Custom Condition Generation for Zero-Shot Human-Scene Interactions Synthesis

Ryosuke Kawamura[1], Zoltán Á. Milacski[2], Fernando de la Torre[2], László A. Jeni[2], Koichiro Niinuma[1]
[1] Fujitsu Research of America, Pittsburgh PA, USA
[2] Robotics Institute, Carnegie Mellon University, Pittsburgh PA, USA

***Abstract*— Existing methods for creating human interactions within scenes show promise for common interactions but often fail with less frequent ones. To overcome this, we introduce a new approach that creates tailored conditions for generating these interactions without previously seen examples. This method leverages the strengths of both large language models (LLMs) and vision-language models (VLMs). Unlike the GenZI, the current state-of-the-art approach, which struggles with rare interactions due to its reliance on VLM inpainting, our method follows a three-step process: first, we generate a preliminary human posture using VLMs and then estimate this posture in three dimensions. Next, we refine the conditions to fit the specific scene and interaction by analyzing the inputs with both LLMs and VLMs. Finally, we fine-tune the placement, orientation, and posture of the human figure using specific optimization techniques. Our experimental results show that this method performs well across a wide range of interactions, including those that are less common.**

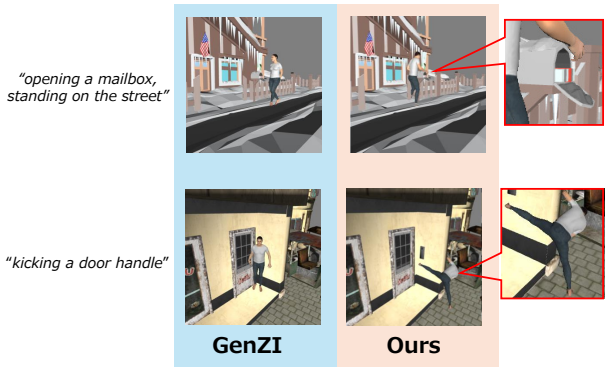

Fig. 1. GenZI [18] struggles to generate uncommon interactions such as "kicking a door handle". Our proposed method successfully synthesizes interactions, including those that are rare and uncommon.

I. INTRODUCTION

In everyday human activities, interactions with surrounding objects range from common actions like sitting on a chair to less typical behaviors such as kicking a chair. The ability to accurately simulate and manage this variety of human-scene interactions is essential for the development of applications in augmented reality (AR), virtual reality (VR), digital twin technologies, computer games, and data augmentation [44].

A key challenge in generating human-scene interactions is ensuring a wide variety of interactions, including both frequent and infrequent ones. For example, interactions with a chair are not limited to "sitting"; "kicking a chair" and "handstanding on a chair" are also possible. Existing works [41], [13], [40], [35] mainly focus on learning-based methods, which face challenges in generating diverse interactions. These methods rely on ground-truth data, such as 3D scenes and 3D motion data of human interactions [12], [41], [39], [15], [21], [4]. However, the high cost of gathering this data limits its availability and restricts the variety of actions and scenes that can be effectively modeled and reproduced.

GenZI [18] addresses the challenge of generating diverse human-scene interactions with a zero-shot synthesis method. It uses natural language descriptions and coarse point locations to guide the inpainting process of Stable Diffusion [26]. This process creates human figures in scene images, which are then lifted to 3D human models. GenZI shows promise in generating standard interactions without relying on datasets. However, it struggles with uncommon interactions like "kicking a chair." This limitation arises from the significant influence of input images and their contained objects on the inpainting process, as discussed in [36].

To address this limitation, we propose a new approach that consists of three steps: first, we generate an initial posture; second, we customize conditions to determine the plausible position and orientation for natural interactions; and third, we optimize the human location, orientation, and posture using loss functions derived from these conditions. This method faces two challenges: generating an initial posture that aligns with each text, and correctly locating the human model in positions and orientations that vary with each interaction described in the text. For example, in "sitting on a chair," the hips should contact the chair, while in "kicking a chair," the foot should touch the chair. First, we generate interaction images from scratch, separate from the scene to get an initial posture, using Text2Image models (T2I) [25], [27], [8], [26] based on the text. This avoids training data bias and allows diverse postures. However, natural positions and orientations within the 3D scene vary with each prompt. Second, we generate custom conditions to achieve plausible positions and orientations for each interaction using both Visual Language Models (VLMs) and Large Language Models (LLMs) [17], [19], [43], [2], [29], [30], [6], [16], [3], [1]. Both VLMs and LLMs are employed to identify body parts involved in each interaction, utilizing VLMs' image analysis capabilities [32] and LLMs' knowledge [42]. We define loss functions based on specific conditions and optimize these customized loss

