as the core component of our framework, ensuring physically plausible object placement. Furthermore, as we prioritize synthesizing natural

Table 3: Further ablation studies are conducted on individual components of the PartCHOI framework. The baseline corresponds to directly placing both the reconstructed human and object at the scene center.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP $\uparrow$	PR($\%$) $\downarrow$
Baseline	17.25	0.7375	0.1438	0.2601	61.5
+ H-GPE	19.55	0.8429	0.1001	0.2927	10.4
w/ PAR w/o O-GRM	20.10	<u>0.8528</u>	0.0956	<u>0.2955</u>	9.1
w/o PAR w/ O-GRM	19.49	0.8520	0.0977	0.2932	6.5
Full model	<u>20.04</u>	0.8532	<u>0.0967</u>	0.2962	4.9



Fig. 7: Visualization of spatial layouts and corresponding **PR (%)** results under different optimization strategies. For a fair comparison, both O-GRM and PAR are disabled, and all experiments are conducted using fixed S .

and realistic scenes over fully reconstructing the input, enabling O-GRM slightly lowers some quantitative metrics but significantly reduces object penetration. Activating PAR not only facilitates structural alignment and improves part-pose accuracy but also enhances the fidelity of the reconstructed scenes.

The effectiveness of H-GPE. As shown in Fig. 5, the proposed H-GPE effectively incorporates TRELLIS’s modeling of the interaction between human and object parts, achieving reasonable layouts. Furthermore, as illustrated in Fig. 7, we explore alternative layout optimization strategies based on Score Distillation Sampling (SDS) loss [4], depth alignment loss [43], and multi-view supervision with TRELLIS-rendered images (M-VRGBs). However, these approaches converge more slowly and produce less accurate pose estimation. Quantitatively, H-GPE achieves the lowest penetration rate, highlighting its superiority in efficient and precise layout modeling.

The effectiveness of O-GRM. As visualized in Fig. 8(a), we present the structural variations in the reconstructed human model before and after applying O-GRM. Following object-guided surface deformation, the human garment remains closely aligned with the object’s surface, showing more natural and physically plausible variations in the interaction region. Moreover, both qualitative and quantitative results indicate that O-GRM substantially reduces mutual penetration between the human and objects, confirming its effectiveness in producing reasonable interactions. Additionally, as shown in the second row of Fig. 5 (the contact between the human’s shorts and the skateboard) and fourth row (the contact between the human’s pants and the chair), our method may not al-

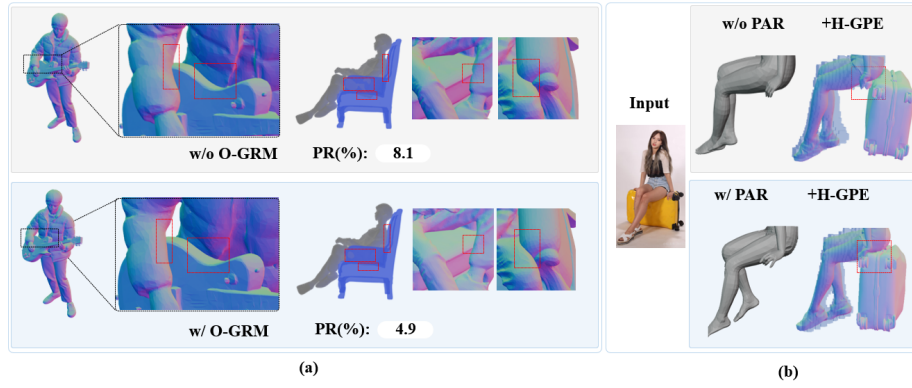


Fig. 8: (a)Effect of O-GRM on object-guided deformation. Normal details are visualized to clearly show structural variations. In the translucent projection map, gray regions within the blue object indicate human body parts intersecting the object. (b) Effect of PAR on pose refinement and spatial alignment. We overlay TRELIS voxel renderings with human surface normals to show spatial alignment.

ways faithfully match the input image under the influence of O-GRM. However, it ensures reasonable, non-penetrative contact, thereby maintaining the overall physical plausibility of the scene.

The effectiveness of PAR. As illustrated in Fig. 8(b), we visualize the structural changes of the human SMPL model and the final generated scene after applying PAR. The results show that the optimized SMPL-X parameters align more accurately with the poses in the input images, while preserving strong geometric consistency with the human voxels produced by TRELIS. In contrast, disabling PAR leads to inaccurate human poses and misaligned spatial arrangements of object parts, highlighting its essential role in refining pose and interaction consistency.

4.5 Applications

As shown in Fig. 1 (C), PartCHOI naturally supports diverse interactive extensions, with corresponding strategies outlined below:

- (a) Disable the human generation branch. The most central object is treated as the primary subject, with the remaining objects modeled as its parts.
- (b) Treat one human as the subject and the other as a part.
- (c) Process complex scenes hierarchically by first reconstructing sub-scenes and then assembling them.
- (d) Enable object replacement and addition by refining contacts via O-GRM and estimating poses via H-GPE.
- (e) Use video input to drive editing by reconstructing or deforming humans [26] frame by frame, updating object poses with H-GPE.

