GenHSI: Controllable Generation of Human-Scene Interaction Videos

Zekun Li Rui Zhou Rahul Sajnani Xiaoyan Cong Daniel Ritchie Srinath Sridhar
Brown University

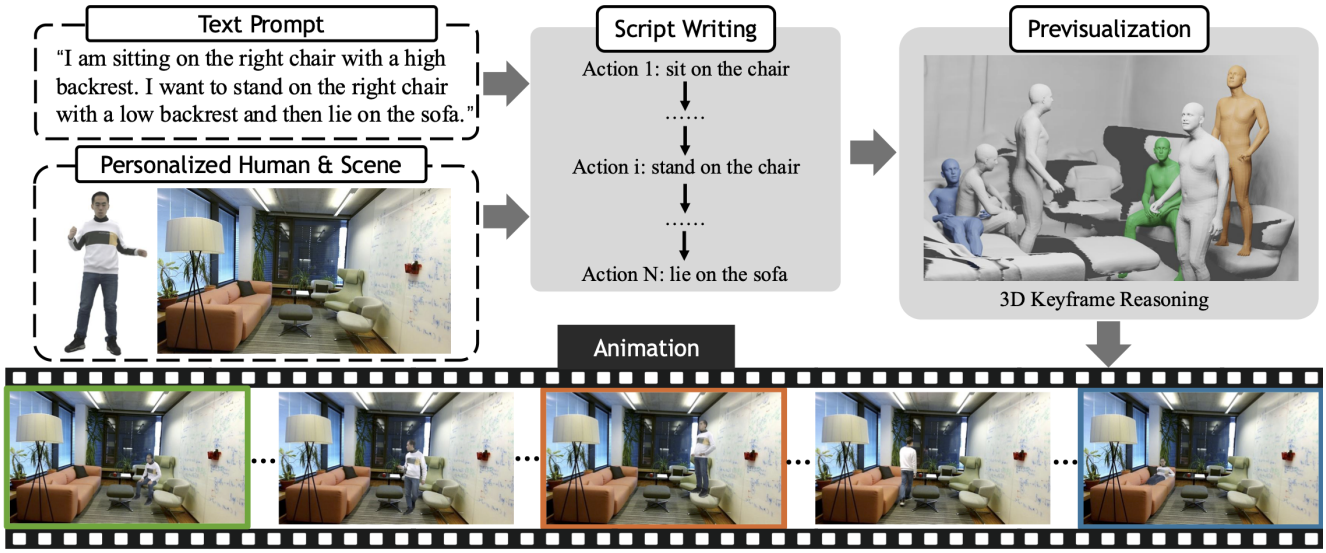

Figure 1. **GenHSI** is a 3D-aware controllable human-scene interaction (HSI) video generation method. We mimic the real-world film-making procedure, *i.e.*, Script Writing, Previsualization, and Animation, to generate an extendable HSI video clip with arbitrary lengths of action chains. Given images of the scene and character with the action sequence prompt, our method will render multiple 3D-aware keyframes based on the posed 3D Gaussian avatar and 3D Gaussian scene. Finally, we interpolate them into a continuous video using the pretrained video diffusion model. The frames with colored borders are selected 3D-aware keyframes that map to the color human meshes.

Abstract

Large-scale pre-trained video diffusion models have exhibited remarkable capabilities in diverse video generation. However, existing solutions face several challenges in generating long videos with rich human–scene interactions (HSI), including unrealistic dynamics and affordance, lack of subject identity preservation, and the need for expensive training. To this end, we propose GenHSI, a training-free method for controllable generation of long HSI videos with 3D awareness. Taking inspiration from movie animation, we subdivide the video synthesis into three stages: (1) script writing, (2) pre-visualization, and (3) animation. Given an image of a scene and a character with a user description, we use these three stages to generate long videos that preserve human identity and provide rich and plausible HSI. Script writing converts a complex text prompt involving a chain of HSI into simple atomic actions that are used in the pre-visualization stage to generate 3D keyframes. To synthesize plausible human interaction poses in 3D keyframes, we utilize pre-trained 2D inpainting diffusion models to generate plausible 2D human interactions based on view canonicalization, which eliminates the need for multi-view fitting in previous works. We then extend these interactions to 3D using robust iterative optimization, informed by contact cues and reasoning from VLMs. Prompted by these 3D keyframes, the pretrained video diffusion models can better generate consistent long videos with plausible dynamics and affordance in a 3D-aware manner. We are the first to synthesize a long video sequence with a chain of HSI actions without training based on the image references of the scene and character. Experiments demonstrate that our method can generate HSI videos that effectively preserve scene content and character identity with plausible human-scene interaction from a single image scene.

1. Introduction

Rapid progress has been made in the last few years on the problem of photorealistic image [10, 16, 40, 61] and video [5, 9, 23, 25, 45, 50, 71, 78] generation, especially using diffusion models [28, 67]. These generative image and video models can synthesize high-quality and diverse content and even support multi-modal control using text instructions [6, 20, 29], audio [68], camera poses [27, 73], and other modalities [11, 22, 51, 65, 79, 83]. These developments have opened up broad applications: for instance, in personalization [13, 62], style transfer [14], and editing [64, 66, 88]. Particularly popular are applications that enable pose-controllable generation of humans in videos [7, 18], human-object interactions [53, 59, 80, 90], and 3D human motion generation [41].

Despite this progress, current video diffusion models (VDMs) face several challenges, particularly when dealing with human physical interactions with the environment. Since the dynamics are implicitly formalized in visual generative models, they often result in unrealistic physical phenomena [54], especially when multifaceted dynamics happen, *e.g*. unrealistic HSI. Additionally, without explicit 3D modeling, VDMs frequently fail to ground text instructions in accurate spatial reasoning, resulting in artifacts such as characters heading in incorrect directions or hallucinated 3D affordances during HSI. Furthermore, it is hard to generate videos with a consistent person identity when composing image references of the human and scene implicitly, even with expensive training or fine-tuning [12, 36].

