InterFusion: Text-Driven Generation of 3D Human-Object Interaction

Sisi Dai[1] Wenhao Li[1] Haowen Sun[2] Haibin Huang[3] Chongyang Ma[3]
Hui Huang[2] Kai Xu[1⋆] Ruizhen Hu[2⋆]

[1] National University of Defense Technology
[2] Shenzhen University
[3] Kuaishou Technology
https://sisidai.github.io/InterFusion/

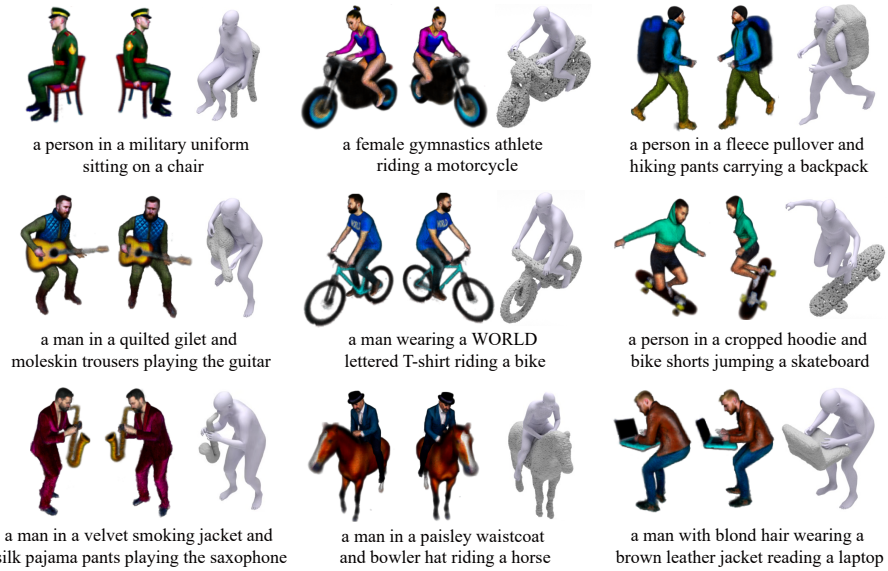


Fig. 1: Given a text prompt, our method InterFusion can generate diverse 3D scenes of a person interacting with an object.

Abstract. In this study, we tackle the complex task of generating 3D human-object interactions (HOI) from textual descriptions in a zero-shot text-to-3D manner. We identify and address two key challenges: the unsatisfactory outcomes of direct text-to-3D methods in HOI, largely due to the lack of paired text-interaction data, and the inherent difficulties in simultaneously generating multiple concepts with complex spatial relationships. To effectively address these issues, we present InterFusion,

⋆ Corresponding authors: kevin.kai.xu@gmail.com; ruizhen.hu@gmail.com.

a two-stage framework specifically designed for HOI generation. InterFusion involves human pose estimations derived from text as geometric priors, which simplifies the text-to-3D conversion process and introduces additional constraints for accurate object generation. At the first stage, InterFusion extracts 3D human poses from a synthesized image dataset depicting a wide range of interactions, subsequently mapping these poses to interaction descriptions. The second stage of InterFusion capitalizes on the latest developments in text-to-3D generation, enabling the production of realistic and high-quality 3D HOI scenes. This is achieved through a local-global optimization process, where the generation of human body and object is optimized separately, and jointly refined with a global optimization of the entire scene, ensuring a seamless and contextually coherent integration. Our experimental results affirm that InterFusion significantly outperforms existing state-of-the-art methods in 3D HOI generation.

Keywords: Text-Driven Generation · Zero-Shot Generation · 3D Human-Object Interaction Generation

1 Introduction

The generation of 3D human-object interactions (HOI) stands as a critical challenge in the fields of computer vision and computer graphics, with far-reaching implications in virtual reality, augmented reality, animation, and embodied AI [41, 49]. This task entails the creation of realistic 3D scenes where human figures interact with objects in ways that are not only physically plausible but also contextually relevant.

Despite its potential, the field has faced significant obstacles, primarily due to the scarcity of large-scale interaction data. Traditional approaches have predominantly relied on motion capture (mocap) datasets or physics-based simulations for generating these interactions. Mocap datasets [15, 26, 55] are limited by the specific scenarios they capture and are both costly and labor-intensive to produce. These limitations have resulted in a notable gap in generating diverse and contextually rich HOI scenes, especially for novel or complex interactions. Conversely, recent advancements have introduced text-to-3D methods [40,56], marking a significant shift in the field. These methods harness the power of textual descriptions to generate 3D objects without direct 3D supervision, presenting a novel approach to 3D content creation.

In this study, we explore text-to-3D method in HOI task within a zero-shot manner, *i.e.*, generating 3D scenes from textual descriptions using 2D diffusion models. Our key observations are two-fold: first of all, a direct application of text-to-3D method in HOI often leads to unsatisfied results like blurry textures and incorrect interactions, which is caused by the relative lack of paired text-interaction data during training and the difficulties of generating multiple concepts for diffusion-based methods [23]. Secondly, while collecting extensive interaction data is challenging, estimating human poses based on described in-

teractions is more feasible. These pose estimations can serve as geometric priors in the HOI generation process. By integrating pose estimation, our method substantially simplifies the text-to-3D process, particularly in the geometry optimization stage for human body generation. It also provides additional constraints for object generation, ensuring accurate placement and alignment with the text's semantic content. More importantly, we can separate the generation of human body and object, and jointly refine the details with the estimated pose, allowing for a more coherent and detailed synthesis of the interaction scene.

