ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation

Hongjie Li* Hong-Xing Yu* Jiaman Li Jiajun Wu

Stanford University

Figure 1. **Zero-shot human-scene interaction motion synthesis.** Our zero-shot method distills 4D interactions from video generation models to generate natural HSIs in various 3D environments. We demonstrate our method's effectiveness on real-world scenes reconstructed from the Mip-NeRF 360 (Garden, Bicycle, Room) and Tanks and Temples (Truck) datasets, showcasing diverse interactions with both static environments (walking, sitting, cleaning car) and dynamic objects (watering plants, lifting vase, operating mower, playing guitar).

Abstract

Human-scene interaction (HSI) generation is crucial for applications in embodied AI, virtual reality, and robotics. Yet, existing methods cannot synthesize interactions in unseen environments such as in-the-wild scenes or reconstructed scenes, as they rely on paired 3D scenes and captured human motion data for training, which are unavailable for unseen environments. We present ZeroHSI, a novel approach that enables ***zero-shot*** *4D human-scene interaction synthesis, eliminating the need for training on any MoCap data. Our key insight is to distill human-scene interactions from state-of-the-art video generation models, which have been trained on vast amounts of natural human movements and interactions, and use differentiable rendering to reconstruct human-scene interactions. ZeroHSI can synthesize realistic human motions in both static scenes and environments with dynamic objects, without requiring any ground-truth motion data. We evaluate ZeroHSI on a curated dataset of different types of various indoor and outdoor scenes with different interaction prompts, demonstrating its ability to generate diverse and contextually appropriate human-scene interactions. Project page:* https://awfuact.github.io/zerohsi.

1. Introduction

Generating realistic human motions that interact with 3D environments is fundamental to computer graphics, VR/AR, embodied AI, and robotics. Humans constantly engage with their surroundings through both static interactions—sitting

*Equal contribution. Work was done while H. Li was a visiting student at Stanford University. H. Li is now with Peking University.

on chairs, lying on sofas, leaning against ladders—and dynamic interactions, such as watering plants, playing musical instruments, or manipulating objects. These interactions are remarkably diverse and pervasive in our daily lives, encompassing countless objects and countless ways of interacting with them. Despite significant advances in motion synthesis, realistically simulating this wide spectrum of human-scene interactions remains a fundamental challenge.

Prior work in human-scene interaction synthesis primarily follows two directions. The first focuses on interactions with static 3D scenes [25], with recent advances driven by motion diffusion models [33, 34] trained on paired scenes and motion capture data. While these models can generate realistic motions for common activities like navigation and sitting, they struggle to generalize even within action categories. The second direction explores manipulation of dynamic objects [48, 49, 63], showing success in generalizing within object categories but failing to handle significant geometric variations. Both approaches share a fundamental limitation: **they cannot synthesize interactions in unseen environments** (such as in-the-wild scenes or reconstructed scenes), as they rely on paired 3D scene and motion capture data, which are unavailable for unseen scenarios. This motivates us to address the problem of *zero-shot human-scene interaction*, synthesizing HSIs in various scenes without training, and eliminating the need for 3D HSI data.

Our key insight is to distill human-scene interactions from state-of-the-art video generation models. These models have been trained on vast amounts of video data, capturing a wide range of natural human movements and interactions in diverse environments. This allows us to generate contextually appropriate human motions for various 3D scenes, whether synthesized or reconstructed. For example, given a reconstructed 3D garden scene, our method can generate natural motions of a person watering plants in the garden (Figure 1).

Our approach, ZeroHSI, enables zero-shot 4D human-scene interaction synthesis by integrating video generation and neural human rendering. In a nutshell, ZeroHSI takes a 3D scene as input, initializes an animatable human avatar in the scene, generates a video of the human interacting with the scene, and then extracts the interaction motion via differentiable neural rendering. ZeroHSI can synthesize appropriate human motions in both static scenes and environments containing dynamic objects, without requiring any ground truth motion data. Our design leverages pretrained video generation models, maintaining compatibility with various generation approaches and allowing seamless integration of future advancements in video generation. Our primary contributions can be summarized as follows:

- We introduce the novel task of zero-shot HSI motion generation, addressing the fundamental limitation of requiring paired motion-scene training data.
- We propose ZeroHSI, which integrates video generation with differentiable human rendering to tackle the challenging task, supporting both static scenes and environments with dynamic objects.
- We curate a dataset of various indoor and outdoor scenes with different interaction prompts and plausible initial human poses to evaluate zero-shot HSI. We demonstrate that ZeroHSI can generate diverse and contextually appropriate human-scene interactions across these environments.

2. Related Work

Text-Guided Motion Generation. The availability of large-scale motion capture datasets like AMASS [56], enhanced with action labels and language descriptions through BABEL [66] and HumanML3D [20], has enabled significant advances in language-guided motion synthesis [20, 65, 78]. Early approaches demonstrated success using VAE architectures [21, 64]. More recently, diffusion models have emerged as a powerful framework for motion generation [3, 11, 32, 47, 48, 67, 70, 71, 80, 95, 99], leading to various text-conditioned approaches [14, 37, 79, 97]. In contrast to these methods that focus on generating isolated human motions, our work synthesizes contextually appropriate human-scene interactions.

