GenZI: Zero-Shot 3D Human-Scene Interaction Generation

Lei Li Angela Dai
Technical University of Munich

Figure 1. Given an arbitrary 3D scene, GenZI can synthesize virtual humans interacting with the 3D environment at specified locations from a brief text description. Our approach does *not* require any 3D human-scene interaction training data or 3D learning. By distilling interaction priors from powerful 2D vision-language models, we optimize for 3D human-scene interaction synthesis in a flexible fashion, with simple language-based control and high generality to various types of scene environments.

Abstract

Can we synthesize 3D humans interacting with scenes without learning from any 3D human-scene interaction data? We propose GenZI[1], the first zero-shot approach to generating 3D human-scene interactions. Key to GenZI is our distillation of interaction priors from large vision-language models (VLMs), which have learned a rich semantic space of 2D human-scene compositions. Given a natural language description and a coarse point location of the desired interaction in a 3D scene, we first leverage VLMs to imagine plausible 2D human interactions inpainted into multiple rendered views of the scene. We then formulate a robust iterative optimization to synthesize the pose and shape of a 3D human model in the scene, guided by consistency with the 2D interaction hypotheses. In contrast to existing learning-based approaches, GenZI circumvents the conventional need for captured 3D interaction data, and allows for flexible control of the 3D interaction synthesis with easy-to-use text prompts. Extensive experiments show that our zero-shot approach has high flexibility and generality, making it applicable to diverse scene types, including both indoor and outdoor environments.

[1]Project page: craigleili.github.io/projects/genzi

1. Introduction

Scenes are typically constructed to enable human interactions with the environment, like sitting on a couch, playing a piano, or opening a mailbox. Understanding these interactions between humans and scenes, also known as affordances [9, 17], has gained increasing attention in the fields of computer vision and graphics. In particular, achieving controllable and generalizable synthesis of human-scene interactions (or HSIs) holds immense potential for various applications, such as robotics, architectural design, video games, and virtual reality experiences, among many others.

Generating realistic humans interacting with a 3D scene is a challenging task, requiring holistic semantic understanding of both the environment and possible human actions therein. Existing approaches to HSI synthesis [11, 46, 50, 51] rely heavily on supervised training using meticulously captured data of real people interacting in 3D environments. Unfortunately, collecting large-scale datasets of 3D scenes and human interactions is exorbitantly difficult. It not only demands accurate tracking and reconstruction of both people and their environments, but also needs to ensure sufficient diversity in subjects and scenes. Existing HSI datasets currently contain very limited quantities of scenes and actions, for example, PROX [10, 51] only consisting

of 12 indoor scenes and humans interacting with 11 object categories. The scarcity of ground truth data for supervision has thus strongly limited the applicability and generalization of the learning-based approaches for synthesizing diverse sets of actions in arbitrary 3D scenes.

We thus consider an alternative perspective to HSI synthesis and pose the following question: *Can we achieve plausible HSI synthesis without using any captured 3D interaction data?* To this end, we present a novel *zero-shot* approach to 3D HSI generation. We propose the first method to leverage the powerful capabilities exhibited by recent vision-language models (VLMs) [19, 20, 30, 33, 34, 37] to synthesize plausible 2D images of human interactions, and introduce a robust optimization to distill inferred 2D pose information into 3D human synthesis in a 3D scene.

More concretely, given a 3D scene, a text prompt and a coarse point location of the desired interaction, GenZI optimizes for the pose and shape of a 3D human performing the action in the scene, guided by a large VLM [34]. We first leverage the VLM to imagine possible 2D humans by inpainting images from multiple rendered views of the scene. We automate this 2D human insertion process with a dynamic masking scheme that automatically updates proposed masks through the inpainting process, eliminating the need for manual specifications of human inpainting regions. We then lift these 2D interaction hypotheses to 3D and optimize for a parametric 3D human body model [23, 29] that is most consistent with the 2D pose guidance. We further refine the generated 3D human in the scene by iterating through the VLM-based 2D inpainting and robust 3D lifting stages. We demonstrate the flexibility and generality of GenZI in various types of 3D environments (Fig. 1), encompassing both indoor and outdoor scenes.

In summary, our contributions are as follows:

- We introduce GenZI, the first zero-shot approach to generating realistic 3D humans interacting with a 3D scene from natural language prompts. GenZI does not require any supervision from 3D interaction data, thus enabling flexible synthesis across diverse scenes and actions.
- We propose a dynamically-masked inpainting scheme that allows for the synthesis of plausible 2D human-scene compositions via VLMs without requiring manually-specified human inpainting masks.
- We develop a robust 3D pose optimization to lift various inferred images of human interactions to a view-consistent, realistic 3D HSI synthesis.

2. Related Work

3D Human-Scene Interaction Synthesis. Synthesizing humans in scenes is an important, challenging task in computer vision and graphics, as it models complex high-level semantic understanding such as affordances and interactions. Existing approaches for human-scene interaction synthesis [11, 12, 16, 21, 27, 39, 43, 44, 50, 51] focus on learning priors from available collected data of people interacting with 3D indoor scenes [10] in a supervised manner.

The early work PiGraphs [39] introduced a probabilistic graphical model of human-object interactions, generating static human-object placements by pose sampling and model retrieval. Zhang *et al*. [50] introduced a generative model of human-scene interaction, using a conditional variational autoencoder (CVAE) to model the distribution of 3D human poses conditioned on scene depth and semantics. PLACE [49] explicitly represented human-scene interactions with Basis Point Sets (BPS) encoding [32], training a CVAE to synthesize such representations for generating natural human poses and contact relations within scenes. POSA [11] proposed an ego-centric representation by augmenting the SMPL-X model [29] with contact labels and scene semantics, using a CVAE conditioned on SMPL-X vertex positions to model potential interactions. Recently, COINS [51] presented a method for compositional human-scene interaction synthesis with high-level semantic control, using transformer-based CVAEs conditioned on provided 3D objects and interaction semantics to regress human body poses and contact features.