To this end, we present **GenHSI**, a training-free method for controllable generation of long human-scene interaction (HSI) videos via 3D-aware keyframe prompting given an image of the scene and person with the HSI text prompt (see Fig. 1). Specifically, we break down the HSI video generation problem into three stages: (1) script writing, (2) pre-visualization, and (3) animation. In the **script writing stage**, GenHSI prompts a VLM [55] to generate a more detailed *script* with step-by-step instructions that decompose complex interactions into a sequence of simpler atomic tasks. In the **pre-visualization stage**, we achieve accurate human-object contacts and interactions by creating 3D keyframes for each atomic task. This novel keyframe generation step synthesizes human-object interactions using an inpainting model in canonical view and optimizes contacts between them in 3D for improved affordance. A key advantage of our work is the benefit we get from large reconstruction models [70, 77] (even when they provide inaccurate geometries) to reconstruct 3D scenes from real-world images, alleviating the need for accurate scene geometry assumed in prior works [15, 33, 44, 74, 87]. Finally, during the **animation stage**, the scene [34] and character [60] are modeled as 3D Gaussians, which inherently accommodate visual occlusion in 3D. We interpolate the rendering results based on parsed HSI scripts to generate controlled 3D-aware videos using an off-the-shelf video generation model [3].

As there are no existing solutions that generate long videos with accurate human-object interactions, we compare individual components of our system against prior works. Our evaluation includes comparing human-object interaction estimation against diffusion-based solutions [57] and 3D human-scene interaction methods [44, 74, 87] on real-world scanned scenes [26]. Additionally, we also compare against commercial solutions [2]. Extensive qualitative and quantitative experiments demonstrate that our method produces more physically plausible results in 3D human-scene interactions and achieves superior visual quality in preserving human-scene consistency.

In summary, our main contributions are summarized as:

- We propose a training-free method **GenHSI** for controllable generation of long human-scene interaction videos given only a scene image, a person image, and an interaction text description.
- Rather than deal with the HSI video generation using a single stage, we break it down into three stages: **script writing**, **pre-visualization**, and **animation** that enable better 3D control through 3D-aware keyframe prompting, while being training-free and retaining person identities.

These contributions open the doors to diverse applications in controllable long HSI video generation, training data generation, and video personalization.

2. Related Works

Human-Scene Interaction Video Generation: Current video models can generate impressive human motions [17] with personalized visual characters [47, 85, 86], with controls such as human poses [19, 80], masks [89], and audio [46]. However, it is difficult to control the *interaction* of the person with the environment. Some existing models [12, 21] still struggle to generate HSI with plausible affordances and dynamics, even when trained on large datasets with extensive GPU resources for fine-tuning. To reduce the hallucination in HSI video generation, some approaches [30, 52] leverage the existing video as the reference and replace either the human or the object to achieve customization. However, these approaches rely on real videos as a foundation and are thus limited to editing existing content, lacking the ability to generate human-scene interactions from scratch. Recent work [37] distills the interaction prior into a special token [62] to achieve customized HSI video generation from image references of scene and character, which still involves training and is limited to a small scale of video generation model. Different from GenHSI, our 3D-aware keyframe prompting is training-free and model-agnostic general solution for plausible HSI video generation.

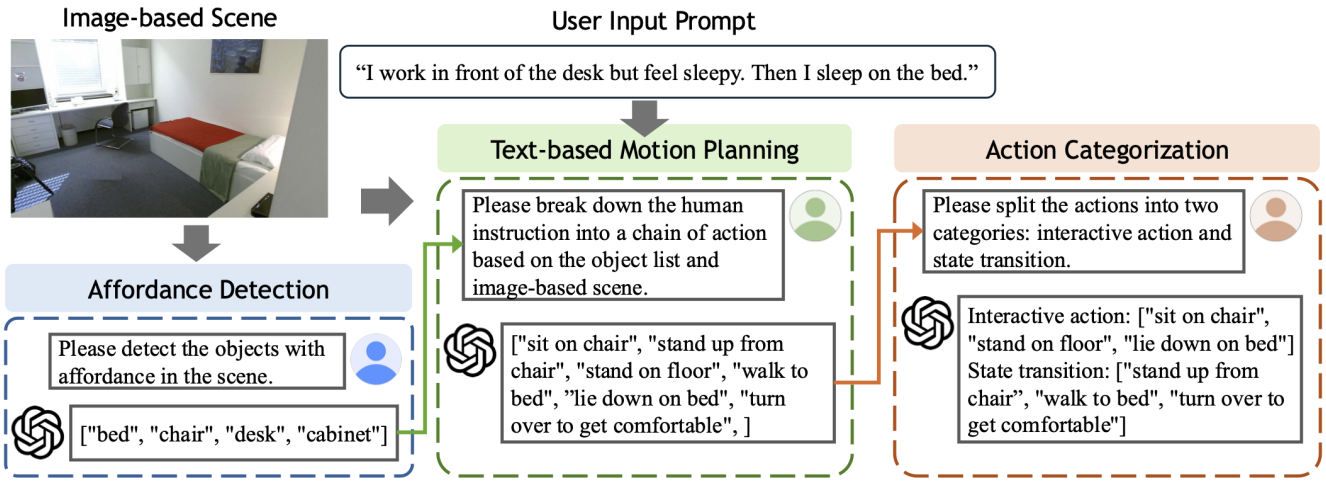

Figure 2. **Script Writing Stage:** Complex high-level text descriptions from users do not provide a detailed scene and task understanding for the desired long video generation. The script writing stage first **identifies and segments objects** that the human can interact with in the scene. These objects, along with the given human prompt, are used to **perform text-based motion planning** from a VLM [55] that provides us with **interactive actions & state transitions for keyframing** in the **Pre-Visualization Stage**.