Human-Scene Interaction Synthesis. Research in human-scene interaction has progressed along multiple directions. With paired scene-motion datasets [1, 23, 24, 85, 102] and object-motion data [25, 98], researchers have developed methods [1, 25, 43, 82, 83, 98] for generating scene-aware human motions like sitting and reaching. In the domain of object manipulation, research has evolved from primarily focusing on hand motion synthesis [13, 96, 101] to incorporating full-body motions [8, 17, 51, 75, 87] with the fuel of full-body interaction datasets [15, 74]. Building on paired human-object motion data [5, 48, 81], recent methods focus on predicting interactions from past states [81, 90] or object motion sequences [48]. While these approaches have shown promising results, they typically require paired motion-scene data. In contrast, we focus on zero-shot interaction synthesis, eliminating the need for scene-specific motion capture data while supporting diverse interaction in both static and dynamic environments. Besides interaction motion synthesis, prior work also explored zero-shot generation of static interactions at a single time step [39, 50]. These methods are complementary to our work as their output static poses can be used as the input initial poses in our approach.

Another line of work leverages reinforcement learning [45, 88] to train scene-aware policies for navigation and interaction in static 3D scenes, as well as object manipulation such as lifting and moving [26, 57, 89]. These approaches can be trained with relatively small amounts of motion data. However, they are limited to specific interaction types and restricted in generalization to diverse scenes and objects.

Video Generation. Recent advances in diffusion models

have revolutionized video generation [2, 6, 7, 9, 18, 29, 30, 73], enabling high-quality generation of human actions [22] and scene dynamics [9]. Video generation models can now take both text and image as input conditions [76, 92], where text descriptions guide the overall motion and actions while image conditions provide scene context and geometry, controlling fine-grained interaction details. These models have demonstrated remarkable capabilities in synthesizing temporally coherent videos, and recent efforts have extended video generation models to have increased controllability in terms of motion [22, 86] and camera viewpoints [28]. Our work bridges the gap between generating realistic 2D videos and synthesizing 3D human motions. By integrating video generation as a drop-in module, our approach directly benefits from the rapid progress in video generation quality and controllability.

Neural Rendering for Humans. Neural rendering techniques have significantly advanced the synthesis of realistic human appearances [12, 36, 42, 44, 46, 54, 61, 62, 84]. Building on the success of neural radiance fields (NeRF) [58], NeuralActor [53] uses texture maps defined on the SMPL model [55] to guide the learning of deformable radiance fields. NeuralBody [62] utilizes structured latent codes linked to SMPL [55] for novel view synthesis from sparse multi-view videos. AnimatableNeRF [61] introduces a neural blend weight field and achieves superior novel view and novel pose synthesis results. Additionally, efficiencies have been significantly improved for rendering avatars [35] through multiresolution hash encoding representation [59]. Recent studies have explored 3D Gaussian splatting [38], an explicit and efficient representation for modeling animatable humans [16, 31, 35, 52, 94, 103]. Animatable Gaussians [52] leverages powerful 2D StyleGAN-based CNNs and 3D Gaussian splatting to create high-fidelity avatars from multi-view RGB videos. In our work, we leverage this differentiable and controllable representation to optimize SMPL parameters using an image-matching loss, bridging the gap between 2D video generation and 3D human motion.

3. ZeroHSI

Formulation. Our goal is to generate a plausible 3D human-scene interaction motion sequence $\tau = \{(\mathcal{M}_t, \mathbf{P}_t)\}_{t=1}^{T}$, conditioned on a 3D scene $\mathcal{S}$, an interactable dynamic object $\mathcal{O}$, a text prompt describing the interaction c, and the initial states of human and objects τ_0, where $\mathcal{M}_t$ represents the human pose in frame t, $\mathbf{P}_t \in \mathbb{R}^6$ represents the object's 6D pose in frame t, and T represents the length of the sequence. We use SMPL-X [60], a widely used parameterized human model, to represent the human pose. In each frame, $\mathcal{M}_t$ consists of root translation $\mathbf{r}_t$, global orientation $\boldsymbol{\phi}_t$, and body pose $\boldsymbol{\Theta}_t$.

Overview. We present an illustration of our zero-shot human-scene interaction motion generation method in Fig. 2. To address the diversity and generalizability issues encountered by most existing learning-based methods, we first generate HSI video using off-the-shelf video generation models [76] (Sec. 3.2). We then reconstruct the generated 2D HSI video into 4D human-scene-interaction sequence (Secs. 3.3 and 3.4). In addition, we incorporate a refinement process using human pose priors to improve the results (Sec. 3.5).

The main challenge lies in converting the generated 2D HSI video into 4D interaction motion. While existing methods [10, 72, 93] can estimate plausible human motions, they have two key limitations for human-scene interaction scenarios. First, they struggle to estimate precise root translations, which often leads to penetration artifacts between the human and 3D scenes. Second, these methods do not address the reconstruction of object motions that are crucial for interaction with dynamic objects. To reconstruct plausible 4D HSI from a single video, we propose an optimization-based method using differentiable rendering. Specifically, we represent scene and object with 3D Gaussians [38], and leverage a Gaussian Avatar model [52] to map human poses to a set of Gaussian particles (Sec. 3.1). We denote the Gaussian scene, object and human as $\mathcal{G}_{\mathcal{S}}$, $\mathcal{G}_{\mathcal{O}}$, and $\mathcal{G}_{\mathcal{H}}$, respectively. We then optimize the camera pose, human pose, and object's 6D pose through rendering loss. Additionally, we incorporate techniques to handle low-quality generated videos, ensuring robust performance with imperfect input footage.