However, these works all require supervision from 3D human-scene interactions – such datasets are difficult and expensive to capture and annotate, for instance, the widely used PROX dataset [10, 51] only has 8 reconstructed indoor scenes for training and 4 for testing, capturing human interactions with 11 object categories. As such, these methods are limited to in-domain synthesis, prohibiting applicability to general 3D scene settings with arbitary objects and arrangements. In this work, we deviate from the learning-based approach and propose to take advantage of established large vision-language models [20, 33, 34, 37] for human-scene interaction synthesis, thereby bypassing the requirement for data capture and 3D learning.

3D Human Estimation from RGB Images. Over the past decades, significant progress has been made in 3D human estimation from RGB images [7, 26, 31, 38]. Among those, the prominent work SMPLify-X [29] fits the SMPL-X model to 2D joints estimated from a single image through optimization. Hassan *et al*. [10] build upon SMPLify-X and estimate 3D human poses by incorporating additional physical contact and penetration contraints between human and 3D scene. Learning-based approaches [3, 5, 18, 22, 35, 48, 53] have received much research attention in recent years towards tackling 3D human estimation from monocular images with improved hand and face estimation. Our work leverages 2D pose reasoning from multiple different 2D view hypotheses, and proposes a robust 3D formulation for aggregating the various 2D hypotheses to a consistent

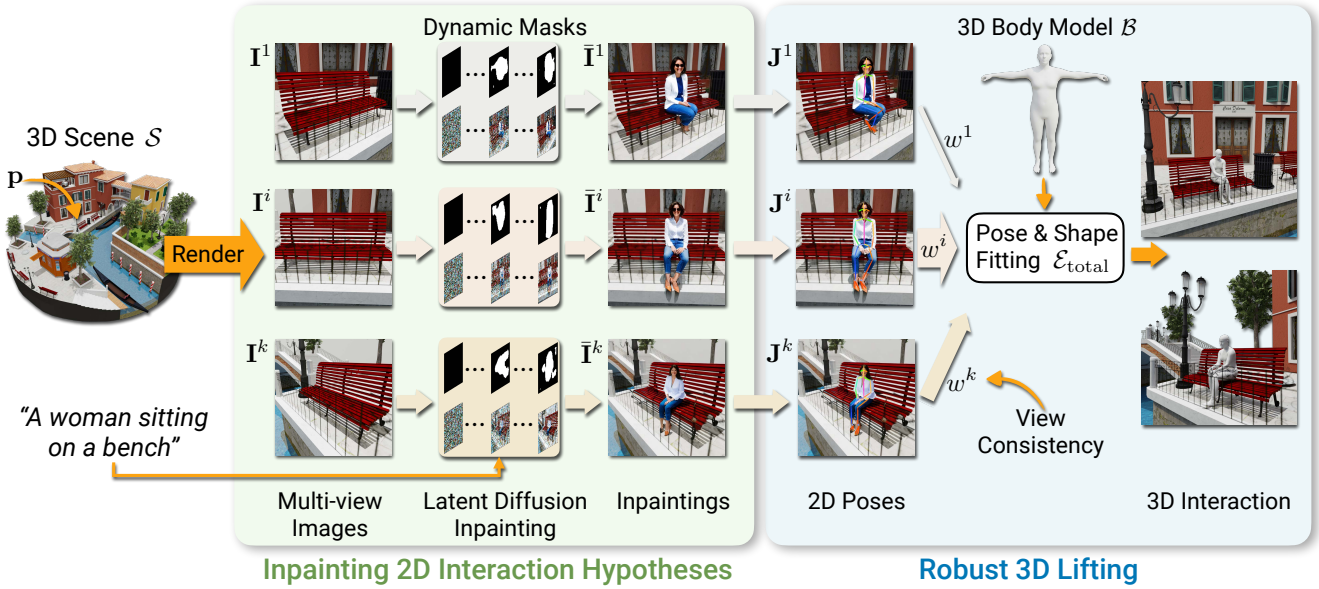

Figure 2. GenZI distills information from vision-language model for 3D human-scene interaction. We first leverage large vision-language models to synthesize possible 2D humans interactions with the 3D scene $\mathcal{S}$ by employing latent diffusion inpainting [34] on multiple rendered views of the environment at location $\mathbf{p}$ using our dynamic masking scheme to automatically estimate inpainting masks. We then lift these 2D hypotheses to 3D in a robust optimization for a 3D parametric body model $\mathcal{B}$ (SMPL-X [29]) that is most consistent with detected 2D poses in the inpainted 2D hypotheses. This produces a semantically consistent interaction that respects the scene context, without requiring any 3D human-scene interaction data.

3D human body interacting with a 3D scene.

Distilling Prior from Vision-Language Models. Recent advances in powerful vision-language models [20, 33, 34, 37] have also inspired various works aiming to distill information learned through the models to various tasks, including 2D panoptic segmentation [45], 3D semantic segmentation [36], 3D scene generation [15], and synthesis of images with hand-object interactions [47]. Our approach leverages the 2D generative capacity of these models to convey information about possible human-scene interactions, which we then lift to a 3D, consistent interaction.