functions. Our approach synthesizes diverse and realistic human-scene interactions across various scenarios, as illustrated in Fig. 1.

Our contributions are threefold:

- We introduce a novel method that generates common and even uncommon interactions in a zero-shot manner. This method enables the generation of human-scene interactions that current methods cannot achieve.
- We propose customizing loss functions for each interaction using VLMs and LLMs. Our method generates specific conditions for each interaction, defining customized loss functions that refine the generated interactions in terms of posture, location, and orientation.
- We validate the efficacy of our approach on a new dataset that includes both common and uncommon interactions.

II. Related Works

A. 3D Human-Scene interactions generation

To synthesize realistic human-scene interactions, various methods have been developed, each with unique approaches to modeling contact and semantics.

Learning-Based Methods: PiGraphs [28] learns a probabilistic model connecting human poses and object geometries from real-world interactions collected with RGB-D sensors. These interactions are encoded as prototypical interaction graphs that capture physical contact and visual attention linkages.

Conditional Variational Autoencoders (cVAEs) are essential in Human-Scene Interaction Synthesis. POSA [13] enhances the SMPL-X [22] model by encoding contact probability and semantic scene labels for each mesh vertex, using a cVAE to learn from body pose and shape. Similarly, PSI [40] uses a cVAE to predict plausible 3D human poses conditioned on latent scene representations, refining them with scene constraints for feasible interactions. PLACE [38], inspired by the Basis Point Set (BPS) [23] method, models proximity between the human body and the 3D scene by synthesizing minimum distances from scene mesh points to the human body surface, using a two-stage BPS encoding scheme and a cVAE to generate natural interactions.

A transformer-based cVAE network is employed in both COINS [41] and Narrator [35]. COINS uses a cVAE to encode 3D human body points and objects in a unified latent space, defining interaction semantics as action-object pairs. This enables compositional human-scene interactions without composite data. Narrator uses a cVAE for generating interactions based on scene graphs, pairing actions with body parts. It includes a mechanism for multi-human interaction generation and a Joint Global and Local Scene Graph (JGLSG) for guiding interactions by spatial relationships.

Wang et al. [33] and Huang et al. [14] employ a learning-based approach to generate human motion that aligns with the scene.

Learning-based methods are powerful strategies for interaction synthesis. However, their capability of human-scene interaction generation is limited by the scarcity of ground truth datasets and their need for supervision.

Zero-Shot Methods: GenZI [18] is the first zero-shot approach for generating 3D human-scene interactions. Given a 3D scene, a text prompt, and a coarse point location, GenZI optimizes the pose and shape of a 3D human using Stable Diffusion (SD) [26] for inpainting. Initially, SD generates possible 2D humans by inpainting images from multiple rendered views, using a dynamic masking scheme to automate updates. These 2D interaction hypotheses are then converted to 3D, optimizing a parametric 3D human body model to match the 2D pose guidance. The model is further refined through iterative SD-based 2D inpainting and 3D lifting stages. GenZI demonstrates flexibility across various 3D environments, bypassing the need for captured 3D interaction data and allowing flexible control of interaction synthesis using text prompts. However, SD tends to generate actions closely associated with specific objects, like "sitting" with a "chair," regardless of the actions specified in the prompt. This limits its range of interactions and applications.

Overall, while learning-based methods offer robust interaction synthesis, their reliance on ground truth data poses limitations. Zero-shot approaches like GenZI provide a promising alternative by enabling interaction generation without extensive training data, though they face challenges in depicting a wide variety of interactions.