Discussion. We justify the advantages of using off-the-shelf video generation model. Our method remains agnostic to video generation approaches, enabling compatibility with diverse models and easy integration of future advancements. By focusing on 2D-to-4D HSI lifting rather than being tied to a specific video generation model, we maintain flexibility to leverage ongoing research. We demonstrate this adaptability across different models in Appendix C.9.

3.1. Preliminaries

3D Gaussian Splatting. 3DGS [38] is an explicit 3D representation comprising a set of Gaussian particles, each characterized by its position $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{1}$$

Each particle is parameterized by its position $\boldsymbol{\mu}$, opacity α, rotation $\mathbf{r}$, scale $\mathbf{s}$, and color $\mathbf{c}$. The covariance matrix $\boldsymbol{\Sigma}$ is computed by $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top$, where $\mathbf{R}$ is the rotation matrix constructed from $\mathbf{r}$, and $\mathbf{S} = \mathrm{diag}([s_x, s_y, s_z])$ is the scaling matrix. The 3D Gaussians are splatted onto 2D plane during rendering, and the color is calculated by alpha-blending of N ordered particles overlapping the pixel,

$$\mathbf{C} = \sum_{i=1}^{N} \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)\mathbf{c}_i. \tag{2}$$

We use $\mathcal{R}$ to represent this rendering process of 3DGS.

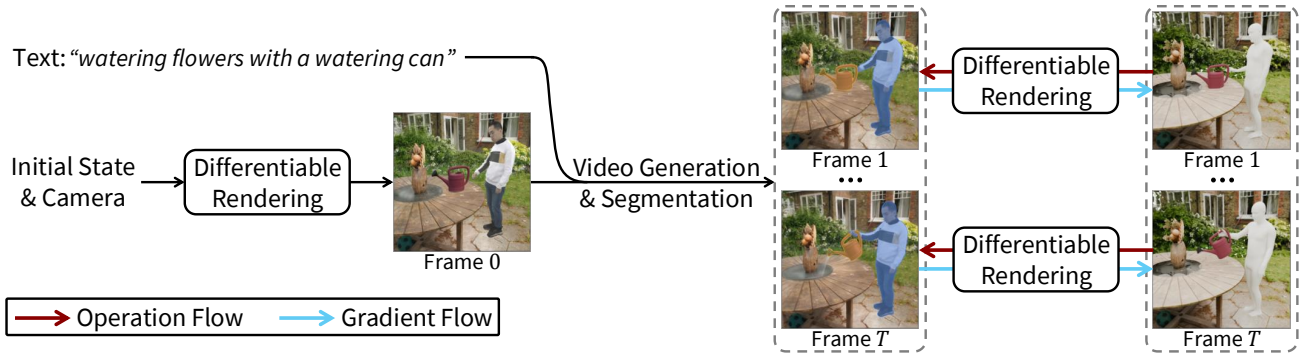

Figure 2. **Overview of ZeroHSI.** Our approach begins with HSI video generation conditioned on the rendered initial state and text prompt. Through differentiable neural rendering, we optimize per-frame camera pose, human pose parameters, and object 6D pose by minimizing the discrepancy between the rendered and generated reference videos.

Animatable Gaussians. Animatable Gaussians [52] is a neural human rendering method that maps SMPL-X parameters [60] to an animatable avatar represented by 3DGS [38]. The appearance of the posed avatar can be differentiably rendered to an image $\mathbf{I}$ as in Eq. (2), with a given camera view $\mathbf{T} \in \mathbb{R}^{4\times4}$. Given the driving SMPL-X pose $\mathbf{\Theta}$, Animatable Gaussians first deforms the character-specific template using a linear blend skinning (LBS) function. It then predicts pose-dependent Gaussian maps conditioned on the driving pose $\mathbf{\Theta}$ and the camera view $\mathbf{T}$:

$$\{\Delta\boldsymbol{\mu}_p\}_{p=1}^{P} \leftarrow \mathcal{F}_{position}(\mathbf{\Theta}), \tag{3}$$

$$\{\alpha_p, \mathbf{s}_p, \mathbf{r}_p\}_{p=1}^{P} \leftarrow \mathcal{F}_{other}(\mathbf{\Theta}), \tag{4}$$

$$\{\mathbf{c}_p\}_{p=1}^{P} \leftarrow \mathcal{F}_{color}(\mathbf{\Theta}, \mathcal{V}(\mathbf{T})), \tag{5}$$

where $\Delta\boldsymbol{\mu}_p$ is the offset position relative to the deformed template, and $\mathcal{V}$ is the view direction feature extractor. The root translation $\mathbf{r}$ and the global orientation ϕ are finally applied to $\{\boldsymbol{\mu}_p, \mathbf{s}_p\}_{p=1}^{P}$ to achieve global transformation. We denote the mapping from human pose and camera view to Gaussian particles proposed by Animatable Gaussians as

$$\mathcal{G}_{\mathcal{H}} \leftarrow \mathcal{A}(\mathbf{r}, \phi, \mathbf{\Theta}; \mathbf{T}). \tag{6}$$

3.2. HSI Video Generation and Processing

Given the initial human pose $\mathcal{M}_0 = (\mathbf{r}_0, \phi_0, \mathbf{\Theta}_0)$ and an additional camera pose input $\mathbf{T}_0$, we first map them to the initial human Gaussians $\mathcal{G}^0_{\mathcal{H}}$ through Eq. (6). The initial object 6D pose $\mathbf{P}_0$ is then applied to the Gaussian object $\mathcal{G}_{\mathcal{O}}$, denoted as $\mathcal{G}^0_{\mathcal{O}} = \mathcal{G}_{\mathcal{O}}(\mathbf{P}_0)$. As depicted in Fig. 3, the resulting Gaussians are concatenated with the Gaussian scene $\mathcal{G}_{\mathcal{S}}$ and rendered into the initial HSI image $\mathbf{I}_0$ through Eq. (2) under camera view $\mathbf{T}_0$.