3. Method

3.1. Overview

Our objective is to synthesize plausible 3D humans interacting with a 3D scene, guided by input text descriptions, in the absence of 3D interaction data capture for learning. We present GenZI, a novel optimization-based multi-view approach that leverages large VLMs to infer spatial relations of interactions between a human and the scene. Fig. 2 shows an illustration of our approach.

GenZI takes as input a 3D scene $\mathcal{S}$, a text prompt Γ describing the desired interaction, and an approximate point location $\mathbf{p} \in \mathbb{R}^3$ in the scene around which the interaction should occur. Our approach generates a posed 3D human $\mathcal{B}$ as output, performing the specified action in the scene. We adopt SMPL-X [23, 29] to parameterize the 3D human $\mathcal{B}$, as it provides a fully differentiable function mapping a set of pose and shape parameters $(\mathbf{R}, \mathbf{t}, \mathbf{\Theta}, \mathbf{\Phi})$ to a 3D human mesh with vertices $\mathbf{V}$ and faces $\mathbf{F}$. We thus optimize for $(\mathbf{R}, \mathbf{t}, \mathbf{\Theta}, \mathbf{\Phi})$ characterizing $\mathcal{B}$ in the scene $\mathcal{S}$, with $\mathbf{R} \in \mathbb{R}^6$ denoting the global orientation as a continuous rotation representation [52], $\mathbf{t} \in \mathbb{R}^3$ the global translation, $\mathbf{\Theta} \in \mathbb{R}^{32}$ the body pose in the latent space of VPoser [29] which decodes to body joint rotations $\hat{\mathbf{\Theta}} \in \mathbb{R}^{21\times3}$, and $\mathbf{\Phi} \in \mathbb{R}^{10}$ representing the body shape as blend shape coefficients.

GenZI synthesizes the desired interaction between the 3D human $\mathcal{B}$ and the scene $\mathcal{S}$ by distilling information from VLMs through 2D human inpainting followed by robust 3D lifting. We first generate a collection of plausible 2D human-scene compositions by employing a large VLM to inpaint humans into multiple rendered views of the scene $\mathcal{S}$ (Sec. 3.2). We then introduce a robust 3D lifting procedure that optimizes for the pose and shape of the human $\mathcal{B}$, guided by consistency with the 2D interaction hypotheses (Sec. 3.3). We further refine the posed 3D human $\mathcal{B}$ by iteratively updating the 2D inpaintings and 3D optimization.

3.2. Inpainting Multi-view Interaction Hypotheses

We first leverage a VLM to generate 2D hypotheses of potential human interactions in the scene by automatically inpainting humans into multiple rendered views of $\mathcal{S}$.

Multi-view Rendering. To capture scene context for 2D

human inpainting, we render multiple views of the 3D scene $\mathcal{S}$ from k virtual cameras looking at $\mathbf{p}$. Cameras are randomly sampled on a hemisphere and filtered according to the visibility of $\mathbf{p}$; we refer to the supplementary material for additional camera setup details. We denote the rendered scene images as $\{\mathbf{I}^i\}_{i\in[1,k]}$. To simplify notation, we omit the superscript i indexing each view in this section.

Inpainting with Dynamic Masking. Given a rendered scene image $\mathbf{I}$ and the text prompt Γ, we leverage a state-of-the-art latent 2D diffusion model [34] to generate a new image $\bar{\mathbf{I}}$, where a human is inpainted into the scene image $\mathbf{I}$ while adhering to the specified interaction and 2D scene context. In practice, we opt for the popular Stable Diffusion Inpainting model [1] as the latent diffusion implementation.

The latent diffusion model, denoted as $\Omega(\mathbf{z}_t, \mathbf{M}, \mathbf{I}, \Gamma, t)$, performs image inpainting by progressively denoising a noisy latent $\mathbf{z}_t$ at each time step t. During the denoising diffusion process, the binary mask $\mathbf{M}$ defines the inpainting region in the image $\mathbf{I}$; however, specifying this mask typically requires manual effort [1]. Thus, we develop a fully-automated inpainting process by automatically generating the mask. Note that using a random or fixed human mask naively can lead to incorrect inpaintings, as this often results in scene context incorrectly masked, *e.g.*, the objects affording the desired action may be entirely masked out, leading to the generated 2D human poses to be incoherent with the scene, producing undesirable 3D HSI synthesis.

We propose a masking scheme that dynamically adapts the mask through the denoising process by leveraging the internal cross-attention maps [4, 13, 42] from Ω to propose masks. The cross-attention maps capture rich semantic correlations between image pixels and input text tokens, playing a crucial role in guiding image generation. Let $\mathbf{A} \in \mathbb{R}^{hw\times n}$ denote the cross-attention map between an image feature map of hw pixels and the text prompt Γ with n tokens, normalized row-wise using `softmax`. Here, $\mathbf{A}[i,j]$ signifies the influence of the j-th token on the i-th pixel, and $\mathbf{A}[:,j]$ forms a heat map of image regions to be filled with content related to the j-th token.

Using the cross-attention map $\mathbf{A}_t$ at each time step t of the diffusion process, we can dynamically derive a mask related to human tokens. Specifically, at time step t, we obtain $\mathbf{z}_{t-1}, \mathbf{A}_t \leftarrow \Omega(\mathbf{z}_t, \mathbf{M}_t, \mathbf{I}, \Gamma, t)$ after denoising. To create a human inpainting mask $\mathbf{M}_{t-1}$ for time step $t-1$, we extract heat maps from $\mathbf{A}_t$ corresponding to the tokens referring to the human (*e.g.*, "woman" or "man"), followed by summation of these extracted heat maps across the tokens and binarization. At the initial time step T, we initialize $\mathbf{M}_T$ as an empty mask. Our dynamic masking approach enables the synthesis of a 2D human-scene composition image $\bar{\mathbf{I}}$ without the need for manually-specified human inpainting masks. We illustrate our dynamic masking scheme in Fig. 3.