Inspired by these observations, we introduce InterFusion, a novel two-stage framework designed for HOI generation. Specifically, in the first stage, rather than relying on precise 3D interaction data, our approach instead collects a comprehensive dataset of synthesized images that depict a wide range of interactions. From these images, we employ advanced 3D pose estimation technique [8] to extract 3D human poses. Building upon these image-pose pairs, InterFusion develops a sophisticated codebook that establishes a mapping between interaction descriptions and 3D human poses, with the integration of the CLIP (Contrastive Language–Image Pretraining) embedding [43]. By leveraging these embeddings, our framework is able to interpret the nuances of interaction descriptions and translate them into accurate 3D pose representations. In the subsequent stage, InterFusion capitalizes on recent advancements in text-to-3D generation [40], as well as neural radiance fields [30], using the estimated human poses to produce 3D HOI scenes with realistic appearances and high-quality geometry. This stage operates in a 'local-global' manner. At the local level, the generation of the human body (SDS-H) and objects (SDS-O) is separately optimized, with the poses serving as additional constraints for the SDS. At the global level, the generation of the entire scene (SDS-I) is also guided by the integrated description and jointly optimized with SDS-H and SDS-O, ensuring a cohesive and contextually accurate representation of the HOI. Our experiments show that the quality of generation can be improved by a large margin and our approach outperforms state-of-the-art methods in HOI generation.

To summarize, our contributions are as follows:

- We introduce a novel two-stage framework InterFusion, for zero-shot 3D human-object interaction generation from text, incorporating 3D pose estimation as geometry priors.
- InterFusion leverages text-to-3D generation with a local-global optimization process. This strategy ensures seamless integration of human bodies and objects, producing realistic and high-quality 3D HOI scenes.
- InterFusion demonstrates significant improvements over existing methods in 3D HOI generation, showcasing its effectiveness in creating detailed, and contextually rich 3D interactions.

2 Related Work

2.1 Human-Object Interaction Synthesis

3D human-object interaction (HOI) generation is a challenging problem that has been studied widely by the computer vision and graphics community. Shape2Pose [18] generates plausible 3D human poses interacting with a given 3D object model by learning an affordance model from synthetic data. PiGraphs [47] learns the distribution between human poses and object arrangements from a collected dataset with 3D scene scans and RGB-D videos, generating the interaction snapshots given action specifications and object models. Recently, the parametric human body models such as SMPL [25,37,45] are employed in interaction synthesis to overcome the lack of realism due to human body representations. Benefiting from the PROX dataset [12] which consists of fitted SMPL models in captured 3D scenes, the more specific human-scene interaction has become an active research direction. Given a 3D scene, PSI [63] and PLACE [62] generate the 3D human body mesh represented as SMPL parameters through a conditional variational autoencoder [19, 51]. POSA [13] learns pose-specific priors to generate contacts conditioned by the given posed human, which can further guide the placement of the body mesh in a scene. COINS [64] enables semantic control on interaction synthesis by embedding the action label together with the interacted object as the condition of the generative model. More recently, fine-grained 3D interaction datasets [1, 7, 53] are captured to promote this field. Relying on these datasets, some methods [6, 38, 58], concurrent to this work, are proposed to explore text-guided 3D HOI generation. However, they remain constrained by distributions within the datasets. While previous methods need ground truth 3D interaction data as supervision, our work, for the first time, attempts to break through the limitation of data requirement. We generate a wider range of realistic and detailed 3D HOI scenes, including both indoors and outdoors.

2.2 Text-to-3D Content Synthesis

Early methods [3, 16, 24] for text-to-3D shape generation require paired data of 3D data and the corresponding textual descriptions to learn the joint embedding space of shape and text for supervision, which limits their generality to unseen object categories. Benefiting from large pre-trained text-to-image models and differentiable rendering techniques, breakthroughs in text-to-3D content generation have been achieved. For example, DreamFields [17] and PureCLIPNeRF [20] combine CLIP [43] with neural radiance fields (NeRF) [30], demonstrating the potential for zero-shot NeRF optimization. Meanwhile, CLIP-mesh [32] and Text2Mesh [28] incorporate CLIP to optimize the 3D mesh representation, starting from an initial sphere mesh and an input base mesh, respectively. Recently, DreamFusion [40] and SJC [56] enable NeRF optimization with guidance from pre-trained text-to-image diffusion models [44, 46] in place of CLIP, achieving more impressive results. To improve DreamFusion, Magic3D [21] proposes

a coarse-to-fine pipeline to generate the fine-grained mesh. TextMesh [54] extends the geometry representation from NeRF to an SDF framework, thereby enhancing detailed mesh extraction and photorealistic rendering. Among follow-up works, Latent-NeRF [27] and Vox-E [48] utilize explicit 3D shapes to provide additional training signals for NeRF optimization. While Latent-NeRF utilizes a rough untextured object for shape sculpture, Vox-E takes multiview images of a fine-grained textured object to edit geometry and appearance. All the methods mentioned above focus on the generation or edition of a single subject, while ignoring the interaction between different subjects.