Human-Scene Interaction 3D Pose Synthesis: Synthesizing humans in scenes is a crucial yet challenging task in computer vision and graphics, requiring the modeling of complex, high-level semantic understanding, such as affordances and interactions. With paired scene-motion datasets [4, 15, 24, 26, 33], previous works [8, 15, 31, 33, 72, 74, 82, 87] encode the scene geometry as latent conditions for human pose and motion generation. However, these methods rely on high-quality 3D scene datasets with motion-captured human interactions, making them difficult to scale and generalize across diverse environments. In contrast, we focus on zero-shot interaction synthesis to generate plausible human-scene interaction from an image diffusion prior without any motion dataset or training, inspired by the comprehensive understanding of human-scene composition priors in diffusion-based image editing systems [39, 57, 63, 69, 81]. Although existing works [35, 44] attempt to lift 3D human poses from multi-view inpainting, they fail to guarantee cross-view consistency because each view is inpainted independently. This 3D inconsistency often leads to instability in subsequent pose lifting optimization. To address this issue, we canonicalize coarse 3D reconstructions of objects and perform character inpainting from the canonical viewpoint that maximizes the visibility of object affordances. Our efficient 3D human lifting approach, grounded in chain-of-contact reasoning and single-view inpainting, effectively avoids these limitations.

3. GenHSI

The goal of GenHSI is to generate a personalized long human-scene interaction video given natural language text descriptions and an image of a scene and a character. To overcome hallucinations during interaction in existing video generative models, we utilize 3D-aware keyframes generated from explicit 3D reasoning to prompt off-the-shelf video generative models. Our solution mimics the real-world filmmaking process through a modularized approach (Script Writing - Previsualization - Animation) as shown in Fig. 1. We first parse a high-level text description into simple atomic tasks using a VLM under image scene context (see Fig. 2), which includes physical interactions and state transitions based on the affordances in the image scene (Sec. 3.1). After obtaining the detailed and structured motion script, we prompt human-object interaction using a novel zero-shot pose generation that leverages a pre-trained inpainting model (Sec. 3.2). We then compose the scene [70, 77] and human [49, 58] in 3D under affordance constraints that lead to physically plausible interactions between the character and the environment (Sec. 3.3). To render them into a holistic 3D-aware keyframes and interpolate as HSI video, we use the estimated depth point cloud [70] as initialization for scene 3D Gaussian fitting and feed-forward 3D Gaussian avatar generation model [60] to obtain character assets (Sec. 3.4).

3.1. Script Writing Stage

The HSI text descriptions from users are usually structureless and need a chain-of-action to execute. Hence, we first perform script writing that takes a complex human description and a scene image to reason [55] and localize [38, 76] the target interaction object using a segmentation mask. Based on the image understanding, we break the text description into atomic action scripts based on the scene understanding in Fig. 2, which serves as the basis for generating 3D keyframes in the **Pre-visualization** stage and interpolating keyframes in the **Animation** stage.

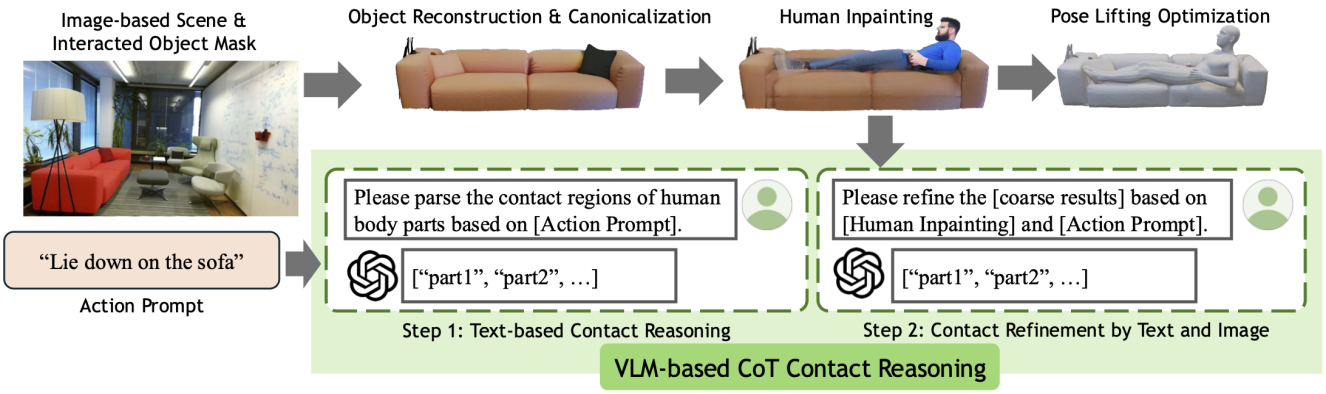

Figure 3. **3D Keyframe Generation for Pre-visualization.** GenHSI synthesizes 3D human-scene interaction pose based on the pretrained 2D image inpainting diffusion model to create a 3D keyframe as an intermediate step for HSI video generation. Our method lifts the 2D human inpainting result in the canonical view of the target object based on contact cues reasoned by the VLM chain-of-though.

Scene Understanding and Object Detection: Before generating the video, it is essential to understand the scene and identify objects that humans can interact with. We first provide the input scene image to the VLM and prompt it to describe the scene as well as list objects that support human-object interaction in Fig. 2. According to the user input prompt, we perform open-set detection [48] & segmentation [38] to locate the involved objects. The object masks m_o will be used in 6D pose and scale estimation.