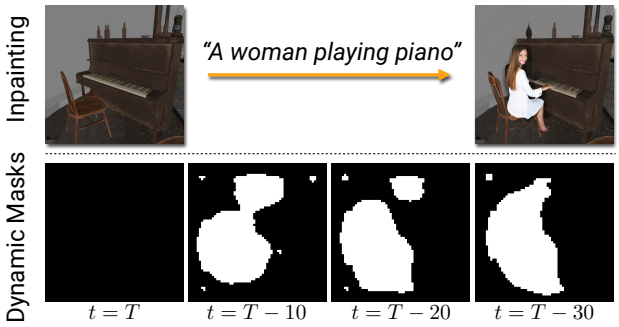

Figure 3. Human inpainting with dynamic masking. Top: Given a scene image and a text prompt, a human is inpainted into the image *without* a mask specifying the inpainting region for latent diffusion. Bottom: The masks generated by our dynamic masking scheme based on cross-attention maps at different diffusion time steps adaptively shift to find the region of interest.

3.3. Robust Lifting to a 3D Interaction

Given the multi-view scene images inpainted with human subjects $\{\bar{\mathbf{I}}^i\}_{i\in[1,k]}$, our aim is to optimize for the pose and shape parameters $(\mathbf{R}, \mathbf{t}, \mathbf{\Theta}, \mathbf{\Phi})$ of a 3D human $\mathcal{B}$, guided by the multi-view interaction cues. We optimize for $\mathcal{B}$ by matching it with the 2D poses extracted from $\{\bar{\mathbf{I}}^i\}$. Since the 2D hypotheses may not be consistent across views, we formulate this as a robust optimization, simultaneously optimizing for the set of views most consistent with $\mathcal{B}$.

2D Pose Estimation To distill interaction guidance for the 3D HSI synthesis, we compute a 2D pose representation for the inpainted human in each image $\bar{\mathbf{I}}^i$. We use AlphaPose [6], an off-the-shelf pose estimation approach that infers a set of 2D joint positions $\mathbf{J}^i$ and the corresponding joint confidence scores $\mathbf{c}^i$ for the human subject in image $\bar{\mathbf{I}}^i$.

These 2D pose hypotheses $\{\mathbf{J}^i, \mathbf{c}^i\}_{i\in[1,k]}$ are then used to steer the interaction synthesis between the 3D human $\mathcal{B}$ and the scene $\mathcal{S}$. We aim to minimize the following objective function $\mathcal{E}_{\text{total}}$:

$$\begin{aligned}\mathcal{E}_{\text{total}} =& \lambda_{\text{PF}}\mathcal{E}_{\text{PF}} + \lambda_{\text{VS}}\mathcal{E}_{\text{VS}} + \lambda_{\text{BP}}\mathcal{E}_{\text{BP}} + \\ & \lambda_{\text{BS}}\mathcal{E}_{\text{BS}} + \lambda_{\text{SC}}\mathcal{E}_{\text{SC}} + \lambda_{\text{SP}}\mathcal{E}_{\text{SP}},\end{aligned} \tag{1}$$

where the λ denote scalar weights for the energy terms: pose fitting $\mathcal{E}_{\text{PF}}$, view selection $\mathcal{E}_{\text{VS}}$, body pose $\mathcal{E}_{\text{BP}}$, body shape $\mathcal{E}_{\text{BS}}$, scene penetration $\mathcal{E}_{\text{SC}}$, and self-penetration $\mathcal{E}_{\text{SP}}$.

Robust View-consistent 3D Pose Fitting. Our primary energy term $\mathcal{E}_{\text{PF}}$ minimizes the discrepancy between the projections of the 3D pose of $\mathcal{B}$ and the inpainted poses from multiple views. However, 2D pose hypotheses can often be inconsistent across different views, due to the stochastic nature of diffusion models, leading to conflicting optimization signals for $\mathcal{B}$. To address this, we employ a robust optimization strategy that optimizes additionally for view selection weights in $\mathcal{E}_{\text{PF}}$ that promote consistency between the 3D pose and the most consistent 2D pose hypotheses.

We thus introduce a new set of *optimizable* variables $\mathbf{w} = \{w^i | w^i \in [0,1]\}_{i\in[1,k]}$, representing view consistency scores, and apply a robust kernel ρ to the per-view joint fitting constraints. The weights $\mathbf{w}$ allow the solver to adaptively focus on views with consistent inpainted poses and downweight the inconsistent ones. Our pose fitting energy $\mathcal{E}_{\text{PF}}$ with view consistency is formulated as follows:

$$\mathcal{E}_{\text{PF}} = \frac{\sum_i w^i \sum_j \mathbf{c}_j^i \rho\big(\Pi(\hat{\mathbf{J}})_j^i - \mathbf{J}_j^i\big)}{\sum_i w^i}, \tag{2}$$

where $\hat{\mathbf{J}}$ denotes the 3D joint positions of the SMPL-X model [29] differentiably computed from the pose and shape parameters $(\mathbf{R}, \mathbf{t}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$. The function $\Pi(\cdot)_j^i$ represents projection of the j-th joint in the i-th camera view, and ρ is the robust Geman-McClure error function [8].

Regularizations. Below, we describe the regularization terms adopted in our optimization.