2.3 Compositional Scene Generation

Representing scenes as compositions of object representations facilitates enhanced controllability. Numerous techniques incorporate additional information at the object level to perform compositional modeling of scenes, effectively separating object representations from the overall scene imagery. For instance, some methods incorporate 2D semantic information such as segmentation labels [65], instance masks [59,61], or features from a pre-trained vision-language model [31]. Some other methods [10, 36, 52] use 3D layout information by object-centric bounding boxes with canonical coordinates. When it comes to scene generation, [34, 35, 60] use compositional representations to generate scenes in a controllable manner. More recent approaches [4, 22, 39] generate compositional 3D scenes from input text prompts. Different from existing methods, our HOI scene is generated automatically without the requirement of input layout. Moreover, the spatial relationship between human and object during interaction is much more complex and cannot be simply characterized using their bounding boxes.

3 Our Method

3.1 Overview

In this section, we formally introduce InterFusion with a focus on the zero-shot and text-driven generation of 3D human-object interactions (HOI). Specifically, the input is a triplet of text descriptions, $T = \{T^H, T^O, T^I\}$, specifying the desired human style, object style, and interaction type. The goal is to generate a detailed 3D scene, $\psi = \{\psi^H, \psi^O\}$, comprising a human model and an object model that not only adhere to the specified appearance styles but also exhibit a tailored spatial relationship to accurately reflect the described interaction.

As illustrated in Figure 2, our method consists of two primary stages: anchor pose generation and anchor pose guided HOI generation. We first generate an interaction pose based on the input text, termed an anchored pose. This pose then serves as a geometric constraint, guiding the subsequent generation of detailed HOI. In the second phase, the human and object models are optimized separately and refined simultaneously with a global context, ensuring a cohesive and accurate representation of the interaction as described by the input text.

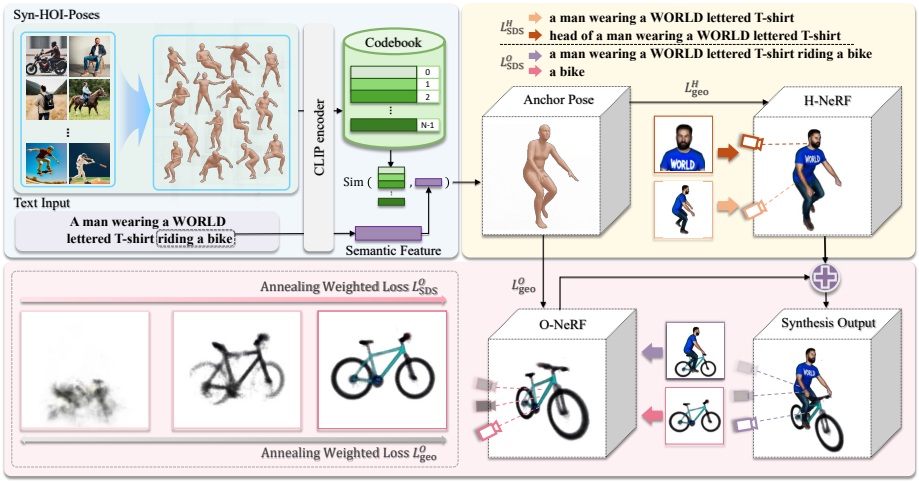


Fig. 2: InterFusion is a two-stage framework that transforms textual descriptions into detailed 3D human-object interactions, initially synthesizing anchor poses (upper left) and then optimizing the human model (upper right) and object model (bottom) with constraints from estimated pose and textual prompts.

3.2 Preliminaries

SDS. Score Distillation Sampling (SDS) has been introduced by DreamFusion [40]. While $x = g(\psi, \zeta)$, x is the 2D image rendered by a differentiable renderer g with model parameters ψ (the volumetric renderer and MLPs correspondingly in NeRF), under a desired camera pose ζ. By injecting the sampled noise ϵ into x at a time step t, the noisy image x_t is produced. The pre-trained 2D text-to-image diffusion model ϕ provides a denoising network $\hat{\epsilon}_\phi(x_t; y, t)$ that predicts the noise $\hat{\epsilon}$ given the noisy image x_t, time step t, and text embedding y. SDS then optimizes the model parameters ψ by minimizing the difference between the predicted noise and the added noise:

$$\nabla_\psi \mathcal{L}_{\text{SDS}}(\phi, x) = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon)\frac{\partial x}{\partial \psi}], \tag{1}$$

where $w(t)$ is the weighting term at the time step t.

3.3 Anchor Pose Generation

InterFusion starts by generating the anchor pose from the input text T^I, which is a non-trivial problem and highly limited by the available pose datasets. Existing datasets (*e.g.*, HumanML3D [9], BABEL [42]) are mocap based datasets and are typically conducted in a controlled laboratory environment, leading to lacks in action diversity and spatial coverage.