B. LLM for 3D human model

Various methods have leveraged Large Language Models (LLMs) to address tasks related to the analysis and generation of 3D human models. Wang et al. [31] infer action-conditioned contact information using LLMs like GPT-3. Their method distinguishes actions like sitting and standing from human poses based on a database and obtains part-wise contact information and object scale using GPT-3. They propose a two-stage approach to estimate the shapes and poses of humans and objects before jointly reasoning about their 3D spatial arrangements.

Similarly, SINC [5] uses language models to map actions to body parts. By prompting GPT-3 [6] with action-related queries and examples, SINC combines body parts from two motions to spatially compose actions, addressing data scarcity with a GPT-guided synthetic data generation scheme. Cen et al. [7] focus on generating human motion in 3D scenes from text descriptions by constructing scene graphs and using ChatGPT to analyze relationships between scene descriptions and instructions. Their approach localizes the target object with the assistance of ChatGPT and subsequently generates human motion.

Xiao et al. [34] introduce UniHSI, which supports diverse 3D human-scene interactions through language commands. Defining interactions as a Chain of Contacts (CoC), UniHSI uses an LLM Planner to translate language prompts into task plans and a Unified Controller to execute these tasks to produce interaction movements.

Previous studies have used LLMs to address tasks related to 3D human models in various ways, achieving notable

results. Unlike these studies, our method introduces a novel approach that customizes the loss function for each interaction.

III. Method

A. Overview

Our objective is to synthesize a 3D human model that interacts with a scene based on specific instructions. Fig. 2 shows the outline of our method. The inputs to our framework include the scene S, a coarse point location of the desired interaction in a 3D scene $L \in \mathbb{R}^3$ which is manually specified, and instruction text prompt T. The output is a synthesized 3D human B that interacts naturally with the scene. We employ SMPL-X [22] to model the 3D human B, parameterizing posture and position. This model provides differential parameters (R, t, θ), where $R \in \mathbb{R}^6$ represents orientation, $t \in \mathbb{R}^3$ represents position, $\theta \in \mathbb{R}^{21\times3}$ represents posture. The vertices of the human body, modeled by SMPL-X, are represented as $\hat{V}$, and the vertices of the scene mesh are represented as V.

Our approach can be divided into two steps: initial posture generation and interaction conditions generation. The first step begins with generating an image using DALLE-2 [24] with the input text T, which illustrates the described interaction. We then estimate the 3D posture of the human in the image. This estimated posture is used as our initial posture. In the second step, we specify the contact parts of the human by analyzing the text T and the scene S with GPT-4 [1]. Based on the specified contact parts, we establish loss functions. Subsequently, the SMPL-X parameters (R, t, θ) are optimized using these loss functions to synthesize a 3D human B that interacts authentically with the object in the designated scene.

B. Initial Posture Generation

Our approach begins with generating an initial posture for the human model. This step is crucial and distinguishes our method from GenZI, which is limited by the inpainting process's capabilities. Our method leverages DALLE-2 to generate 2D images from the input text T that illustrate specified actions. To better constrain the generation process by DALLE-2, each input text T (e.g., "kicking a car") is prefixed with "a person" and suffixed with ", full body". Although the objects in the generated images may differ from objects in the scene S, the posture of the human in the 2D image aligns approximately with our target posture in scene S and its associated semantics. From these images, we estimate the SMPL-X pose parameters using Pymafx [37]. This initial posture serves as the foundation for subsequent refinements. Note that existing Text2Pose models [9], [11], [10], which generate SMPL [20] posture parameters based on text, are inadequate for this task due to the limited variety imposed by the scarcity of text-posture datasets.

C. Custom Condition Generation

To ensure the generated interactions are plausible, we define specific conditions for each interaction and design custom loss functions that account for the variability in location and orientation. Interactions involve contact with the scene, and the specific body parts that should be in contact vary with each interaction. By utilizing GPT-4, we can effectively select the relevant body parts and their corresponding areas for distance calculations.