Text-based Motion Planning and Categorization: After obtaining the scene description and segments of possible objects of interest, our next step is to plan the progress of HSI in a plausible dynamics procedure. This text-based motion plan is crucial for breaking down complex motion descriptions into simpler, detailed instructions that are used to create keyframes and maintain smooth motion in video generation. Hence, we utilize the VLM to convert complex, high-level human text descriptions into a script of low-level, atomic actions that ensure smooth and natural transitions. Each low-level action clearly describes the relationship between the human and the object instance in the scene using a verb or preposition. However, as shown in Fig. 2, not all the low-level actions from text-based motion planning outputs represent a physical interaction between the human and the scene, which motivates us to categorize them into actions that describe human-object interactions (interactive actions) and body movement without changing interaction (state transitions). As the interactive actions model physical constraints in human-scene interactions, we use them to create 3D keyframes to enhance the plausibility in video generation. Additionally, utilizing state transitions, along with interactive actions, ensures a logical and smooth progression of human motion during keyframe interpolation.

3.2. Pre-visualization Stage: 2D Human Inpainting

In the second stage **Pre-visualization**, our method generates keyframes to prompt off-the-shelf video generative models. To reduce hallucinations in spatial understanding during HSI video generation, such as 3D affordance and interaction targets, we propose a novel and efficient 3D HSI pose generation method using a pretrained 2D image inpainting model and contact-guided 3D pose lifting optimization. As shown in Fig. 3, given the image of the scene with the target object segment and the interactive action text prompt c obtained from **Script Writing** stage, we synthesize a 3D human mesh $\mathcal{M}_h$ parameterized by pose and shape parameters (θ, β), performing the specified interaction with the object in the scene.

Variant Performance of 2D Human Inpaint: Different from regular inpainting tasks, inpainting humans in scene images requires localizing the human in the scene and prompting the model to interact with the human and the object of interest. We input the object image and a prompt detailing the desired human-object contact to extract human-object contact priors from the diffusion model. However, directly obtaining human-object interaction in random views of objects does not always perform well (see Fig. 4). This issue occurs primarily because: 1) the diffusion model is biased towards forward-facing views of objects (Fig. 4a) and 2) inpainting in random views often occludes human-object interactions, leading to sub-optimal human poses (Fig. 4b). For instance, inpainting hallucinates the human sitting in the air in Fig. 4a when the back of the chair is visible in the scene, and Fig. 4b displays a human sitting in an uncomfortable pose looking towards the camera, even when the couch is rotated sideways.

Effective Human Inpainting in the Canonical View: Our solution to curb the afore-mentioned issues is to first canonicalize the object of interest and then synthesize human interaction poses. We hypothesize that diffusion models can readily capture the canonical view of objects, making it more accurate to estimate human object interactions

(a) Hallucinate Affordance (b) Inplausible Interaction

Figure 4. Failure cases when inpainting a human from different object views, demonstrating that the performance of pose generation from an image diffusion prior is view-dependent.

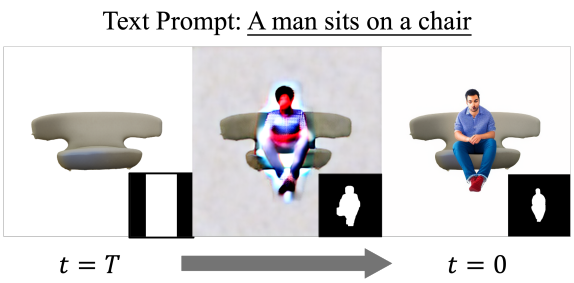

Figure 5. Human-Object Interaction from 2D Inpainting Models in the canonical view of objects. We progressively update the human mask (bottom) while denoising the inpainting result (top).

in this view. Hence, we prompt inpainting models to synthesize human-object contacts in the canonical view of the object. We reconstruct the 3D object using the image-to-3D model Trellis [77] and canonicalize it via OrientAnything [75]. Next, we render the object to obtain the 2D human interaction in this canonical view using an inpainting model [1] (see Figs. 3 and 5) instead of rendering in random multiple views as previous works [35, 44]. The canonicalized rendered image of the object is input to the diffusion model with a coarse mask m_t to inpaint. We then predict the clean image at every denoising step using Tweedle's formula and use Detectron [76] to estimate the human mask. This human mask is then used to update the input mask of the inpainting model for the next denoising step, leading to easy, efficient, and fine-grained interaction (see Fig. 5) as follows:

$$\mathbf{z}_{0|t} = \frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\Theta(\mathbf{z}_t, \mathbf{z}_0^*, m_t, c, t)}{\sqrt{\bar{\alpha}_t}} \quad (1)$$

$$m_{t-1} = \text{Segment}(\text{Decode}(\hat{\mathbf{z}}_{0|t})) \quad (2)$$

$$\hat{\mathbf{z}}_{0|t} = (1 - \downarrow m_{t-1}) \odot \mathbf{z}_0^* + \downarrow m_{t-1} \odot \mathbf{z}_{0|t} \quad (3)$$

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_t}\hat{\mathbf{z}}_{0|t} + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (4)$$

where $\downarrow m$ is the downsampled mask with the aligned shape with noise latent $\mathbf{z}$ and $\mathbf{z}_0^*$ is the original object rendering image latent extracted by the U-net encoder in the pretrained diffusion model. We then estimate the SMPL-X parameters by HybrIK-X [42, 43] in the canonical view that offers rich visual affordances and reduced self-occlusions.

3.3. Pre-visualization Stage: 3D Pose Lifting

Since we extract the human pose from a single-view image, there is depth-scale ambiguity after estimation. To compose the human and target object seamlessly in 3D, we formulate an optimization framework that corrects the scale s_h, translation t_h, and global rotation r_h of the human to resolve inaccurate scale & depth in 3D. This optimization (1) uses the contact cues from VLM to compose the human and object in 3D, (2) minimizes the penetration between the human and the coarse 3D object reconstruction, and (3) matches the silhouette of the projected human mesh with the inpainted 2D interaction result.