Recently, text-to-image diffusion models have shown the versatility and effectiveness of generating realistic and diverse images, through the integration of visual and linguistic feature spaces. We thus instead utilize advancements in this technology, specifically models like Stable Diffusion, to create a large-scale dataset of Human-Object Interaction (HOI) images. These images depict a wide range of human interactions. To ensure the diversity of interaction types, we utilize ChatGPT to generate prompts about human daily events or actions forming "verb-ing a/an/the object". A total of 235 result prompts are generated, covering most interactions in daily life. After filtering synthesized images without humans, we then estimate 3D human poses using the pre-trained PIXIE [8] model, creating a comprehensive Syn-HOI pose dataset consisting of a total 55K 3D pseudo-SMPL poses.

While SMPL model maps pose and shape parameters to a triangulated mesh, to align these poses with our text-to-interaction generation task, we render images from multiple perspectives and use CLIP's image encoder to derive pose feature embeddings. The average feature across these multiple perspective images represents the pose feature and attaches the dataset with pairs of averaged CLIP embedding and pose parameters. We further construct a codebook using K-Means clustering to identify key pose centroids based on the feature embeddings. 2,048 cluster centroids are clustered compositing our pose code-book, where each cluster comprises a subset of poses that represents similar interaction as the key poses of the centroids.

Once the codebook is built, for a given input text, we extract its feature embedding using CLIP's text encoder. This serves as a query to retrieve the most similar poses from the codebook, based on feature similarity. Specifically, given query text T^I, top k poses $\theta_k^{T^I}$ could be matched by pose embeddings θ_E, the averaged CLIP embeddings among rendered images from key poses:

$$\theta_k^{T^I} = \mathrm{TOP}_k \left(f_{text}(T^I), \theta_E \right) \tag{2}$$

where f_{text} is the text encoder of CLIP and $\mathrm{TOP}_k\left(X, Y\right)$ returns top k poses with the top k highest cosine similarities, and we use $k = 7$ for suitable poses in our experiments.

We then utilize GPT-4V to select the most precise pose as the final queried key pose. Depending on the requirement, we can instead select the poses sampled in the cluster, corresponding to the key pose, to guide the generation process for diverse results. This approach ensures rich and contextually aligned anchor poses for our text-to-interaction generation task.

To restrict the optimization of geometry and appearance using human structural priors from the acquired pose, we further incorporate COAP [29], a neural occupancy representation of the articulated human body based on SMPL parameters [25]. Given a pose θ and a shape β, COAP offers an occupancy prediction network $f(x; \beta, \theta)$ that maps a 3D query point p to an occupancy value, directly indicating whether the spatial point resides within the 3D body.

3.4 Pose-Guided HOI Generation

Once get the interaction pose, we further use it along with the input text to guide the generation of a detailed 3D HOI scene. In this phase, the estimated pose serves both as spatial constraints for the scene's geometry and as an anchor that aligns the human and object models. This approach ensures that the human model and the object model can be generated separately, while they are cohesively aligned to form the final 3D HOI scene.

The anchor pose provides specific spatial constraints for each component in the scene. For the human model, it establishes a basic geometric structure, while for the object model, it defines the areas that should remain unoccupied. This clear distinction is essential for rendering the scene both accurately and realistically. Simultaneously, the input text undergoes a complex processing procedure to offer distinct semantic guidance for each component. This is accomplished through various text conditioning techniques applied in SDS with the pre-trained DeepFloyd model [5]. The text is carefully crafted to direct the generation of both the human and object models to ensure all elements are in harmony with the semantic nuances of input descriptions.

Additionally, we introduce a novel camera tracing module to enhance the optimization process under varying text conditions. This module adaptively adjusts the camera pose, focusing on relevant elements within the scene at each optimization stage. This adaptive camera positioning is instrumental in ensuring that each aspect of the scene is optimally rendered according to the text descriptions, resulting in a more dynamic and contextually accurate 3D HOI scene.

We now provide details of each module in this stage, including the Neural Radiance Field representations for the human model (H-NeRF) and object model (O-NeRF), the camera tracing mechanism, and the guided optimization process.

H-NeRF. We use NeRF to represent the human model, noted as H-NeRF. The generation of H-NeRF is guided by the text description specifying the human style, conducting pose-specific human avatar generation with SDS. To enhance the quality of our renderings, particularly in terms of resolution, we incorporated a specialized optimization process that focuses on the head region of the human avatar. The location of the head for any given pose is determined with COAP [29]. This allows us to augment the text prompt specifically for the head region, using the notation * *the head of* *, to ensure that the head receives detailed attention during the generation process. The loss function for this optimization process is as follows, designed to balance the fidelity of the head region with the overall pose and style of the human figure.

$$\begin{aligned}\nabla_{\psi^H}\mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t^H; y^H, t) - \epsilon)\frac{\partial x^H}{\partial \psi^H}] \\ + \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t^{H,h}; y^{H,h}, t) - \epsilon)\frac{\partial x^{H,h}}{\partial \psi^H}],\end{aligned} \tag{3}$$