In our approach, we categorize selected body parts into three types: Key Contact, Auxiliary Contact, and Ground Contact. These categories are designed to specify necessary conditions accurately and reduce errors caused by LLMs and VLMs. Key Contact Body Parts calculate the distance to a specific location in the scene. However, relying solely on Key Contact Body Parts can lead to unnatural positions and orientations because the segmentation of parts is not sufficiently detailed. To address this, Auxiliary Contact Body Parts are selected to calculate the distance to the scene, ensuring a more natural appearance. Finally, to prevent the human model from floating in certain interactions, we select Ground Contact Body Parts to calculate the distance to the ground. For example, consider the text prompt "sitting on a chair." In this scenario, the Key Body Parts can be the hips, the Auxiliary Contact Body Parts can be the left and right upper legs, and the Ground Contact Body Parts can be the left and right feet. The detailed steps of this selection process are explained below.

1) Key Contact Body Parts Selection: Identification of the contact scene area and the corresponding body part is crucial for generating effective human-scene interactions. For example, when a person is opening a mailbox, their hand should be in contact with the mailbox. We assume each interaction involves a single pair of the scene area and body parts. These critical elements are defined as Key Scene Areas and Key Contact Body Parts. Key Scene Areas are determined from a specified point location L, defined as $V_k = \{v \in V_s \mid \|v - L\|_2 \leq d_{(k)}(L)\}$ where $d_{(k)}(L)$ represents the k-th smallest distance from L to points in V_s, and V_s denotes the vertices of the scene S. Consequently, V_k represents the k nearest vertices to the given interaction location. The identification of Key Contact Body Parts depends on the specific interaction. Therefore, we use three main steps for this process:

- **Assessing Specific Area Contact Need** First, we query GPT-4 to determine if the interaction described in the text prompt requires contact with a specific area on the object. The answer is "Yes" or "No", denoted as A_k.
- **Identifying the the name of Key Scene Area** To obtain accurate answers about which body parts should be in contact with Key Scene Area, it is necessary to provide the names of the area where the interactions occur. We use GPT-4 to identify the name of that area. We employ multi-view rendering [18] to generate multiple images centered on the location L and query GPT-4 to determine the exact name of the area using a majority voting strategy.The majority voting is used to reduce errors in identifying the name of Key Scene Area.
- **Selecting Key Contact Body Parts** Finally, we ask GPT-4 to identify which body part should be in contact with the identified location. We pose the follow-

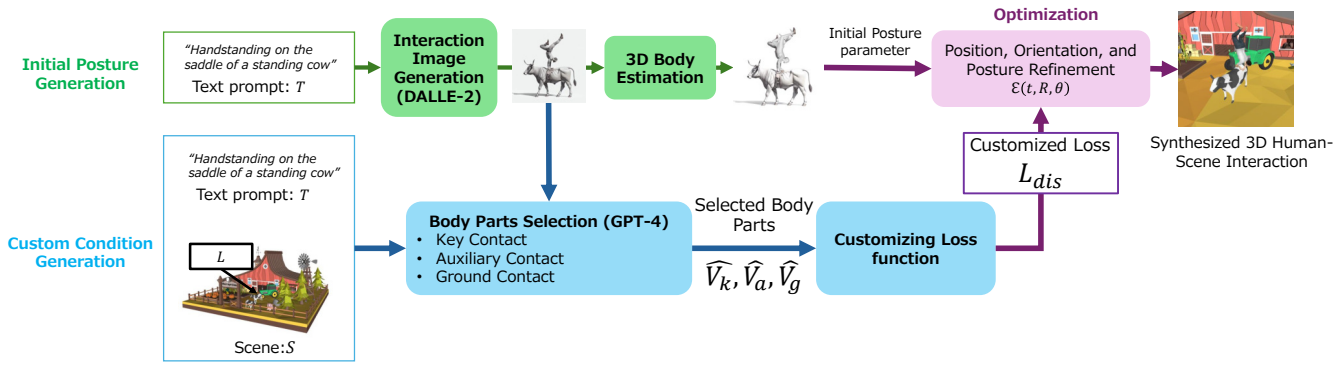

Fig. 2. Outline of our method: In initial posture generation step, an interaction image is generated using DALLE-2, and 3D posture parameters are estimated. In custom condition generation step, GPT-4 analyzes the image, text prompt, and scene to identify contact body parts for customizing loss functions. Finally, a 3D human is synthesized by optimizing these customized loss functions to ensure realistic 3D human interactions.

ing question: *"Generally speaking, when a person is {PROMPT}, which body part should be in contact with the {AREA'S NAME}? Please select just one body part from the following options: <list of body parts>."*, where the list is ['rightHand', 'rightUpLeg', 'leftArm', 'head', 'leftLeg', 'leftFoot', 'rightFoot', 'rightArm', 'rightLeg', 'leftForeArm', 'rightForeArm', 'neck', 'leftUpLeg', 'leftHand', 'hips']. The answer is the name of one body part.