The HSI video $\{\mathbf{I}_t\}_{t=0}^{T}$ is generated through the KLING image-to-video model [76], conditioned on the rendered initial frame $\mathbf{I}_0$ and text prompt c. We employ Segment Anything Model 2 (SAM2) [69] to segment dynamic foreground (human and object) and static background within the generated video for subsequent uses. The prompt points for SAM2 in the initial frame are automatically proposed through farthest point sampling on non-occluded regions of both human and object. We denote the segmentation masks for human and object as $\{\mathbf{M}^t_{\mathcal{H}}\}_{t=0}^{T}$ and $\{\mathbf{M}^t_{\mathcal{O}}\}_{t=0}^{T}$, respectively.

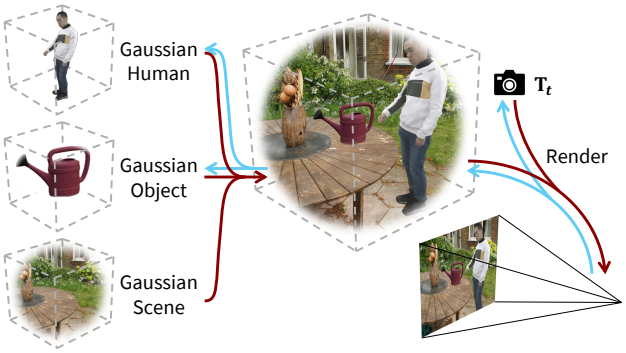

Figure 3. **Illustration of the differentiable rendering process.** The parameterized Gaussian human, transformed Gaussian object, and static Gaussian scene are concatenated and rendered through Gaussian rasterization.

Handling Occlusions. We address the challenge of under-constrained pose parameters caused by body part occlusions by minimizing occlusions in the generated video. Since camera movement remains minimal over short durations (e.g., 5 seconds), we simply select an initial camera view $\mathbf{T}_0$ from a pre-defined camera array that maximizes visibility of both the human and dynamic object in the first frame. Details of the camera view selection process are in Appendix C.3.

3.3. Camera Pose Estimation

We sequentially estimate the relative camera transformation between nearby frames for camera pose estimation. In frame t, we apply a learnable transformation $\mathbf{T} \in \mathbb{R}^{4\times4}$ to the Gaussian scene $\mathcal{G}_{\mathcal{S}}$, denoted as $\mathcal{G}_{\mathcal{S}}(\mathbf{T})$. We render image with the estimated camera pose $\mathbf{T}_{t-1}$, and optimize the relative transformation $\mathbf{T}$ with the photometric loss between frame t, focusing on the static background region:

$$\mathbf{T}_* = \arg\min_{\mathbf{T}} \mathcal{L}_2\Big(\mathcal{R}\big(\mathcal{G}_{\mathcal{S}}(\mathbf{T}); \mathbf{T}_{t-1}\big) \odot \mathbf{M}_t, \mathbf{I}_t \odot \mathbf{M}_t\Big), \tag{7}$$

where $\mathbf{M}_t$ represents the mask of static background. The camera pose for frame t is calculated by $\mathbf{T}_t = \mathbf{T}_*^{-1}\mathbf{T}_{t-1}$.

Handling Incorrect Contents. Video generation models produce incorrect contents in regions initially occluded by humans or objects that later become visible (Appendix D.1). Instead of using only the current frame's static background mask, we address this by aggregating dynamic foreground masks from frames 0 to t to eliminate incorrect contents:

$$\mathbf{M}_t = \mathbf{1} - \bigcup_{i=0}^{t} \left(\mathbf{M}_{\mathcal{H}}^i \cup \mathbf{M}_{\mathcal{O}}^i\right). \tag{8}$$

3.4. HSI Optimization

The optimization of 4D human-scene interactions is carried out sequentially on a frame-by-frame basis. For object 6D pose estimation, we also optimize the relative transformation $\mathbf{P} \in \mathbb{R}^6$ between nearby frames. We optimize the human pose with a different strategy, as optimizing relative transformations leads to rapidly increased cumulative errors, particularly when the generated video exhibits quality issues such as body part disappearance (Appendix D.2). We directly optimize the human pose parameters $\mathcal{M}_t = (\mathbf{r}_t, \boldsymbol{\phi}_t, \boldsymbol{\Theta}_t)$, where the root translation $\mathbf{r}_t$ and global orientation $\boldsymbol{\phi}_t$ are initialized using their respective values $\mathbf{r}_{t-1}$ and $\boldsymbol{\phi}_{t-1}$ from the previous frame. Additionally, the body poses $\{\boldsymbol{\Theta}_t\}_{t=1}^T$ are initialized using independent frame-wise estimates from a pose estimation model [10].