In shaping H-NeRF, our geometric constraint ensures that the human model evolves directly from the anchor pose. Points within the anchor are required to be occupied, establishing a firm base for the model. Meanwhile, points outside the anchor can also be occupied, but with probabilities that decrease as they move away from the anchor's surface. This approach allows for the addition of geometric details over the anchor, ensuring the model aligns with the human style described in the text while maintaining a coherent structure rooted in the anchor pose. The loss function used is as follows:

$$\begin{aligned}\mathcal{L}_{\text{geo}}^{H} &= CE_{p_i \in \mathbb{P}_{\text{in}}}(\alpha_i, f(p_i)) \\ &+ CE_{p_j \in \mathbb{P}_{\text{out}}}(\alpha_j, f(p_j))(1 - e^{-\frac{d}{2\eta^2}}),\end{aligned} \tag{4}$$

where d represents the point distance from the anchor surface, and η is the hyperparameter to control the extent of decaying, similar as in [27].

O-NeRF. We also employ NeRF to model the object component, referred to as O-NeRF. This model is designed to interact seamlessly with the human model and to embody the desired style as specified by the input text. To guide the generation of both the interaction type and the object style, we utilize SDS with tailored text prompts:

$$\begin{aligned}\nabla_{\psi^O} \mathcal{L}_{\text{SDS}} &= \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t^I; y^I, t) - \epsilon)\frac{\partial x^I}{\partial \psi^O}] \\ &+ \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t^O; y^O, t) - \epsilon)\frac{\partial x^O}{\partial \psi^O}],\end{aligned} \tag{5}$$

For the interaction scene x^I generation, we use an alpha-composited rendering to integrate H-NeRF and O-NeRF. This method calculates an alpha value from the density at each point, determining its contribution to the scene's color. Higher alpha values indicate a greater influence on the rendering, allowing for a nuanced and realistic integration of the human and object models in the final interaction image:

$$\begin{aligned}x^I &= \sum_i w_i^I c_i^I, w_i^I = \alpha_i^I \prod_{j=1}^{i-1}(1 - \alpha_j^I), \\ c_i^I &= \frac{\alpha_i^H}{\alpha_i^H + \alpha_i^O} c_i^H + \frac{\alpha_i^O}{\alpha_i^H + \alpha_i^O} c_i^I.\end{aligned} \tag{6}$$

In our NeRF-based approach, alpha values, capped at 1, are derived from density for composite rendering. This setup allows semantic guidance gradients, conditioned by the interaction image, to optimize the density and color of both the human and object models. However, we noted a tendency for the model to prioritize object generation at the expense of the human model's quality. To counteract this, we implemented gradient truncation towards the human model to maintain a balanced optimization between the two components.

For O-NeRF's geometric constraints, we aim to prevent occupancy of points within the anchor pose, denoted as $x_i \in \mathbb{X}_{in}$, by the object model. Inspired by the physical collision prevention concept, we define a specific loss function that ensures these points remain unoccupied by O-NeRF, effectively preventing overlap between the human and object models in the 3D space:

$$\mathcal{L}^{O}_{\text{geo}} = CE_{p_i \in \mathbb{P}_{\text{in}}}(\alpha_i, 1 - f(p_i)) \tag{7}$$

By minimizing the above loss term, model parameters are optimized to enhance the geometric consistency between H-NeRF and O-NeRF.

Camera tracing. To enhance the generation of O-NeRF, we introduce a dynamic camera tracing module within the SDS framework. This module automatically adjusts the camera pose to focus on the target's center, either the entire interaction scene or just the object during object generation. The camera aligns with the average position of voxels with an occupancy probability above 0.5, ensuring a consistent focus on the most significant parts of the scene or object for optimal detail capture and realism.

Optimization. The total loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}^{H}_{\text{SDS}} + \lambda_1 \mathcal{L}^{O}_{\text{SDS}} + \mathcal{L}^{H}_{\text{geo}} + \lambda_2 \mathcal{L}^{O}_{\text{geo}} + \lambda_3 L_{\text{reg}}, \tag{8}$$

where λ_1, λ_2 and λ_3 are the corresponding loss weights. Figure 2 (left bottom) shows several intermediate states of the object model during the generation steps and we can see that the object is gradually generated with fine details. Weight annealing is adopted during the guided optimization process. We leave the details in the supplement.

4 Experiments

Figure 3 (left) shows results with various interaction poses, showcasing the strength of our method. Our method can support diverse interaction poses within a single type. Meanwhile, with interaction type and pose fixed, our method can further generate more numerous results under different human styles or object styles, as shown in Figure 3 (right). Moreover, our InterFusion also supports controllable text-conditioned editing, providing users more control over the already generated 3D models, which are presented in the supplementary materials.

We both qualitatively and quantitatively evaluate our method against alternative baseline methods. To verify the effectiveness of individual components in our method, we conduct ablation studies. Furthermore, we discuss the application potential, limitations and future work of our InterFusion. The implementation details, more evaluations, and further discussions are presented in the supplementary materials.