The vertices of the selected body part are denoted as $\hat{V}_k$. Using $\hat{V}_k$, we calculate the distance loss function d_{key}. The d_{key} is defined as:

$$d_{\text{key}} = \begin{cases} \sum_{\hat{v}_k \in \hat{V}_k} \min_{v_k \in V_k} \|\hat{v}_k - v_k\|_2 & \text{if } A_k = \text{"yes"} \\ \sum_{\hat{v}_k \in \hat{V}_k} \min_{v_s \in V_s} \|\hat{v}_k - v_s\|_2 & \text{otherwise} \end{cases} \tag{1}$$

where $\hat{v}_k$ denotes the location of a vertex in $\hat{V}_k$. If the answer to Assessing Specific Area Contact Need, A_k, is "No," indicating that the body does not need to be in contact with a specific area, the distance between $\hat{V}_k$ and the entire scene is then calculated.

The second and third steps of this process are illustrated in Fig. 3. Here is an example of how these processes are conducted. When the text prompt is "sitting on the saddle of a cow", the human should sit on a specific location of the cow, specifically the cow's back, not the head or neck. Therefore the answer to Assessing Specific Location Contact Need, A_k becomes "yes". Then based on multi-view rendered images, the interaction area is identified as "cow's back", and "hips" is selected as the body part that should contact the "cow's back".

These steps ensure consistency between scenes and synthesized bodies. Scene information is obtained by rendering images around the Key Scene Area. Key Contact Parts are then selected based on this scene information to ensure semantic consistency. We have conducted experiments to assess the impact of these restrictions, and the results are provided in the supplementary materials.

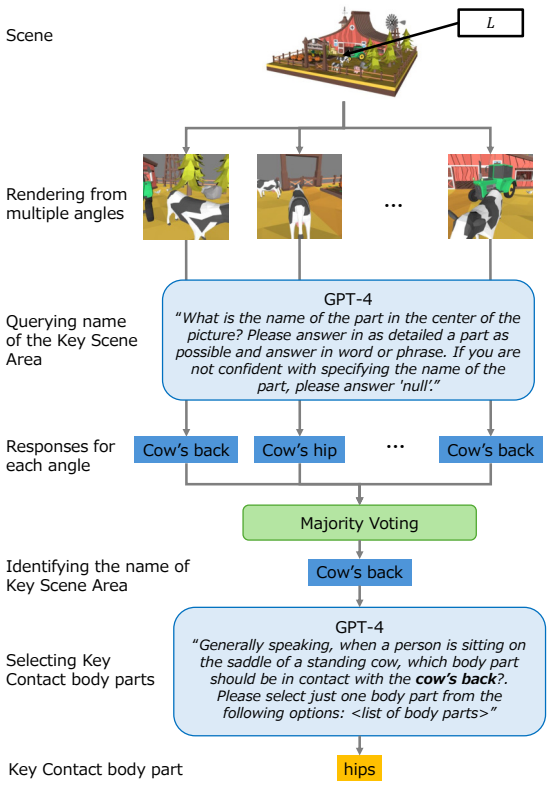

Fig. 3. A part of Key Contact Body Parts Selection (identifying the interaction area and selecting Key Contact Body Parts).Multiple images centered on the interaction location are rendered, and the name of the location is identified using GPT-4 and majority voting. This identified name is then passed to GPT-4 to select Key Contact Body Parts.