1) We encourage the per-view weights $\mathbf{w}$ to focus on at least τ views:

$$\mathcal{E}_{\text{VS}} = \max(\tau - \sum_i w^i, 0). \tag{3}$$

2) To ensure a natural 3D pose, we impose the following energy term on the body pose parameter:

$$\mathcal{E}_{\text{BP}} = \|\boldsymbol{\Theta}\|^2 + \mathcal{E}_{\text{JA}}(\hat{\boldsymbol{\Theta}}). \tag{4}$$

The first term is the VPoser body prior [29] regularizing the latent pose $\boldsymbol{\Theta}$, and the second term $\mathcal{E}_{\text{JA}}$ is a simple angle prior on the body joint rotations $\hat{\boldsymbol{\Theta}}$, decoded from the latent pose $\boldsymbol{\Theta}$ by VPoser, to penalize extreme bending. The formulation of $\mathcal{E}_{\text{JA}}$ is provided in the supplementary material.

3) We reguarlize the shape parameter $\boldsymbol{\Phi}$ to obtain a plausible body shape via

$$\mathcal{E}_{\text{BS}} = \|\boldsymbol{\Phi}\|^2, \tag{5}$$

which measures the Mahalanobis distance between $\boldsymbol{\Phi}$ and the body shape distribution used in SMPL-X [29].

4) To ensure physical contact but also avoid penetration between the 3D human $\mathcal{B}$ and the scene $\mathcal{S}$, we formulate the spatial constraints as:

$$\mathcal{E}_{\text{SC}} = \begin{cases} \min_{\mathbf{v}\in\mathbf{V}} \Psi(\mathbf{v}), & \Psi(\mathbf{v}) > 0 \; \forall \mathbf{v} \in \mathbf{V} \\ \sum_{\mathbf{v}\in\mathbf{V}} |\min\big(\Psi(\mathbf{v}), 0\big)|, & \text{otherwise} \end{cases} \tag{6}$$

where Ψ is a pre-computed signed distance field (SDF) [10] of the scene $\mathcal{S}$. When $\Psi(\mathbf{v})$ has a negative sign, it indicates that the body vertex $\mathbf{v}$ is located inside the nearest scene object (*i.e.*, penetration). Conversely, a positive sign means that $\mathbf{v}$ is positioned on the outside.

5) Finally, to resolve penetration within the human body $\mathcal{B}$ itself, we include a self-penetration energy $\mathcal{E}_{\text{SP}}$ based on detecting colliding body triangles using Bounding Volume Hierarchies (BVH) [40]. We refer the reader to [2, 10, 41] for more formulation details on $\mathcal{E}_{\text{SP}}$.

3.4. Iterative Refinement

A posed 3D human $\mathcal{B}$ interacting with the 3D scene $\mathcal{S}$ is generated after applying the VLM-based 2D inpainting (Sec. 3.2) and the robust 3D lifting (Sec. 3.3). To improve the synthesis and consistency of the interaction results, we employ an iterative refinement scheme over the aforementioned inpainting and 3D lifting. During refinement, we render the silhouette of the posed 3D human $\mathcal{B}$ in each camera view, and use it as a more precise and consistent mask $\mathbf{M}$ in the latent diffusion inpainting Ω [34], replacing dynamic masking. By doing so, the consistency among the 2D pose hypotheses inpainted into the multi-view scene images gradually improves, thus leading to an improved 3D HSI synthesis outcome.

3.5. Implementation Details

We implemented GenZI with PyTorch [28]. For VLM-based 2D inpainting, we use $k = 16$ cameras for multi-view rendering. We use 50 denoising steps in Stable Diffusion Inpainting with a state-of-the-art diffusion sampler [25]. For dynamic masking during inpainting, we aggregate cross-attention maps with a resolution of 16×16. Input text prompts Γ (*e.g.*, "sitting on a bench") are all appended with fixed prefix "a woman" and fixed suffix "wearing a white shirt and blue pants, full body" to better constrain generation. For robust 3D lifting, we set $\tau = 3$, and optimize the energy $\mathcal{E}_{\text{total}}$ using gradient descent [24] for 1.6K steps, which takes ~ 3 minutes on an NVIDIA A100 GPU. In the iterative refinement, we dilate the rendered human silhouette with a kernel size of 11×11 for mask generation, and the refinement is performed once.

4. Experiments

We demonstrate the effectiveness and generality of our approach GenZI on a diverse collection of 3D scene models from Sketchfab.com. We conduct both quantitative and qualitative evaluations to compare GenZI with alternative baselines approaches [10, 51] to our new task.

Dataset. In our Sketchfab dataset, we gathered 8 large-scale 3D scenes encompassing a variety of indoor and outdoor environments with diverse geometric structures, including a realistic Venice city, a gym, and a cartoon-style food truck. We collected 4-5 text prompts per scene describing human interactions with the scene for specified approximate point locations, resulting in 38 actions for evaluation.

Baselines. To the best of our knowledge, there are no baselines that estimate 3D human-scene interactions based on natural language text input in a zero-shot fashion. We thus perform comparisons with related baseline approaches:

- COINS [51] is a state-of-the-art approach estimating 3D humans in indoor 3D scans with a fixed vocabulary of actions and objects, given object segmentations. It takes as input $\langle \text{action}, \text{object} \rangle$ pairs for semantic control, and requires full supervision in its CVAE training, using captured 3D interaction data with both instance segmentations of 3D objects and action labelling. Due to being trained on a small, closed set of indoor interactions, we adapt COINS to a subset of the most similar Sketchfab actions by manually segmenting corresponding 3D objects from the scenes similar to its indoor training data and mapping the text prompts to its $\langle \text{action}, \text{object} \rangle$ input. Note that our approach GenZI does *not* require any 3D scene segmentations.
- Hassan *et al*. [10] perform 3D human estimation from a single RGB image. To adapt this to Sketchfab, we reuse the multi-view scene images $\{\bar{\mathbf{I}}^i\}_{i\in[1,k]}$ inpainted with humans from our dynamic masking scheme (Sec. 3.2), where the view with the best image-text cosine similarity (*i.e*., CLIP score [33]) is used as the input for 3D human estimation under the known virtual camera parameters.
- Ours-Single View: Finally, we consider a baseline leveraging 3D body estimation from our method, but limited to only one inpainted view. The same best view image, as described above, is also as input for this baseline.