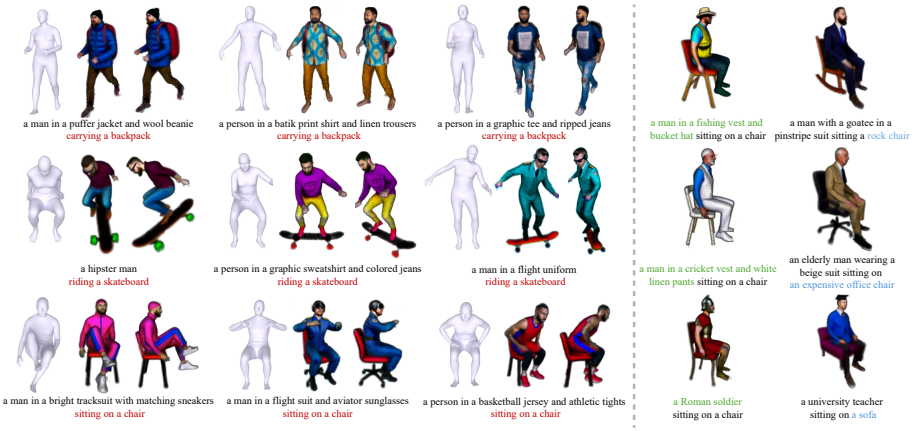


Fig. 3: More results generated by InterFusion. Diverse integration poses are supported. Numerous human and object styles are also supported.

Evaluation prompts and criteria. Similar as the first stage for generating prompts about various interaction types, we also use ChatGPT to randomly generate prompts about human styles and collect the text prompts for evaluation. We select 61 distinct and diverse prompts, ensuring the classes of Human (*e.g.*, a policewoman, a teenager) with various styles (*e.g.*, in a graphic tee and ripped jeans), Objects (*e.g.*, a motorcycle, a guitar) and Interactions (*e.g.*, riding, playing) are reasonable and evenly distributed. In total, we have 13 different types of interactions, covering contact areas across the whole body, which demonstrates the effectiveness of our method on diverse types of interactions compared to baseline methods. We calculate the CLIP scores between the input text prompts and different views of generated 3D human-object interactions, and then compare the means based on different evaluation prompts. The CLIP score measures the similarity between a prompt text for an image and the actual content of the image. We also provide an assessment using GPT-4V for selection, named GPT-4V select, which evaluates the completeness of objects and correctness of physical interactions across multiview rendered images. We refer to the supplementary material for assessment details and additional evaluation criteria.

Baseline approaches. To the best of our knowledge, the proposed method is among the first to generate 3D human-object interactions based on text inputs in a zero-shot manner. We thus compare our method to alternative text-to-3D methods, including DreamFusion [40], Magic3D [21], and TextMesh [54]. Additional comparisons with MVDream [50] and ProlificDreamer [57] are presented in the supplementary material. To justify our key idea of using human pose to guide the generation, we design an object-centric baseline (Ours-OC) as a variant of our method. Instead of using human pose as the prior, we retrieve an object with the given semantic category from ShapeNet dataset [2] and use it as

Table 1: Quantitative evaluation of CLIP score and GPT-4V select.

Method	DreamFusion [40]	Magic3D [21]	TextMesh [54]	Ours-OC	Ours-HC
CLIP score	0.3027	0.3179	0.2761	0.3203	**0.3308**
GPT-4V select(%)	8.20	11.48	1.64	13.11	**65.57**

geometry constraints to guide both human and object generation. As official implementations are unavailable for some of these baselines, we use the third-party re-implementations provided by threestudio [11] for a fair comparison. Note that all re-implementations use multi-resolution hash-grid [33] for 3D representation and DeepFloyd [5] for guidance.

4.1 Comparison Results

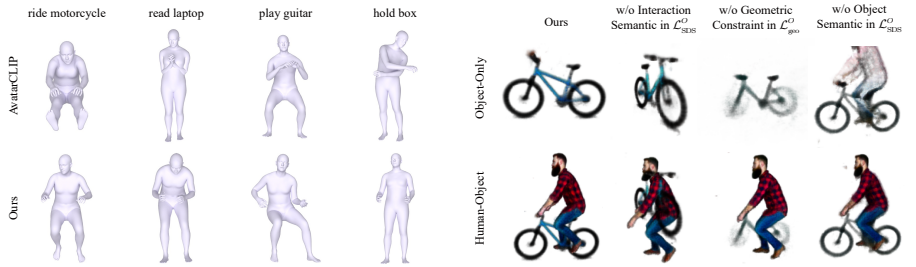
Qualitative evaluations. Some representative visual comparisons are shown in Figure 4, and additional comparisons including those with our object-centric baseline are presented in the supplementary material. The overall results of the baselines reveal common uncertainties in both geometry and appearance attributed to the confusion introduced by multi-concept guidance. Specifically, baseline models may lean towards one specific target, as seen in the example of "a man wearing a red baseball cap" with only the red baseball cap generated, the example of "a policewoman" with only the upper half of the human body generated, or the example of "a man in a puffer jacket and wool beanie" where the shopping cart is failed to be generated. Even with a relatively uniform attention distribution, the baseline model may encounter challenges in producing complete results, as demonstrated in the saxophone and violin examples (the 3rd and 4th column in Figure 4), which exhibit deficiencies such as incomplete human body parts, and mixing of human legs. Moreover, the baseline model sometimes struggles to effectively choose the focus of generation when the input text describes a less common target like "a person in a military uniform", resulting in poor outcomes.