Similar to Sec. 3.2, we render the Gaussians human, object, and scene at frame t through

$$\begin{aligned}\hat{\mathbf{I}}_t &= \mathcal{R}(\mathcal{G}_{\mathcal{H}}^t, \mathcal{G}_{\mathcal{O}}^t, \mathcal{G}_{\mathcal{S}}; \mathbf{T}_t) \\ &= \mathcal{R}\big(\mathcal{A}(\mathbf{r}_t, \boldsymbol{\phi}_t, \boldsymbol{\Theta}_t; \mathbf{T}_t), \mathcal{G}_{\mathcal{O}}(\mathbf{P}_t), \mathcal{G}_{\mathcal{S}}; \mathbf{T}_t\big),\end{aligned} \tag{9}$$

where $\mathbf{P}_t$ denotes the object's 6D pose in frame t, computed as the composition of the relative transformation $\mathbf{P}$ and the previously optimized object pose $\mathbf{P}_{t-1}$.

Using the generated video as the reference, we optimize the human pose parameters $\mathcal{M}_t$ with the photometric loss in each frame:

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1(\hat{\mathbf{I}}_t, \mathbf{I}_t) + \lambda\mathcal{L}_{\text{D-SSIM}}(\hat{\mathbf{I}}_t, \mathbf{I}_t). \tag{10}$$

We introduce two additional loss terms to enhance the accuracy of object 6D pose optimization. For frame t, we compute the center point position $C_{\mathcal{O}}^t \in \mathbb{R}^2$ of the object's segmentation mask $\mathbf{M}_{\mathcal{O}}^t$ and the center point position of the rendered object region $\hat{C}_{\mathcal{O}}^t$. These positions are normalized to the range [0,1], and the object center point position loss is defined as

$$\mathcal{L}_{\text{center}} = \mathcal{L}_2(\hat{C}_{\mathcal{O}}^t, C_{\mathcal{O}}^t). \tag{11}$$

A depth regularization term is incorporated based on the assumption that the object's depth remains relatively constant throughout a short time window (e.g., 5 seconds), and thus, the object depth within the time window should be close to the depth value at the first frame. The average object depth in the first frame is computed as

$$D_{\mathcal{O}}^0 = \frac{\texttt{sum}(\mathbf{D}_0 \odot \mathbf{M}_{\mathcal{O}}^0)}{\texttt{sum}(\mathbf{M}_{\mathcal{O}}^0)} \in \mathbb{R}, \tag{12}$$

where $\mathbf{D}_0$ represents the rendered depth map in frame 0. The average depth of the rendered object region $\hat{D}_{\mathcal{O}}^t$ is calculated similarly, and we define the depth regularization loss as

$$\mathcal{L}_{\text{depth}} = \mathcal{L}_2(\hat{D}_{\mathcal{O}}^t, D_{\mathcal{O}}^0). \tag{13}$$

The object's relative transformation $\mathbf{P}$ is optimized using the composite loss function:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}. \tag{14}$$

Handling Appearance Change. Video generation models do not guarantee a consistent human appearance (Appendix D.3), which can potentially compromise human pose optimization. Given that human appearance exhibits continuous changes in the generated video, we fine-tune the color net defined in Eq. (5) during optimization using $\mathcal{L}_{\text{rgb}}$.

3.5. Refinement

Limited supervision from single-view video poses challenges for natural human motion reconstruction. We refine our results in the latent space of VPoser [60], a variational human pose prior trained on the AMASS dataset [56], which preserves robust human pose priors. For each frame, the reference joint positions $\hat{J}_t$ are computed by passing the raw human pose parameters obtained through Sec. 3.4 to the SMPL-X layer [60]. We then optimize the root translation $\mathbf{r}_t$, global orientation $\boldsymbol{\phi}_t$, and the body pose latent $\mathbf{z}_t \in \mathbb{R}^{32}$ with the fitting loss:

$$\mathcal{L}_{\text{fit}}^t = \mathcal{L}_2\Big(\hat{J}_t, J_t\big(\mathbf{r}_t, \boldsymbol{\phi}_t, \mathcal{D}(\mathbf{z}_t)\big)\Big), \tag{15}$$

where $\mathcal{D}$ denotes the VPoser decoder. Following prior works [32, 90, 91], additional physics losses are incorporated to enhance physical plausibility, yielding the composite loss:

$$\mathcal{L} = \frac{1}{T}\sum_{t=0}^{T}\mathcal{L}_{\text{fit}}^t + \lambda_{\text{physics}}\mathcal{L}_{\text{physics}}. \tag{16}$$

Detailed formulation of $\mathcal{L}_{\text{physics}}$ is in Appendix C.4.

4. Experiments

Experimental Settings. We evaluate our HSI generation framework across two distinct settings: static scene interactions and dynamic object interactions. The ***static setting*** focuses on human motion synthesis within fixed environments, where the scene geometry remains unchanged throughout

Figure 4. **Qualitative comparison of interactions with static scenes on AnyInteraction.** ZeroHSI generates 4D HSIs that are more realistic and better aligned with text prompts, demonstrating generalizability across diverse scenes and interaction types compared to baselines.