2) Auxiliary Contact Body Parts Selection: We identify the body parts that should contact the scene by querying GPT-4. Specifying a single pair of body and scene parts often does not suffice to define natural interactions. For example, relying solely on the distance between the hip (Key Contact Body Parts) and a cow's back when sitting on a cow's saddle can result in unrealistic interactions due to the hip's large segmentation area. Fig. 4 illustrates this failure case. To

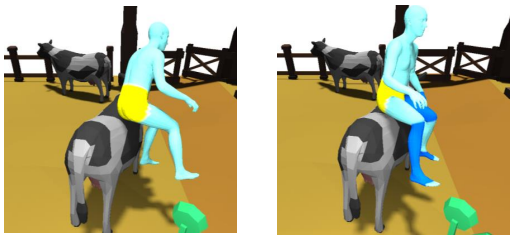

Fig. 4. Examples showing the effect of Auxiliary Contact. The yellow and blue areas represent the segmentation of the Key Contact and Auxiliary Contact Body Parts, respectively, with the hip selected as the Key Contact body part. In the left image, without Auxiliary Contact, the body is angled unnaturally. In the right image, the human model is not angled due to the use of Auxiliary Contact Body Parts, including the left upper leg, right leg, and both feet.

address this, we select Auxiliary Contact Body Parts that should contact the scene, ensuring more natural positions and orientations by regulating the distances between relevant parts and the scene. The query is: *"In the given image, a person is {PROMPT}. Which body parts should be in contact with the OBJECT? Please select from the following parts: <list of body parts>."*. The answer is a list of body parts. Based on the responses, in order to obtain more precise part, we determine the specific areas (front or back) of these body parts that should be in contact with the object by querying GPT-4. The vertices of the selected body parts are denoted as $\hat{V}_a = (\hat{V}_a^0, ..., \hat{V}_a^m)$ where $\hat{V}_a^m$ is vertices of mth selected parts. The distance loss function $d_{\text{auxiliary}}$ is calculated as:

$$d_{\text{auxiliary}} = \frac{1}{|\hat{V}_a|} \sum_{\hat{V}_a^m \in \hat{V}_a} \sum_{\hat{v}_a \in \hat{V}_a^m} \min_{v_s \in V_s} \|\hat{v}_a - v_s\|_2 \quad (2)$$

where $\hat{v}_a$ denotes the location of a vertex in $\hat{V}_a^m$, V_g denotes the vertecis of the ground within the scene S.

3) Ground Contact Body Parts Selection: To avoid unrealistic scenarios such as floating models, we identify Ground Contact Body Parts that should be in contact with the ground. This is achieved by querying the GPT-4 to determine if the person is in contact with the ground and, if so, which parts are involved. This step ensures the 3D human model maintains physical plausibility within the scene. The queries are: *"Given the person in the input image, do you think any of this person's body parts are touching the floor? Please just answer Yes/No."* and *"When the person in the given image {PROMPT}, which body parts should be contacting with the floor? Please select from the following parts: <list of body parts>."* The answer to the first query is denoted as A_g. The vertices of the selected body parts are denoted as $\hat{V}_g = (\hat{V}_g^0, ..., \hat{V}_g^m)$ where $\hat{V}_g^m$ is vertices of mth selected parts. The distance loss function between these selected parts and the ground within the scene is calculated as:

$$d_{\text{ground}} = \begin{cases} \frac{1}{|\hat{V}g|} \sum_{\hat{V}_g^m \in \hat{V}_g} \sum_{\hat{v}_g \in \hat{V}_g^m} \min_{v_g \in V_g} \|\hat{v}_g - v_g\|_2 & \text{if } A_g = \text{"yes"} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where V_g denotes the vertices of the ground within the scene S and $\hat{v}_g$ denotes a vertex in $\hat{V}_g^m$.

4) Customizing Loss Functions: We customize the loss functions based on interaction requirements to ensure realistic and contextually accurate human-scene interactions. Using d_{key},$d_{\text{auxiliary}}$, and d_{ground}, the composite Distance Loss (L_{dis}) is defined as $\mathcal{L}_{\text{dis}} = \alpha_{\text{key}} \cdot d_{\text{key}} + \alpha_{\text{auxiliary}} \cdot d_{\text{auxiliary}} + \alpha_{\text{ground}} \cdot d_{\text{ground}}$ where α_{key}, $\alpha_{\text{auxiliary}}$, and α_{ground} represent the weights assigned to each respective distance.