Evaluation Metrics. To measure 3D HSI quality, we conduct perceptual studies and compute metrics including semantic consistency, diversity, and physical plausibility.

We first carry out two perceptual studies to assess the realism and semantic accuracy of the synthesized interactions. The first is a binary-choice study, where interaction samples generated by two different methods based on the same text prompt are shown, and the participants are asked to choose the sample that is more realistic and better matched the text. The second study is a unary test, where for each interaction sample, the participants are asked to rate the realism and consistency between the shown sample and text prompt from 1 (strongly disagree) to 5 (strongly agree).

To evaluate the semantic consistency between a synthesized 3D interaction and the input text prompt, we calculate the CLIP score [33], where the 3D interaction is re-rendered into k view images, and the image-text cosine similarities from CLIP ViT-B/32 are averaged across all views.

Additionally, we include quantitative metrics from existing works to evaluate diversity and physical plausibility. We observe that these metrics often do not reflect perceptual quality, as generated bodies can be diverse but have interpenetrations without demonstrating any physical or semantic coherence. Nevertheless, to measure the synthesis diversity, we follow [50, 51] and cluster the SMPL-X parameters of the generated humans into 20 clusters with K-means and compute the entropy of the cluster ID histogram of all the samples. We also calculate the cluster size as the mean distance to cluster centers. For the physical plausibility, we evaluate the collision and contact between body meshes and scene meshes, following [50, 51]. The non-collision score is computed as the ratio between the number of body mesh vertices with positive SDF values (Sec. 3.3) and the number of all body mesh vertices. The contact score is defined as the ratio between the number of body meshes with scene contact and the number of all generated body meshes. A body mesh is in contact with the scene if any body mesh vertex has a non-positive SDF value.

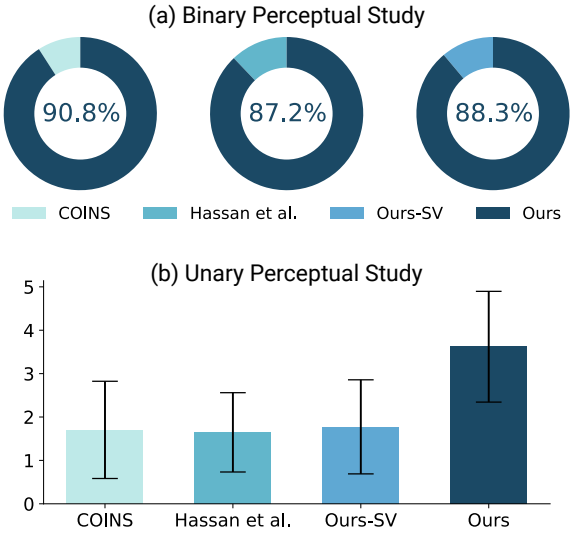

Figure 4. User study of 3D human-scene interaction synthesis on the Sketchfab dataset, where participants show a strong preference for the generations by our approach, in comparison with all baselines, COINS [51], Hassan *et al*. [10], and Ours-Single View.

4.1. Comparison to Alternative Approaches

Quantitative Evaluation. Fig. 4 shows the results of the perceptual studies collected from 30 participants across a binary study and a unary study. In the binary study, we observe that participants overwhelmingly favor the generations by our GenZI compared to all baselines – more than 87% of the time. In the unary test, the average realism rating for our interaction generations is 3.6, the highest compared to the baselines, which are all below 2.0. These perceptual results strongly indicate that our GenZI can synthesize realistic 3D humans interacting with various 3D scenes without requiring any captured 3D interaction data.

Tab. 1 presents the quantitative evaluation of semantic consistency, diversity, and physical plausibility on the full Sketchfab dataset. Our approach achieves the best semantic consistency score, echoing the strong user preference for