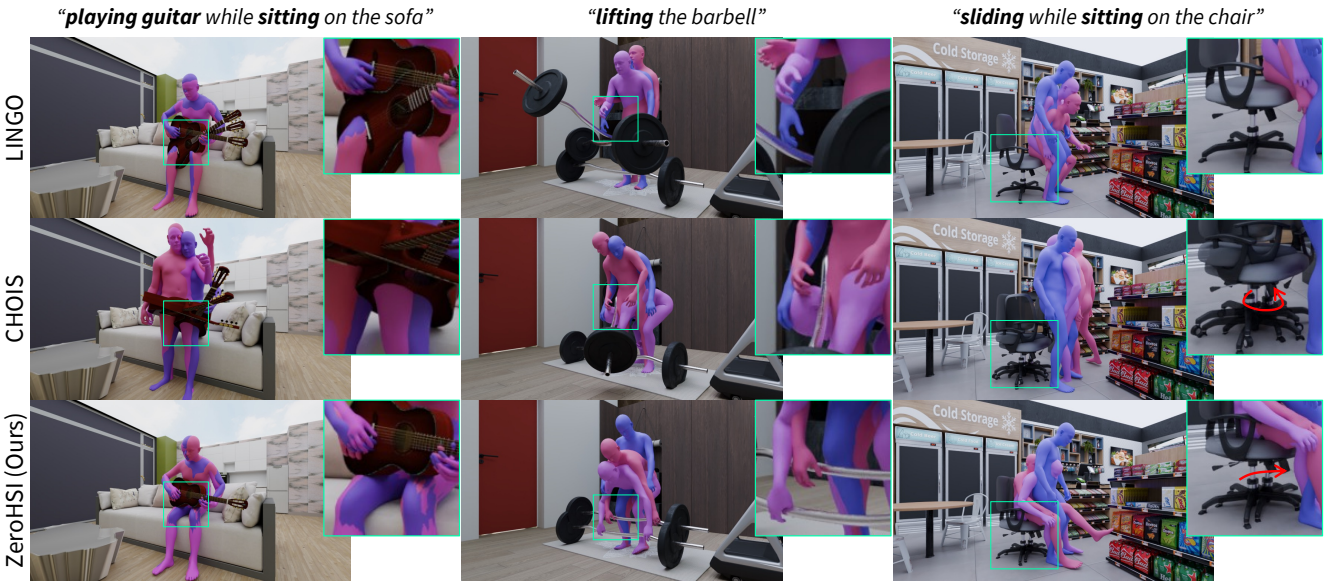

Figure 5. **Qualitative comparison of interactions with dynamic objects in scenes on AnyInteraction.** Our method maintains proper object contact while minimizing penetration, successfully handling challenging interactions like sliding while seated on an office chair.

the interaction. The ***dynamic setting*** extends this to include movable objects, requiring the simultaneous generation of human motion and object pose sequences.

Evaluation Dataset. We introduce AnyInteraction, an evaluation benchmark comprising various 3D environments sourced from existing datasets [34], public asset libraries, and real-world reconstructed scenes from Mip-NeRF 360 [4] and Tanks and Temples [41]. The dataset encompasses 12 distinct 3D environments (7 indoor and 5 outdoor), spanning residential spaces (e.g., bedrooms and living rooms), recreational facilities (e.g., gyms and cafes), and outdoor venues (e.g., greenhouses and playgrounds). For evaluating the dynamic setting, we augment these environments with interactive object models from BlenderKit and 3dsky. Each scene is annotated with 1-3 natural language descriptions of human-scene interactions and corresponding initial positions. To assess generative diversity, we synthesize multiple HSI sequences per text-position pair, yielding 100 distinct evaluation instances. Details and statistics of our AnyInteraction dataset are available in Appendix B.

Baselines. For the the static setting, we use TRUMANS [34] and LINGO [33] as baseline methods. TRUMANS [34]

Figure 6. **Qualitative results of long-term interactions with reconstructed real scenes on AnyInteraction.** ZeroHSI generates long-term interaction sequences with multiple text prompts in reconstructed scenes from the Mip-NeRF 360 (Garden and Bicycle) dataset [4].

generates interactions with static scenes conditioned on a navigation trajectory. LINGO [33] synthesizes scene-aware human motions autonomously based on text instructions and goal locations inside the scene.

In the dynamic setting, we compare our method with LINGO [33] and CHOIS [49]. LINGO [33] generates grasp/put-down actions by designating goal locations for hand-object attachment and release. CHOIS [49] synthesizes human-object interaction guided by language and sparse object waypoints.

Evaluation Metrics. Since we target a zero-shot generation task that does not have ground-truth, we evaluate our approach through rendered videos and quantitative metrics assessing three key aspects: semantic alignment, motion diversity, and physical plausibility. We render the synthesized interactions by applying the generated SMPL-X parameters and 6D object pose sequences to Gaussian avatar and object, visualizing the results via Gaussian rasterization [38].

Semantic alignment: We compute: (i) CLIP score [68] between input text prompts and rendered frames to assess text-motion correspondence, and (ii) frame-wise CLIP consistency to measure temporal coherence through cosine similarity of adjacent frame embeddings.

Motion diversity: We generate five HSI sequences per evaluation instance with identical inputs. We compute the mean per-joint Euclidean distance between each pair of generated sequences, with higher values indicating greater diversity in the synthesized motions.

Physical plausibility: For both static and dynamic settings, we measure scene penetration following [100]: (i) $\text{Pene}_{\%scene}$: percentage of body vertices penetrating the scene, (ii) Pene_{mean}: average penetration depth, and (iii) Pene_{max}: maximum penetration depth. For static settings, we additionally evaluate foot sliding score (FS) adapted from NeMF [27]. In dynamic settings, we assess object interactions using CHOIS [49] metrics: hand-object contact ratio (Cont.) and average object penetration depth (Pene_{obj}). All penetration metrics are computed using pre-calculated Signed Distance Fields (SDFs) for scenes and objects.