For specific operational steps for each scenario and application, please refer to the supplementary materials.

5 Conclusion

We propose PartCHOI, a method that addresses the challenging task of generating clothed human-object interactions. Unlike existing methods that rely on object templates, large-scale training data, or complex layout optimization, PartCHOI achieves joint generation and interaction modeling from a novel part-aware perspective. Specifically, our framework leverages human geometric priors to guide object-part pose estimation, and further utilizes object-part layouts to refine human surface deformations. With this design, PartCHOI enables efficient and accurate modeling of both coarse spatial layouts and fine-grained physical contacts. Experimental results demonstrate that our method outperforms prior approaches in generation quality and interaction detail. Furthermore, PartCHOI paves the way for a new paradigm in modeling and editing HOIs, with promising potential for future real-world applications.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62376285 and 61673396), the Innovation Fund Project for Graduate Students at China University of Petroleum (East China), and the Fundamental Research Funds for the Central Universities (No. 26CX04038A).

References

1. Aigerman, N., Gupta, K., Kim, V.G., Chaudhuri, S., Saito, J., Groueix, T.: Neural jacobian fields: Learning intrinsic mappings of arbitrary meshes. *arXiv preprint arXiv:2205.02904* (2022)
2. Chen, M., Shapovalov, R., Laina, I., Monnier, T., Wang, J., Novotny, D., Vedaldi, A.: Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 5881–5892 (2025)
3. Chen, W., Li, P., Zheng, W., Zhao, C., Li, M., Zhu, Y., Dou, Z., Wang, R., Liu, Y.: Synchronizing 2d and 3d generative models for single-view human reconstruction. *arXiv preprint arXiv:2510.07723* (2025)
4. Chen, Y., Wang, T., Wu, T., Pan, X., Jia, K., Liu, Z.: Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In: *Proc. of European Conference on Computer Vision*. pp. 128–146. Springer (2024)
5. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: *Proc. of International Conference on Computer Vision*. pp. 19982–19993 (2023)

6. Dai, S., Li, W., Sun, H., Huang, H., Ma, C., Huang, H., Xu, K., Hu, R.: Interfusion: Text-driven generation of 3d human-object interaction. In: Proc. of European Conference on Computer Vision. pp. 18–35. Springer (2024)
7. Fan, S., Huang, W., Cai, X., Du, B.: 3d human interaction generation: A survey. arXiv preprint arXiv:2503.13120 (2025)
8. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: Proc. of 3D Vision Conference. pp. 792–804 (2021)
9. Gao, X., Li, X., Zhang, C., Zhang, Q., Cao, Y., Shan, Y., Quan, L.: Context-human: Free-view rendering of human from a single image with texture-consistent synthesis. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 10084–10094 (2024)
10. He, X., Wu, Z., Li, X., Kang, D., Zhang, C., Ye, J., Chen, L., Gao, X., Zhang, H., Zhuang, H.: Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. In: Proc. of AAAI Conference on Artificial Intelligence. pp. 3437–3445 (2025)
11. Ho, I., Song, J., Hilliges, O., et al.: Sith: Single-view textured human reconstruction with image-conditioned diffusion. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 538–549 (2024)
12. Hong, J.G., Noh, S.Y., Lee, H.K., Cheong, W.S., Chang, J.Y.: 3d clothed human reconstruction from sparse multi-view images. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 677–687 (2024)
13. Hu, Y., Liu, S., Yang, X., Wang, X.: Flash sculptor: Modular 3d worlds from objects. arXiv preprint arXiv:2504.06178 (2025)
14. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. In: Proc. of 3D Vision Conference. pp. 1531–1542 (2024)
15. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography* **32**(5), 922–923 (1976)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., et al.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 16008–16018 (2025)
18. Liu, A., Lin, C., Liu, Y., Long, X., Dou, Z., Guo, H.X., Luo, P., Wang, W.: Part123: part-aware 3d reconstruction from a single-view image. In: Proc. of ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
19. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 7037–7047 (2025)
20. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 9970–9980 (2024)
21. Manukyan, H., Sargsyan, A., Atanyan, B., Wang, Z., Navasardyan, S., Shi, H.: Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In: Proc. of International Conference on Learning Representations. pp. 96301–96330 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/f0156a82b6af6a4e838923ce9c124424-Paper-Conference.pdf

22. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
23. Nam, H., Jung, D.S., Moon, G., Lee, K.M.: Joint reconstruction of 3d human and object via contact-based refinement transformer. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 10218–10227 (2024)
24. OpenAI: Gpt-5. <https://openai.com/gpt-5/> (2025), accessed: 2025-10-28
25. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 165–174 (2019)
26. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625* (2025)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *Proc. of International Conference on Machine Learning*. pp. 8748–8763. PMLR (2021)
28. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 10684–10695 (2022)
30. Tan, J., Xiang, D., Tulsiani, S., Ramanan, D., Yang, G.: Dressrecon: Freeform 4d human reconstruction from monocular video. In: *Proc. of 3D Vision Conference*. pp. 250–260 (2025)
31. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653* (2023)
32. Wang, R., Cao, Y., Han, K., Wong, K.Y.K.: A survey on 3d human avatar modeling—from reconstruction to generation. *arXiv preprint arXiv:2406.04253* (2024)
33. Wang, W., Xiao, J., Zhuang, Y., Chen, L.: Physics-aware human-object rendering from sparse views via 3d gaussian splatting. *arXiv preprint arXiv:2503.09640* (2025)
34. Wang, Z., Zheng, Q., Ma, S., Ye, M., Zhan, Y., Li, D.: End-to-end hoi reconstruction transformer with graph-based encoding. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 27706–27715 (2025)
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
36. Wen, B., Huang, D., Zhang, Z., Zhou, J., Deng, J., Gong, J., Chen, Y., Ma, L., Li, Y.L.: Reconstructing in-the-wild open-vocabulary human-object interactions. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 17426–17436 (2025)
37. Wen, H., Zhang, Y., Li, Z., Chen, J., Zhao, M., Wang, T., Liu, Z.: Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 12345–12355 (2025)
38. Wu, K., Liu, F., Cai, Z., Yan, R., Wang, H., Hu, Y., Duan, Y., Ma, K.: Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343* (2024)

39. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
40. Xie, X., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Template free reconstruction of human-object interaction with procedural interaction generation. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 10003–10015 (2024)
41. Xie, X., et al.: Chore: Contact, human and object reconstruction from a single rgb image. In: Proc. of European Conference on Computer Vision. pp. 125–145. Springer (2022)
42. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 512–523 (2023)
43. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 10371–10381 (2024)
44. Yang, Y., Zhou, Y., Guo, Y.C., Zou, Z.X., Huang, Y., Liu, Y.T., Xu, H., Liang, D., Cao, Y.P., Liu, X.: Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. arXiv preprint arXiv:2507.06165 (2025)
45. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. ACM Transactions on Graphics **44**(4), 1–19 (2025)
46. Yoo, S., Kim, K., Kim, V.G., Sung, M.: As-plausible-as-possible: Plausibility-aware mesh deformation using 2d diffusion priors. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 4315–4324 (2024)
47. Yu, Q., Li, X., Tang, Y., Han, X., Hu, L., Hao, Y., Chen, M.: Fancy123: One image to high-quality 3d mesh generation via plug-and-play deformation. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 595–604 (2025)
48. Zhan, Y., Zhu, Q., Niu, M., Ma, M., Zhao, J., Zhong, Z., Sun, X., Qiao, Y., Zheng, Y.: Tomie: Towards modular growth in enhanced smpl skeleton for 3d human with animatable garments. arXiv preprint arXiv:2410.08082 (2024)
49. Zhang, G., Yao, N., Zhang, S., Zhao, H., Pang, G., Shu, J., Wang, H.: Multigo: Towards multi-level geometry learning for monocular 3d textured human reconstruction. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 338–347 (2025)
50. 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: Proc. of European Conference on Computer Vision. pp. 34–51. Springer (2020)
51. Zhang, J., Zhang, Y., Zhang, H., Zhou, X., Zhou, B., Shao, R., Hu, Z., Liu, Y.: Ins-hoi: Instance aware human-object interactions recovery. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
52. Zhang, J., Li, X., Zhong, H., Zhang, Q., Cao, Y., Shan, Y., Liao, J.: Humanrefgs: Image-to-3d human generation with reference-guided diffusion and 3d gaussian splatting. IEEE Transactions on Circuits and Systems for Video Technology (2025)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of Computer Vision and Pattern Recognition Conference. pp. 586–595 (2018)
54. Zhao, Q., Zhang, X., Xu, H., Chen, Z., Xie, J., Gao, Y., Tu, Z.: Depr: Depth guided single-view scene reconstruction with instance-level diffusion. In: Proc. of International Conference on Computer Vision. pp. 5722–5733 (2025)

- 55. Zhao, Y., Wu, C., Huang, B., Zhi, Y., Zhao, C., Wang, J., Gao, S.: Surfel-based gaussian inverse rendering for fast and relightable dynamic human reconstruction from monocular videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
- 56. Zhou, Y., He, Z., Li, Q., Wang, C.: Layoutdreamer: Physics-guided layout for text-to-3d compositional scene generation. *arXiv preprint arXiv:2502.01949* (2025)
- 57. Zhuang, Y., Lv, J., Wen, H., Shuai, Q., Zeng, A., Zhu, H., Chen, S., Yang, Y., Cao, X., Liu, W.: Idol: Instant photorealistic 3d human creation from a single image. In: *Proc. of Computer Vision and Pattern Recognition Conference*. pp. 26308–26319 (2025)