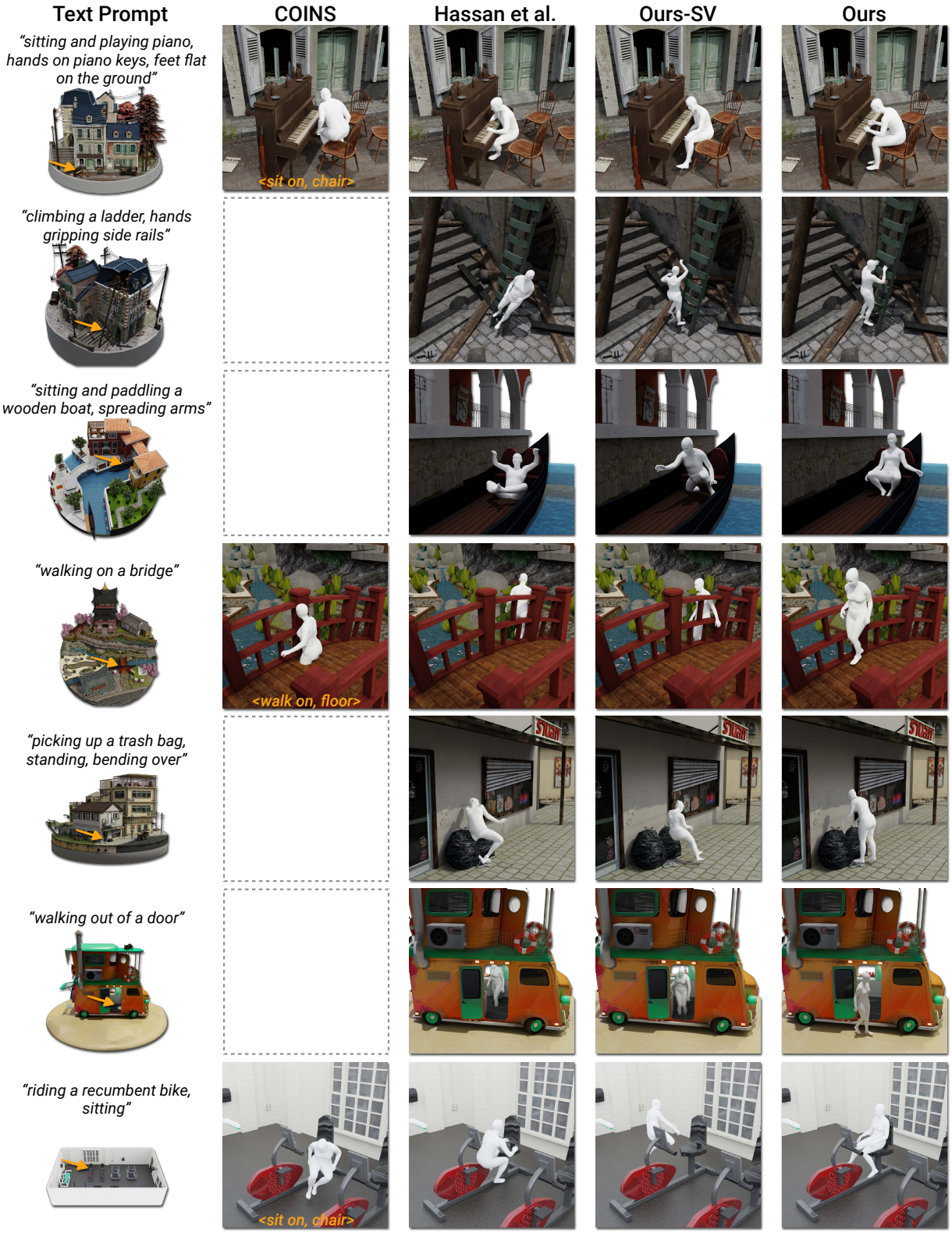

Figure 5. Qualitative results on the Sketchfab dataset. Our GenZI synthesizes more realistic 3D human-scene interactions and generalizes better across diverse scene types, compared to the baselines COINS [51], Hassan *et al.* [10], and Ours-Single View. For COINS, we show the used ⟨action, object⟩ labels from its closed set of indoor interactions; its closed setting can lead to degraded results from out-of-distribution object classes (e.g., curved bridge deck as the floor, chair at a different height or shape than those in the training set).

Method	Semantics CLIP ↑	Diversity Entropy ↑	Diversity Cluster Size ↑	Physical Plausibility Non-collision ↑	Physical Plausibility Contact ↑
Hassan *et al.* [10]	0.2598	2.7014	1.1907	0.8824	0.9669
Ours-SV	0.2613	2.6452	**1.5813**	0.9765	**1.0000**
Ours	**0.2710**	**2.7304**	1.0500	**0.9767**	0.9868

Table 1. Quantitative comparisons on the Sketchfab dataset. Our approach achieves the best semantic consistency, diversity entropy, and non-collision scores, with the contact score on par. Note that single view methods Hassan *et al.* [10] and Ours-SV tend to produce increased diversity at the cost of semantic plausibility.

	w/o DM	w/o VC	w/o IR	Ours
CLIP ↑	0.2639	0.2694	0.2664	**0.2710**

Table 2. Ablation study on the Sketchfab dataset. The semantic consistency of 3D interaction generations degrades without dynamic masking (DM), view consistency (VC), or iterative refinement (IR), compared to our full approach.

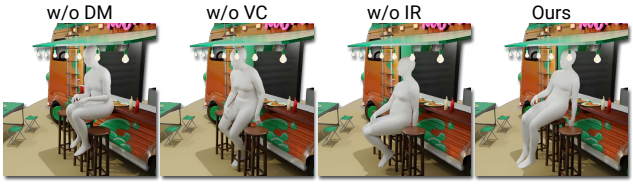

Figure 6. Visualization of our method ablations on Sketchfab dataset for the input text: "*sitting on a bar stool*". Without dynamic masking (DM) or view consistency (VC), the person floats above the middle stool. Without iterative refinement (IR), the person penetrates the stool. Our full approach results in a more realistic synthesis.

GenZI in the perceptual studies. We note that since both Hassan *et al.* [10] and Ours-Single View operate on single inpainted view samples, results can be very diverse but lack 3D plausibility. This suggests that the CLIP score is a more reliable metric for assessing the HSI synthesis quality, compared to the diversity and physical plausibility metrics, which reflect less about the generation realism. Nevertheless, our approach has the best diversity entropy and non-collision scores, with the contact score on par.