4.1. Comparisons

Static Setting. We show examples of zero-shot HSI generation in Fig. 4. TRUMANS [34], which only accepts scene conditions as input in our experiments, merely follows the trajectory without meaningful interactions when encountering novel scenes. While LINGO [33] successfully generates text-aligned HSIs for familiar text prompts, it fails to properly avoid collisions in unseen environments, as indicated in examples of picking snacks and sitting on sofa backs. In addition, LINGO also struggles to synthesize reasonable motions of unseen interaction types. In contrast, ZeroHSI generates plausible 4D HSIs across diverse scenes, demonstrating its generalizability to various environments.

We show the quantitative metrics in Tab. 1. Aligned with visual observations, our ZeroHSI outperforms TRUMANS [34] and LINGO [33] across most evaluation metrics, achieving the highest CLIP score and diversity while significantly reducing scene penetration metrics. These results indicate that our approach generates more diverse and plausible motions with better semantic alignment.

Dynamic Setting. In terms of the dynamic setting, our method excels in interactions involving dynamic objects. As illustrated in Fig. 5, while LINGO [33] and CHOIS [49] generate semantically relevant motions, they suffer from object penetration issues (playing guitar) and poor contact quality (lifting barbell). They also struggle with complex interactions requiring multiple body parts and global translation, such as sliding while seated. Our ZeroHSI consistently handles these challenging settings, generating high-quality

Method	CLIP Score↑	CLIP Consistency↑	Diversity↑	$\text{Pene}_{\%scene}$↓	Pene_{mean}↓	Pene_{max}↓	FS↓
TRUMANS [34]	22.36	0.9934	0.1527	0.046	0.240	1.892	0.196
LINGO [33]	22.61	**0.9942**	0.1698	0.058	0.421	2.056	**0.106**
ZeroHSI (ours)	**23.52**	0.9928	**0.1703**	**0.019**	**0.147**	**1.330**	0.158

Table 1. **Quantitative evaluation of interactions with static scenes.** ZeroHSI achieves better semantic alignment with text inputs (higher CLIP score), motion diversity, and physical plausibility (lower scene penetration) compared to TRUMANS [34] and LINGO [33].

Method	CLIP Score↑	CLIP Consistency↑	Diversity↑	$\text{Pene}_{\%scene}$↓	Pene_{mean}↓	Pene_{max}↓	Cont.↑	Pene_{obj}↓
CHOIS [49]	22.11	0.9871	**0.3382**	0.025	0.191	1.877	0.687	1.581
LINGO [33]	22.99	**0.9965**	0.0914	0.032	**0.089**	**0.446**	0.699	0.242
ZeroHSI (ours)	**24.01**	0.9955	0.1942	**0.022**	0.109	1.062	**0.835**	**0.033**

Table 2. **Quantitative evaluation of interactions with dynamic objects in scenes.** Our method outperforms baselines with stronger semantic alignment with text prompts and better dynamic interaction quality (higher contact ratio and lower object penetration).

Setting	Method	Realism	Alignment
Static	vs. TRUMANS [34]	85.0%	93.1%
	vs. LINGO [33]	75.1%	84.0%
Dynamic	vs. CHOIS [49]	96.5%	99.0%
	vs. LINGO [33]	86.9%	89.1%

Table 3. **Human study on generated 4D HSI motions.** In both static and dynamic scenarios, participants prefer our generated HSIs for their motion realism and semantic alignment by large margins.

interactions with dynamic objects.

We show the quantitative results in Tab. 2. ZeroHSI achieves the highest contact ratio and lowest object penetration compared to LINGO [33] and CHOIS [49]. This precise object interaction is further complemented by strong semantic alignment and reliable scene interaction, demonstrating our method's capability to generate natural and accurate interactions with dynamic objects within scenes while maintaining high motion quality.

Human Perceptual Study. We conduct a human perceptual study using the two-alternative forced choice (2AFC) method. We recruit 400 participants. Participants are given results generated by different methods using identical input and instructed to select the sample they perceive as more realistic and better aligned with the textual description. As reported in Tab. 3, users consistently prefer our results over baselines by significant margins in both static and dynamic settings, aligning well with our quantitative metrics and qualitative examples.

4.2. Long-Term Interaction Synthesis

We show examples of long-term HSIs in reconstructed real scene generated by ZeroHSI in Fig. 6. The Garden example shows a character walks forward, waters flowers, puts down the watering can, and leans on the table. In the Bicycle example, the character sits on a bench, stands up, and practices Tai Chi. These examples showcase ZeroHSI's capability to generate realistic interactions in real-world environments and its application for producing sequences of varying lengths. More details and visualizations are in Appendix C.7.

Method	CS↑	CC↑	$\text{P}_{\%}$↓	P_{mean}↓	P_{max}↓	FS↓
WHAM [72]	22.73	0.9890	0.048	0.505	4.917	**0.054**
WHAC [93]	22.81	0.9903	0.168	15.30	99.68	0.985
ZeroHSI w/o OPT_{body}	23.39	0.9890	0.025	0.165	**1.092**	0.245
ZeroHSI (ours)	**23.52**	**0.9928**	**0.019**	**0.147**	1.330	0.158

Table 4. **Quantitative results of ablation study.** "CS" denotes CLIP Sore, "CC" denotes CLIP Consistency, "$\text{P}_{\%}$" denotes $\text{Pene}_{\%scene}$, "P_{mean}" denotes Pene_{mean}, "P_{max}" denotes Pene_{max}. Our full method achieves higher motion quality and physical plausibility compared to variants.