Interaction Affordance Loss: To compose the human and the target interacting object in the same 3D space, we first use the VLM to reason which predefined body parts contact the target object, as shown in Fig. 3. Given the contact cues, we propose an interaction loss $\mathcal{L}_{hoi}$ to ensure that the regions highlighted by the VLM are in contact. This loss is formulated by minimizing the loss between the points P_h of contact on the human surface that are identified by the VLM and the points P_o on the object surface. We first gather the possible contact points in P_h and P_o using the k-nearest neighbor method to obtain P_h^* and P_o^*. We then ensure that the point pairs in P_h and P_o exist in the k-nearest neighbor subsets of each other. This loss is formulated as

$$\mathcal{L}_{hoi} = \sum_{x \in P_h^*} \min_{y \in P_o^*} \|x - y\|_2^2 + \sum_{y \in P_o^*} \min_{x \in P_h^*} \|y - x\|_2^2. \quad (5)$$

Penetration Loss: Interaction loss alone increases human object contact but leads to human-object penetration. Hence, we encourage the optimization to avoid human-object penetration by constructing a signed distance field (SDF) Φ from the human mesh $\mathcal{M}_h$ and ensuring that points v on the object surface $\mathcal{M}_O$ have a non-negative SDF value *i.e.* no penetration with the human mesh:

$$\mathcal{L}_{pen} = -\mathbb{E}_{v \in \mathcal{M}_O}[\min\left(\Phi(v), 0\right)]. \quad (6)$$

Silhouette Loss: Both $\mathcal{L}_{hoi}$ and $\mathcal{L}_{pen}$ ensure better human-object affordances and contact, but they do not fix the scale of the human. We ensure that the projected mask of the human matches the inpainting mask using an intersection over union (IoU) constraint. Our solution involves rendering two masks from a silhouette rasterizer Π corresponding to 1) the human $m_h = \Pi(\mathcal{M}_h)$ and 2) the human with object occlusions $m_{hoi} = \Pi(\mathcal{M}_h|\mathcal{M}_O)$. We then ensure that these masks overlap with the inital human mesh projection m_h^{init} obtained using SMPL-X and initial human inpainting mask m_{hoi}^*:

$$\mathcal{L}_{mask} = \frac{m_h \cap m_h^{init}}{m_h \cup m_h^{init}} + \frac{m_{hoi} \cap m_{hoi}^*}{m_{hoi} \cup m_{hoi}^*}. \quad (7)$$

Matching the mask without object occlusion alone results in improper projection, where the human mesh may translate away from the object to just match itself. In contrast, matching only the mask with object occlusion does not provide information about the occluded regions. Hence, we use both masks to penalize the optimization. The optimization solves the scale s_h, translation t_h, and global rotation r_h of the human mesh $\mathcal{M}_h$ for every keyframe to compose the human and target object in 3D space by minimizing a combined loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{hoi} + \mathcal{L}_{pen} + \mathcal{L}_{mask}. \quad (8)$$

3.4. Animation Stage

In this stage, we render the 3D-aware keyframes to prompt the pretrained video generative model to generate plausible HSI videos based on the optimized 3D human pose and the parsed chain-of-action.

3D-aware Keyframe Rendering To insert the optimized 3d human pose $\{\theta, r_h, t_h, s_h\}$ into the camera space of the input image scene, we estimate the 6D pose and scale of the object according to the coarse 3D object reconstruction template. The object 6D pose $\{r_o, t_o\}$ is first initialized by DINOv2 feature similarity [56].. For better alignment between the 3D reconstructed object and the depth estimation of the image scene, we recalculate the scale s_o and translation t_o based on silhouette and depth alignment in the original image scene. To insert the character in the human reference image with optimized pose into the image scene with a plausible visual effect, we represent both the human and scene in 3D Gaussians [34], which allows us to resolve occlusions in 3D space naturally. We initialize the 3D Gaussian centers for each pixel based on the monocular depth estimator [70] and optimize only the color, rotation, and scale parameters for each 3D Gaussian by rendering and comparing to the input image view of the scene. The human 3D avatars are obtained via the existing feed-forward Gaussian avatar generation model [60]. Compared with using image editing models to add humans into the image scene [57, 69], leveraging explicit 3D representations allows us to render 3D-aware keyframes with minimal change to the background and high preservation of character identity, while naturally handling occlusions between humans and the environment, making them well-suited for prompting off-the-shelf video generative models.

3D-aware Keyframe Interpolation The Previsualization stage provides us with a 3D storyboard in the form of keyframes for our video. These 3D keyframes provide us with rich human-object contacts for each interactive action, enabling the rendering of human-object interactions in the input view. After rendering the keyframes, we animate them using Kling AI 1.6 [3], which generates transition frames between the start frame and the end frame using the state transition actions obtained from the VLM as described in Sec. 3.1. Interpolating keyframes is analogous to in-context prompting in language generation: by decomposing an HSI video into a sequence of salient keyframes, we introduce richer constraints that mitigate hallucinations. Incorporating detailed atomic state transitions into keyframe interpolation helps maintain human identity consistency and enhances motion realism for HSI video generation.

4. Experiments

Since no solutions exist that achieve the same goal of controlling long video generation with human-scene interactions, we instead conduct quantitative and qualitative evaluations to compare the components of GenHSI with alternative solutions for 3D human pose generation and video synthesis involving human-scene interactions.

Dataset: To fairly compare with previous pose generative methods [44, 74, 87], we utilize a well-used 3D scene dataset PROX-S [84, 87] to demonstrate the effectiveness of our human pose generation. Since our method uses the image as the representation of the scene, we filter out the invisible object interaction test cases in PROX-S based on the real-world images of the scenes for fair comparison. To increase the scene diversity, we render the synthetic 3D scene dataset [33, 44] and use Flux-depth to convert them into photorealistic image scenes. The composed HSI video generation evaluation set contains 50 samples.