Our approach overcomes these issues by conducting optimization in an explicit decomposed way and intelligently guiding attention from SDS jointly in spatial and semantic aspects. Therefore, our method achieves more stable and higher-quality 3D results under multiple-concept guidance.

Quantitative evaluations. The comparison results are shown in Table 1. We can see that our method (Ours-HC) achieves the best performance compared to baselines, showcasing our results exhibit more 3D plausibility with given text prompts. Among the baselines, Ours-OC gets the best performance. Though with object priors, the object-centric approach still does not provide sufficient body priors for human generation, resulting in its inability to achieve complete interaction generation. Magic3D gets the next best performance among the baselines, as it usually generates the whole interaction scene comparing to other two



Fig. 4: Qualitative comparison results with baselines. InterFuion generates more stable and higher-quality results and is more consistent with input interaction descriptions.



(a) Visual comparison between AvatarCLIP and our pose generation.

(b) Visual ablation for loss terms during the pose-guided generation process.

Fig. 5: Qualitative results of ablation studies.

baselines, however, the results are somewhat vague as shown in Figure 4, thus there is still a large performance gap comparing to our method.

The significant enhancement in performance can be attributed to our method's effective concurrent generation of both human and object, along with cohesive interactive information.

4.2 Ablation Studies

We conduct several ablation studies both qualitatively and quantitatively to show the importance of our design as well as introduced loss terms. Results are presented in Figure 5a, Figure 5b and Table 2. We refer to the supplementary material for additional results and details.

Table 2: Quantitative results of ablation studies.

Settings	w/o $\mathcal{L}_{\text{SDS}}^{I}$	w/o $\mathcal{L}_{\text{SDS}}^{O}$	w/o $\mathcal{L}_{\text{geo}}^{O}$	Ours
CLIP score	0.3164	0.3293	0.3171	**0.3308**
GPT-4V select(%)	1.64	4.92	16.39	**77.05**

Syn-HOI-pose. The anchor pose obtained at the first stage provides the geometry constraints for our HOI generation. We thus first demonstrate the effectiveness of synthesizing anchoring poses using our pseudo-pose dataset. We compare the pose generated with the existing approach AvatarCLIP [14], which utilizes CLIP only to retrieve the queried pose from the codebook constructed from the mocap dataset AMASS [26]. Figure 5a shows that the proposed pose generation stage enables a better-fitted anchor pose for input interaction types, which demonstrates our pseudo-poses, reconstructed from synthesized images, has the potential for more diverse poses' requirements than existing mocap datasets.

Semantic guidance. To better illustrate the efficacy of semantic guidance in our NeRF-based generation, we conduct ablations separately on SDS from object and interaction. As shown in Figure 5b and Table 2, in the absence of SDS from object, semantic consistency for the object is compromised, resulting in noisy outcomes influenced by the human, thus the contact region can not be extracted correctly. SDS from interaction plays the most crucial role in the performance, making results more semantically and visually plausible. The absence of SDS from interaction leads to generated objects being confined to the remaining space, but without interaction with the human.

Geometric constraint. Relying solely on semantic guidance is inadequate for achieving the final objectives, as the object should be generated in spaces outside the human body, with its generation targets not positioned at the origin. As illustrated in Figure 5b and Table 2, depending only on semantic guidance may yield results that appear spatially conflict with the human body. Without the spatial constraint, the object generation at the origin would additionally have a conflict with the interaction objective, thus resulting in degenerated objects and final interactions.

5 Conclusion

In conclusion, our work presents InterFusion, a novel framework for zero-shot 3D human-object interaction generation. InterFusion tackles the challenges of limited 3D interaction data and the complexity of generating multiple concepts simultaneously. Our two-stage approach, which synthesizes 3D interaction poses from text and then uses these poses as geometric anchors for detailed HOI scene generation, has demonstrated significant improvements over existing methods.

Acknowledgements. We thank Zejia Su and Yunfan Ye for helpful discussions and paper proofreading. We also thank the anonymous reviewers for their valuable comments. This project was supported by the NSFC (62325211, 62132021, 62322207), Guangdong Natural Science Foundation (2021B1515020085), Shenzhen Science and Technology Program (RCYX20210609103121030), and the Major Program of Xiangjiang Laboratory (23XJ01009).