4.3. Ablation Study

To evaluate the effectiveness of our per-frame optimization approach, we conduct an ablation study in the static scenario. We replace our camera pose estimation and HSI optimization modules with state-of-the-art human motion reconstruction methods WHAM [72] and WHAC [93]. Additionally, we evaluate a variant of our model ("ZeroHSI w/o OPT_{body}") that removes body pose optimization and instead only optimizes root translation and global orientation for 30 iterations per frame, using poses directly estimated by SMPLer-X [10].

Quantitative results in Tab. 4 show that our full model outperforms the ablated variants across most metrics, demonstrating that our per-frame optimization approach effectively reconstructs human-scene interactions and serves as the optimal solution for bridging 2D HSI videos and 4D HSIs. Qualitative examples are presented in Appendix C.8.

5. Conclusion

We present ZeroHSI, a zero-shot approach to 4D human-scene interaction generation that addresses the limitation of requiring paired motion-scene training data. Our method successfully distills HSIs from video generation models through neural human rendering to synthesize contextually appropriate interactions across diverse environments.

Acknowledgments. We thank Zimo He and Nan Jiang for experimental setup. The work was in part supported by ONR YIP N00014-24-1-2117 and ONR MURI N00014-22-1-2740. J. Li was in part supported by the Wu Tsai Human Performance Alliance at Stanford University.

References

[1] Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Deepak Gopinath, Jiajun Wu, Alexander Clegg, and C Karen Liu. CIRCLE: Capture in rich contextual environments. In *CVPR*, 2023. 2

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, 2024. 3

[3] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *ICCV*, 2023. 2

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 6, 7, S1, S2

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 2

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 3

[8] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *3DV*, 2024. 2

[9] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 3, S8, S9

[10] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2024. 3, 5, 8

[11] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 2

[12] Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. In *CVPR*, 2023. 3

[13] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *CVPR*, 2022. 2

[14] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. MoFusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 2

[15] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 2

[16] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *CVPR*, 2023. 3

[17] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 2

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 3

[19] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, 2024. S1, S5

[20] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, 2022. 2

[21] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022. 2

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[23] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, 2021. 2

[24] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 2

[25] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 2

[26] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH*, 2023. 2

[27] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022. 7

[28] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[29] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[30] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[31] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *CVPR*, 2024. 3

[32] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3D scenes. In *CVPR*, 2023. 2, 5

[33] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia*, 2024. 2, 6, 7, 8, S5

[34] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024. 2, 6, 7, 8, S1, S2, S5

[35] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 3

[36] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 3

[37] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. GMD: Controllable human motion synthesis via guided diffusion models. In *ICCV*, 2023. 2

[38] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139, 2023. 3, 4, 7, S1

[39] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *ECCV*, 2025. 2

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. S1

[41] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):78, 2017. 6, S1, S2

[42] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *CVPR*, 2024. 3

[43] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *CVPR*, 2024. 2

[44] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. In *NeurIPS*, 2024. 3

[45] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *ICCV*, 2023. 2

[46] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 3

[47] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023. 2

[48] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM TOG*, 42(6):197, 2023. 2

[49] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2025. 2, 7, 8

[50] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *CVPR*, 2024. 2

[51] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *WACV*, 2024. 2

[52] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 3, 4

[53] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM TOG*, 40(6):219, 2021. 3

[54] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. In *ACM MM*, 2024. 3

[55] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):248, 2015. 3

[56] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 5

[57] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM TOG*, 39(4):39, 2020. 2

[58] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2021. 3

[59] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):102, 2022. 3

[60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3, 4, 5, S3

[61] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 3

[62] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body:

Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3

[63] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3D human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 2

[64] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021. 2

[65] Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 2

[66] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, 2021. 2

[67] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. In *ICLR*, 2024. 2

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[69] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4, S1

[70] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024. 2

[71] Yi Shi, Jingbo Wang, Xuekun Jiang, and Bo Dai. Controllable motion diffusion model. *arXiv preprint arXiv:2306.00416*, 2023. 2

[72] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 3, 8, S6, S7

[73] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[74] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2

[75] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 2

[76] KLING AI Team. Kling image-to-video model. https://klingai.com/image-to-video/, 2024. 3, 4, S1, S7, S8

[77] Wan Team. Wan: Open and advanced large-scale video generative models. 2025. S8, S9

[78] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 2022. 2

[79] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. In *ICLR*, 2023. 2

[80] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. EDGE: Editable dance generation from music. In *CVPR*, 2023. 2

[81] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE RA-L*, 7(2):4702–4709, 2022. 2

[82] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3D human motion and interaction in 3D scenes. In *CVPR*, 2021. 2

[83] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3D human motion synthesis. In *CVPR*, 2022. 2

[84] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 3

[85] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *NeurIPS*, 2022. 2

[86] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024. 3

[87] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 2

[88] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024. 2

[89] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. Hierarchical planning and control for box loco-manipulation. In *SCA*, 2023. 2

[90] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2, 5

[91] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 5

[92] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[93] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *ECCV*, 2024. 3, 8, S6, S7

[94] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *CVPR*, 2023. 3

[95] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 2

[96] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM TOG*, 40(4):121, 2021. 2

[97] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 46(6):4115–4128, 2024. 2

[98] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *ECCV*, 2022. 2

[99] Zihan Zhang, Richard Liu, Rana Hanocka, and Kfir Aberman. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *SIGGRAPH*, 2024. 2

[100] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *ICCV*, 2023. 7

[101] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *CVPR*, 2023. 2

[102] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas Guibas. GIMO: Gaze-informed human motion prediction in context. In *ECCV*, 2022. 2

[103] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. In *3DV*, 2025. 3