Qualitative Evaluation. We show qualitative comparisons in Fig. 5. COINS is severely limited by its training on the closed set of indoor interactions, and thus fails to generalize to outdoor scenes and unseen objects (e.g., no curved floors exist during training, and limited sets of heights and shapes of chairs). As Hassan *et al.* [10] and Ours-Single View operate from single views, they both suffer from insufficient pose constraints from other views for plausible interaction generation. In contrast, our approach demonstrates high flexibility and generality to a diverse set of 3D indoor and outdoor scenes by leveraging large VLMs to imagine multi-view interaction hypotheses and then robust 3D lifting.

4.2. Ablation Studies

We conduct ablation studies on the Sketchfab dataset to validate the effectiveness of our proposed dynamic masking scheme (Sec. 3.2), robust 3D lifting with view consistency (Sec. 3.3), and iterative refinement (Sec. 3.4). Results are presented in Fig. 6 and Tab. 2.

Dynamic Masking (DM). We replace our dynamic masking, used during latent diffusion inpainting, with random masking. We sample a random mask around the image center covering at least 30% of the image area, and use it as a fixed mask input to the latent diffusion model Ω. We observe that random masking results in noticeably worse quality of interaction synthesis (Tab. 2), and incoherence with the scene (Fig. 6, floating on the stool). This indicates that our dynamic masking is effective in incorporating sufficient scene context for human inpainting.

View Consistency (VC). We evaluate the role of view consistency in robust 3D lifting by fixing the optimizable scores to $\mathbf{w} = \mathbf{1}$. Using all inpainted views leads to averaged, less expressive 3D human poses (Tab. 2, Fig. 6) due to potential inconsistent 2D pose hypotheses across views. By allowing the solver to adaptively focus on views with consistent inpaintings, our approach generates more realistic 3D HSIs.

Iterative Refinement (IR). Finally, we show the effectiveness of iterative refinement by applying our VLM-based 2D inpainting and robust 3D lifting only once. Tab. 2 and Fig. 6 show that iterative refinement improves synthesis quality.

Limitations. Our approach is limited by the inpainting capability of latent diffusion models to imagine possible 2D human-scene compositions, and the diffusion models are also known to be slow at inference time due to their iterative nature [14]. Nevertheless, we believe that our approach can directly benefit from the rapid advancement of VLMs for improved HSI synthesis.

5. Conclusion

We have presented the first approach to synthesize general 3D human-scene interactions guided by text inputs. Key to our approach is effective distillation of knowledge from large vision-language models, enabling generating 3D humans in scenes without requiring any 3D interaction data for training. We leverage these powerful vision-language models to generate hypotheses for inpainted 2D human-scene interactions. We then formulate a robust optimization to lift the hypotheses to 3D in a view-consistent fashion by simultaneously optimizing for the most informative 2D hypotheses. Our approach is flexible and can be applied to general scene settings for a variety of actions. We believe this opens up new opportunities for 3D understanding without requring expensive capture of 3D/4D data.

Acknowledgements. This project is funded by the ERC Starting Grant SpatialSem (101076253), and the German Research Foundation (DFG) Grant "Learning How to Interact with Scenes through Part-Based Understanding", and supported in part by a Google research gift.

References

[1] Stable Diffusion v2. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting. Accessed: 2023-09-27. 4

[2] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 5

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. *arXiv*, 2023. 2

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM TOG*, 2023. 4

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 2

[6] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE TPAMI*, 2022. 4

[7] Dariu M Gavrila. The visual analysis of human movement: A survey. *CVIU*, 1999. 2

[8] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. 1987. 5

[9] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978. 1

[10] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019. 1, 2, 5, 6, 7, 8

[11] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *CVPR*, 2021. 1, 2

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

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 8

[15] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2Room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, 2023. 3

[16] 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

[17] Kurt Koffka. Principles of gestalt psychology. 1935. 1

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 3

[21] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3D indoor environments. In *CVPR*, 2019. 2

[22] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023. 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 2, 3

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv*, 2022. 5

[26] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 2006. 2

[27] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. iMapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019. 2

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 5

[29] 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. 2, 3, 5

[30] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. H. Bermano, E. R. Chan, T. Dekel, A. Holynski, A. Kanazawa, C. K. Liu, L. Liu, B. Mildenhall, M. Nießner, B. Ommer, C. Theobalt, P. Wonka, and G. Wetzstein. State of the art on diffusion models for visual computing. *arXiv*, 2023. 2

[31] Ronald Poppe. Vision-based human motion analysis: An overview. *CVIU*, 2007. 2

[32] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, 2019. 2

[33] 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. 2, 3, 6

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5

[35] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV*, 2021. 2

[36] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 3

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2, 3

[38] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *CVIU*, 2016. 2

[39] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: learning interaction snapshots from observations. *ACM TOG*, 35(4):1–12, 2016. 2

[40] Matthias Teschner, Stefan Kimmerle, Bruno Heidelberger, Gabriel Zachmann, Laks Raghupathi, Arnulph Fuhrmann, M-P Cani, François Faure, Nadia Magnenat-Thalmann, Wolfgang Strasser, et al. Collision detection for deformable objects. In *CGF*, 2005. 5

[41] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 5

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[43] 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

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

[45] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 3

[46] Haibiao Xuan, Xiongzheng Li, Jinsong Zhang, Hongwen Zhang, Yebin Liu, and Kun Li. Narrator: Towards natural control of human-scene interaction generation via relationship reasoning. *arXiv*, 2023. 1

[47] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 3

[48] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *IEEE TPAMI*, 2023. 2

[49] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3D environments. In *3DV*, 2020. 2

[50] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *CVPR*, 2020. 1, 2, 6

[51] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, 2022. 1, 2, 5, 6, 7

[52] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[53] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, 2021. 2