Metrics: We use community-accepted video evaluation metrics from [32][1] to measure **Subject Consistency**, **Dynamic Degree**, **Background Consistency**, **Motion Smoothness**, and **Imaging Quality** in the long video results. **Subject** & **Background Consistency** measure the semantic similarity of the subject and the background scene, respectively, using features extracted from DiNO. **Dynamic Degree** & **Motion Smoothness** measure the average optical flow between consecutive frames within the video and their consistency across a video. Additionally, we follow [44, 87] and use community-accepted metrics of **Semantic Clip**, **Contact**, **Non-Collision**, **Entropy**, and **Cluster Size** to evaluate human-scene interaction quality. We refer the reader to the papers referenced above for more details about each metric.

[1] We use the long-VBench to evaluate the HSI video since our generated video duration is larger than 10s.

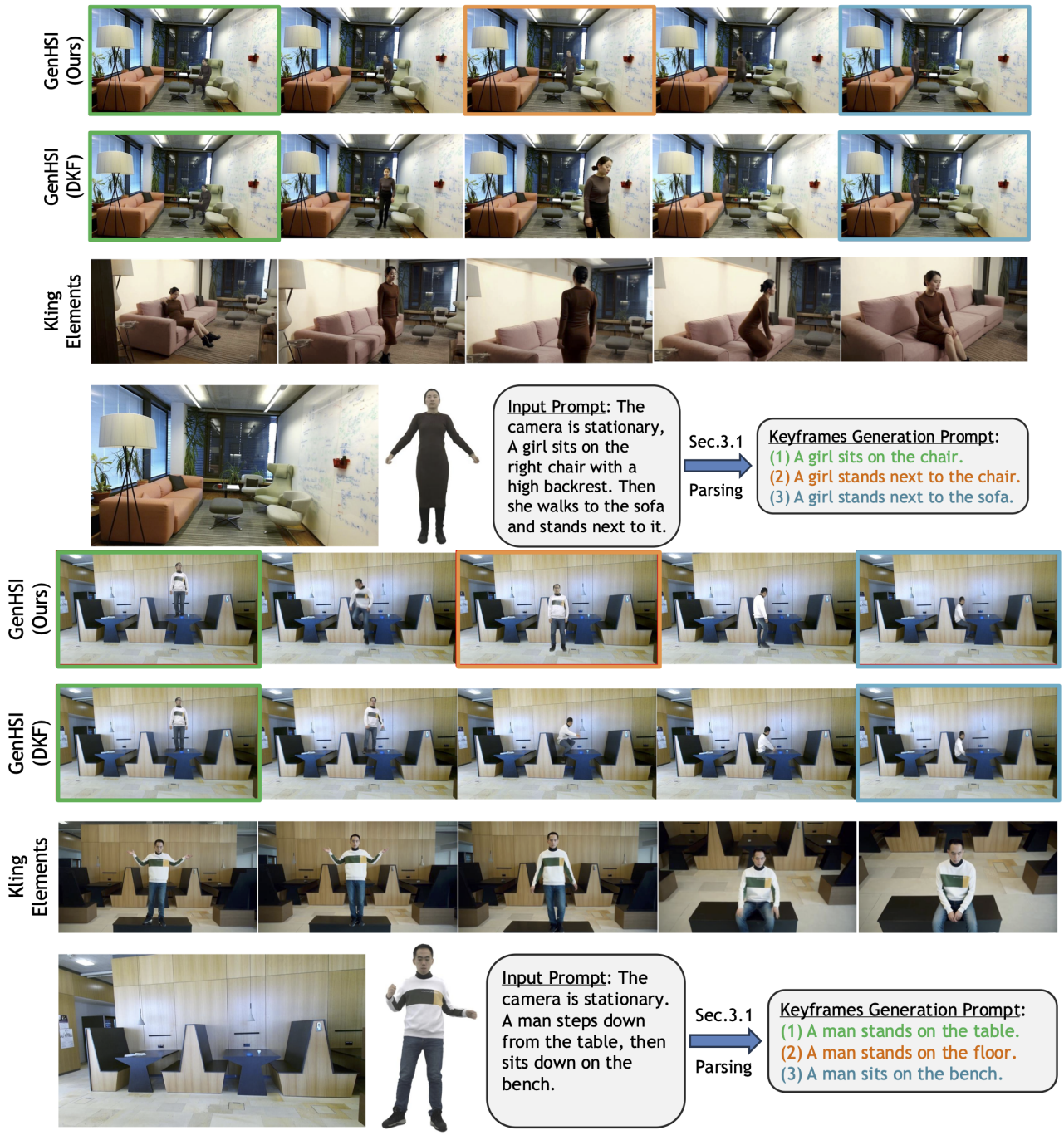

Figure 6. **HSI Video Generation Qualitative Results:** GenHSI (ours) produces the best results with subject identity preservation and good human-object contacts. GenHSI (DKF) only uses two keyframes (start and end frame) and often changes the identity of the subject. Kling Elements drastically change the scene and character. Each video displays the keyframes highlighted with a red bounding box.

4.1. Qualitative Results

We demonstrate some HSI video generation results in different settings, shown as Fig. 6. The frames with colored borders are the 3D-aware keyframes generated based on the atomic action prompt parsed in Sec. 3.1. In GenHSI, we synthesize 3D-aware keyframes for each interactive action to prompt the off-the-shelf video generative model, achieving the least spatial hallucination and highest character identity preservation. To evaluate the effectiveness of 3D-aware keyframes prompting, we conduct an ablation setting with only dual keyframes (DKF), *i.e*. the start frame and end frame. Without detailed keyframe prompts to constrain the video generation, the generative model usually hallucinates the transition dynamics, *e.g*. in the first sample (top), the woman goes out of the screen first and then magi-

cally reappears in the scene; in the second sample (bottom), the man is shown sitting down through an unrealistically narrow space between a table and a bench. To demonstrate the advantage of GenHSI, we compared it with a commercial solution, Kling-Elements [2], that supports composing multi-image references to generate customized videos. Although Kling-Elements can generally generate HSI video with better quality in terms of human appearance details, it cannot preserve the human identity like GenHSI, *e.g.* it changes the color of the cloth in the first test case. Additionally, Kling-Elements seldom follows the text prompt to drive the human character to interact with the image scene. Instead, it usually hallucinates the 3D scene and affordance, *e.g.* the model creates a non-existent table and misinterprets it as the bench to sit on in the second test case.