References

1. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15935–15946 (2022)
2. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
3. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 100–116. Springer (2019)
4. Cohen-Bar, D., Richardson, E., Metzer, G., Giryes, R., Cohen-Or, D.: Set-the-scene: Global-local training for generating controllable nerf scenes. arXiv preprint arXiv:2303.13450 (2023)
5. Deepfloyd IF. https://github.com/deep-floyd/IF, 2023
6. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19888–19901 (2024)
7. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Articulated objects in free-form hand interaction. arXiv preprint arXiv:2204.13662 (2022)
8. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: 2021 International Conference on 3D Vision (3DV). pp. 792–804. IEEE (2021)
9. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
10. Guo, M., Fathi, A., Wu, J., Funkhouser, T.: Object-centric neural scene rendering. arXiv preprint arXiv:2012.08503 (2020)
11. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. `https://github.com/threestudio-project/threestudio` (2023)
12. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2282–2292 (2019)
13. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14708–14718 (2021)

14. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022)
15. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
16. Jahan, T., Guan, Y., Van Kaick, O.: Semantics-guided latent space exploration for shape generation. In: Computer Graphics Forum. vol. 40, pp. 115–126. Wiley Online Library (2021)
17. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
18. Kim, V.G., Chaudhuri, S., Guibas, L., Funkhouser, T.: Shape2Pose: Human-centric shape analysis. ACM Transactions on Graphics (Proc. SIGGRAPH) (Aug 2014)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Lee, H.H., Chang, A.X.: Understanding pure clip guidance for voxel grid nerf models. arXiv preprint arXiv:2209.15172 (2022)
21. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. arXiv preprint arXiv:2211.10440 (2022)
22. Lin, Y., Bai, H., Li, S., Lu, H., Lin, X., Xiong, H., Wang, L.: Componerf: Text-guided multi-object compositional nerf with editable 3d scene layout. arXiv preprint arXiv:2303.13843 (2023)
23. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
24. Liu, Z., Wang, Y., Qi, X., Fu, C.W.: Towards implicit text-guided 3d shape generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17896–17906 (2022)
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
26. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
27. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. arXiv preprint arXiv:2211.07600 (2022)
28. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13492–13502 (2022)
29. Mihajlovic, M., Saito, S., Bansal, A., Zollhoefer, M., Tang, S.: Coap: Compositional articulated occupancy of people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13201–13210 (2022)
30. 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)

31. Mirzaei, A., Kant, Y., Kelly, J., Gilitschenski, I.: Laterf: Label and text driven object radiance fields. In: European Conference on Computer Vision. pp. 20–36. Springer (2022)
32. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022)
33. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
34. Nguyen-Phuoc, T.H., Richardt, C., Mai, L., Yang, Y., Mitra, N.: Blockgan: Learning 3d object-aware scene representations from unlabelled images. Advances in neural information processing systems **33**, 6767–6778 (2020)
35. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)
36. Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F.: Neural scene graphs for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2856–2865 (2021)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
38. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023)
39. Po, R., Wetzstein, G.: Compositional 3d scene generation using locally conditioned diffusion. arXiv preprint arXiv:2303.12218 (2023)
40. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
41. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., et al.: Habitat 3.0: A co-habitat for humans, avatars and robots. arXiv preprint arXiv:2310.13724 (2023)
42. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 722–731 (2021)
43. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
45. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
46. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

47. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
48. Sella, E., Fiebelman, G., Hedman, P., Averbuch-Elor, H.: Vox-e: Text-guided voxel editing of 3d objects. arXiv preprint arXiv:2303.12048 (2023)
49. Sheridan, T.B.: Human–robot interaction: status and challenges. Human factors **58**(4), 525–532 (2016)
50. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
51. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)
52. Song, Y., Kong, C., Lee, S., Kwak, N., Lee, J.: Towards efficient neural scene graphs by learning consistency fields. arXiv preprint arXiv:2210.04127 (2022)
53. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 581–600. Springer (2020)
54. Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., Tombari, F.: Textmesh: Generation of realistic 3d meshes from text prompts. arXiv preprint arXiv:2304.12439 (2023)
55. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 601–617 (2018)
56. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. arXiv preprint arXiv:2212.00774 (2022)
57. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36** (2024)
58. Wu, Q., Shi, Y., Huang, X., Yu, J., Xu, L., Wang, J.: Thor: Text to human-object interaction diffusion via relation intervention. arXiv preprint arXiv:2403.11208 (2024)
59. Wu, Q., Liu, X., Chen, Y., Li, K., Zheng, C., Cai, J., Zheng, J.: Object-compositional neural implicit surfaces. In: European Conference on Computer Vision. pp. 197–213. Springer (2022)
60. Xu, Y., Chai, M., Shi, Z., Peng, S., Skorokhodov, I., Siarohin, A., Yang, C., Shen, Y., Lee, H.Y., Zhou, B., et al.: Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4402–4412 (2023)
61. Yang, B., Zhang, Y., Xu, Y., Li, Y., Zhou, H., Bao, H., Zhang, G., Cui, Z.: Learning object-compositional neural radiance field for editable scene rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13779–13788 (2021)
62. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: Place: Proximity learning of articulation and contact in 3d environments. In: 2020 International Conference on 3D Vision (3DV). pp. 642–651. IEEE (2020)

63. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6194–6204 (2020)
64. Zhao, K., Wang, S., Zhang, Y., Beeler, T., Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: European Conference on Computer Vision. pp. 311–327. Springer (2022)
65. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15838–15847 (2021)