4.2. Quantitative Evaluation

Inpainting Successful Rate in Different Viewpoints: To validate our assumption about 2D image inpainting models struggling with view-dependent performance in Sec. 3.2, we conduct a human judgment with 10 users across 30 inpainting results to evaluate the plausibility of inpainting in different views compared with the canonical view.

Table 1. **Success Rate of Inpainting across Views** $\Delta\theta$ denotes the yaw angle relative to the canonical pose of the reconstructed 3D object.

	$\Delta\theta = 0$	$\lvert\Delta\theta\rvert \in (0, \frac{\pi}{6}]$	$\lvert\Delta\theta\rvert \in (\frac{\pi}{6}, \frac{\pi}{3}]$	$\lvert\Delta\theta\rvert \in (\frac{\pi}{3}, \frac{\pi}{2}]$
Success Rate	93.33%	54.67%	20%	6.67%

3D Human-Scene Interaction: We evaluate the performance of our human-object interaction synthesis against GenZI [44] & COINS [87] on filtered PROX-s, which exclude the invisible object interactions. Tab. 2 compares 3D Human Object Interaction results generated in the pre-visualization stage against baselines. GenHSI improves Semantic Clip and Contact from the SOTA GenZI [44] in zero-shot human scene interaction generation. Benefiting from inpainting humans from the canonical view, the diffusion model efficiently inserts humans with plausible poses by only inpainting from a single view. Our method can also perform comparably on other metrics to both [44, 87] even when we do not have human joint pose optimization based on accurate 3D scenes. Without canonicalization for human inpainting, the single-view human pose generation usually results in implausible HSI poses.

Table 2. **3D HSI Pose Generation** "SV" means only inpaint single view. "MV" means inpaint multiple views.

	Semantic Clip ↑	Entropy ↑	Cluster Size ↑	Non-Collision ↑	Contact↑
COINS (3D) [87]	**0.2624**	2.695	0.813	0.974	0.969
GenZi (3D+MV) [44]	0.2521	**2.779**	**0.914**	**0.983**	0.971
GenHSI (SV)	0.2578	2.601	0.852	0.980	**0.984**
GenHSI (SV) w/o can.	0.2513	2.734	0.877	0.865	0.966

Table 3. **Long-VBench Video Quality** [32] GenHSI beats commercial video customization solution across major metrics, but shows lower Dynamic Degree as the consistent background does not contribute to the optical flow used in the evaluation. GenHSI (DKF) - dual key frame improves over commercial model, but increasing keyframes in GenHSI (Ours) improves consistency, subject identity, motion smoothness, and image quality.

	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑	Dynamic Degree ↑
Kling AI 1.6 Elements [2]	0.961	0.949	0.991	0.726	**0.960**
GenHSI (DKF)	0.972	0.960	0.993	0.740	0.885
GenHSI (Ours)	**0.985**	**0.969**	**0.996**	**0.754**	0.609

Video Quality: GenHSI has the highest score in **Subject Consistency** (**0.985**), **Background Consistency** (**0.969**), **Motion Smoothness** (**0.996**), and **Image Quality** (**0.754**) that beating commercial solutions. Our videos exhibit limited **Dynamic Degree** due to their static backgrounds, which do not contribute to the optical flow used for dynamic degree measurement. And we also evaluate the effectiveness of keyframing based on atomic interactive action parsing in **GenHSI (DKF)**. In this setting, only two keyframes will be used to generate the long video based on the text prompt composed by text-based motion planning. The results in Fig. 6 and Tab. 3 indicate that although we do not optimize the parameters of the video generative model, our keyframe prompting strategy can significantly preserve the character & scene identity over a longer video duration. Please find more details in the video results in supplementary materials.

Conclusion: GenHSI is a training-free approach that generates personalized, controllable HSI videos by leveraging 3D-aware keyframe prompting on off-the-shelf video generative models. Instead of adding new modules to large video generation models with heavy training, our key insight is to divide complex HSI video generation tasks into three stages of script writing, pre-visualization, and animation to generate 3D-aware keyframes and atomic parsed actions for video synthesis. Results demonstrate the efficacy of our single-view 3D-aware keyframing approach for enhancing 3D human-object contact and subject consistency in HSI video generation.

Limitations & Future Work: Our work is limited by the capabilities of existing image inpainting models in inserting humans at a plausible scale and feed-forward Gaussian avatar generation in high-fidelity appearance modeling. When generating 3D-aware keyframes, we naively render the composed 3D Gaussians of the avatar scene. Although visual quality could be further enhanced by large image editing models for harmonization and lighting effects, our lower-quality keyframe prompts still yield substantial improvements in HSI video generation.

Acknowledgment: This work was supported by Meta, and an AWS Cloud Credits award.

References

[1] Realistic vision inpainting. https://huggingface.co/Uminosachi/realisticVisionV51_v51VAE-inpainting. 5

[2] Kling AI. Kling ai 1.6 elements. https://klingai.com/image-to-video/multi-id/, . 2, 8

[3] Kling AI. Kling ai 1.6 frames. https://klingai.com/image-to-video/frame-mode/, . 2, 6

[4] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 3

[5] 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. 2

[6] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2

[7] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Y. Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7611–7620, 2023. 2

[8] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1855–1866, 2024. 3

[9] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao-Liang Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7310–7320, 2024. 2

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ArXiv*, abs/2310.00426, 2023. 2

[11] Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. *arXiv preprint arXiv:2412.02690*, 2024. 2

[12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. *arXiv preprint arXiv:2501.06187*, 2025. 2

[13] Weiliang Chen, Fangfu Liu, Diankun Wu, Haowen Sun, Haixu Song, and Yueqi Duan. Dreamcinema: Cinematic transfer with free camera and 3d character. *arXiv preprint arXiv:2408.12601*, 2024. 2

[14] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8795–8805, 2024. 2

[15] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024. 2, 3

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[17] Haopeng Fang, Di Qiu, Binjie Mao, Pengfei Yan, and He Tang. Motioncharacter: Identity-preserving and motion controllable human video generation. *ArXiv*, abs/2411.18281, 2024. 2

[18] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, and Xuansong Xie. Dreamoving: A human video generation framework based on diffusion models. *ArXiv*, abs/2312.05107, 2023. 2

[19] Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. 2025. 2

[20] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2

[21] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 2

[22] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Y. Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ArXiv*, abs/2307.04725, 2023. 2

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 3

[25] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2

[26] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2, 3

[27] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[30] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025. 2

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3

[32] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6, 8

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 2, 3, 6

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 2, 6

[35] 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 *European Conference on Computer Vision*, pages 400–419. Springer, 2024. 3, 5

[36] Hyeonwoo Kim, Sangwon Beak, and Hanbyul Joo. David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. *ArXiv*, abs/2501.08333, 2025. 2

[37] Taeksoo Kim and Hanbyul Joo. Target-aware video diffusion models. *arXiv preprint arXiv:2503.18950*, 2025. 2

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3, 4

[39] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17089–17099, 2023. 3

[40] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux. Accessed: 2024-09-24. 2

[41] Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. *ArXiv*, abs/2412.18600, 2024. 2

[42] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 5

[43] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 5

[44] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20465–20474, 2024. 2, 3, 5, 6, 8

[45] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiao wen Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model. *ArXiv*, abs/2412.00131, 2024. 2

[46] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. 2025. 2

[47] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. 2025. 2

[48] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4

[49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned

multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[50] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 2

[51] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Cinemo: Consistent and controllable image animation with motion diffusion models. *arXiv preprint arXiv:2407.15642*, 2024. 2

[52] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21181–21191, 2025. 2

[53] Chaerin Min and Srinath Sridhar. Genheld: Generating and editing handheld objects. *arXiv preprint arXiv:2406.05059*, 2024. 2

[54] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 2

[55] OpenAI. Chatgpt-4o, 2025. Accessed: 2025-03-08. 2, 3

[56] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pages 163–182. Springer, 2024. 6

[57] Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R Venkatesh Babu. Text2place: Affordance-aware text guided human placement. In *European Conference on Computer Vision*, pages 57–77. Springer, 2024. 2, 3, 6

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

[59] 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*, abs/2312.06553, 2023. 2

[60] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 2, 3, 6

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[62] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[63] Nataniel Ruiz, Yuanzhen Li, Neal Wadhwa, Yael Pritch, Michael Rubinstein, David E Jacobs, and Shlomi Fruchter. Magic insert: Style-aware drag-and-drop. *arXiv preprint arXiv:2407.02489*, 2024. 3

[64] Rahul Sajnani, Jeroen Vanbaar, Jie Min, Kapil Katyal, and Srinath Sridhar. Geodiffuser: Geometry-based image editing with diffusion models. *arXiv preprint arXiv:2404.14403*, 2024. 2

[65] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Y. Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Da, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *ArXiv*, abs/2401.15977, 2024. 2

[66] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 2

[67] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[68] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023. 2

[69] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. *arXiv preprint arXiv:2411.07232*, 2024. 3, 6

[70] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *ArXiv*, abs/2410.19115, 2024. 2, 3, 6

[71] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Pe der Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Y. Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. *ArXiv*, abs/2309.15103, 2023. 2

[72] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 3

[73] 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. 2023. 2

[74] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can:

Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 2, 3, 6

[75] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024. 5

[76] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 5

[77] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2, 3, 5

[78] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *ArXiv*, abs/2310.12190, 2023. 2

[79] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 2

[80] Ziyi Xu, Ziyao Huang, Juan Cao, Yong Zhang, Xiaodong Cun, Qing Shuai, Yuchen Wang, Linchao Bao, Jintao Li, and Fan Tang. Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation. *ArXiv*, abs/2411.17383, 2024. 2

[81] ChangHee Yang, ChanHee Kang, Kyeongbo Kong, Hanni Oh, and Suk-Ju Kang. Person in place: Generating associative skeleton-guidance maps for human-object interaction image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8164–8175, 2024. 3

[82] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2024. 3

[83] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[84] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6194–6204, 2020. 6

[85] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7747–7756, 2023. 2

[86] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magic mirror: Id-preserved video generation in video diffusion transformers. *ArXiv*, abs/2501.03931, 2025. 2

[87] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022. 2, 3, 6, 8

[88] Yan Zheng, Zhenxiao Liang, Xiaoyan Cong, Yuehao Wang, Peihao Wang, Zhangyang Wang, et al. Oscillation inversion: Understand the structure of large flow model through the lens of inversion method. *arXiv preprint arXiv:2411.11135*, 2024. 2

[89] Qiang Zhou, Shaofeng Zhang, Nianzu Yang, Ye Qian, and Hao Li. Motion control for enhanced complex action video generation. *ArXiv*, abs/2411.08328, 2024. 2

[90] Thomas (Hanwen) Zhu, Ruining Li, and Tomas Jakab. Dreamhoi: Subject-driven generation of 3d human-object interactions with diffusion priors. *ArXiv*, abs/2409.08278, 2024. 2