Additionally, we define a Penetration Loss, $\mathcal{L}_{\text{pen}} = \sum_{\hat{v} \in \hat{V}} \min(\Psi(\hat{v}), 0)$, to prevent unrealistic interpenetration of the human model with the environment, with $\Psi(\hat{v})$ denoting the signed distance of body vertex $\hat{v}$ to the scene. When $\Psi(\hat{v})$ has a negative sign, it indicates that the body vertex $\hat{v}$ is located inside the nearest scene object, signifying penetration. For computational efficiency, we use a precomputed SDF grid for each scene. Finally, the Pose Regularization Loss, $\mathcal{L}_{\text{reg}} = \|\theta - \theta_{\text{init}}\|_2$, penalizes SMPL-X pose parameter θ deviating from their initialization. The optimization objective is defined as:

$$\mathcal{E}(t, R, \theta) = \mathcal{L}_{\text{dis}} + w_{\text{reg}}\mathcal{L}_{\text{reg}} + w_{\text{pen}}\mathcal{L}_{\text{pen}} \quad (4)$$

where the weights w_{reg} and w_{pen} balance the contributions of regularization and penetration losses, respectively.

D. Synthesizing through Optimization

The synthesis process begins by setting the initial body location and orientation parameters. The initial location is randomly chosen from around the given interaction location, and the initial orientation is selected from various angles around the vertical axis. Through iterative optimization, we adjust (t, R, θ) to minimize the loss functions. This process is repeated N times, and the parameters that achieve the minimum loss are selected as the final output, ensuring the generation of realistic and diverse human-scene interactions.

IV. Experiment

To demonstrate the efficacy of our method for generating diverse human-scene interactions, including uncommon scenarios, we conduct both quantitative and qualitative evaluations, following the methodologies in previous works [41], [18]. We compared our approach with a baseline method for this novel task.

A. Dataset

To evaluate a variety of interaction types, including both common and uncommon ones, we constructed a dataset based on GenZI's Sketchfab dataset [18]. Our dataset includes not only frequent interactions (e.g., "pulling dumbbells, standing") similar to those in GenZI's Sketchfab

dataset but also infrequent interactions (e.g., "kicking a door handle"). This enables us to evaluate a diverse range of interactions with one dataset, setting it apart from existing datasets, like GenZI's dataset and PROX [12], [41], which mainly focus on common interactions.

Due to copyright constraints, we selected 6 out of the original 8 scenes from GenZI's Sketchfab dataset. We generated 3-5 text prompts per scene, describing human interactions at specified approximate point locations, with the aid of GPT-4 to suggest both common and uncommon interactions involving specific objects. In total, 30 prompts for 26 point locations were utilized in the experiment. Fig. 6 presents some examples of the prompts and locations.

B. Baseline

In this study, we use GenZI as a comparative method. To the best of our knowledge, GenZI is the only baseline method that estimates 3D human-scene interactions based on natural language input in a zero-shot manner.

While GenZI used COINS [41] and Hassan et al. [13] as their baselines, we do not use them as our baseline in this study due to their limited variations of actions and scenes. COINS, a state-of-the-art method for estimating 3D humans in indoor scans, relies on a fixed vocabulary of actions and objects, using ⟨action, object⟩ pairs for semantic control, and requires full supervision with captured 3D interaction data for CVAE training. However, its applicability is limited by its training on a dataset with a small vocabulary, making it unsuitable for our dataset. Hassan et al. perform 3D human estimation from a single RGB image and require the presence of humans in the image, which is incompatible with our method, as it does not generate or require such images.

C. Evaluation Metrics

We evaluate our method by conducting perceptual studies and assessing interaction performance, including the interactions' physical plausibility, semantic consistency, and diversity.

1) Perceptual Studies: Our perceptual studies follow the methodology employed by GenZI and consist of two measurements: Binary-Choice and Unary.

Binary-Choice Study: In the Binary-Choice measurement, participants compare images generated by two different methods based on the same text prompt and select the image that best matches the prompt.

Unary Study: For the Unary study, participants rate how well the displayed image matches the text prompt used to generate it, using a 5-level Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

To prevent participants from being influenced by seeing the same interaction more than once, we divided 30 prompts randomly into two groups of 15 each. Participants were then instructed to respond to each study accordingly.

2) Assessing Interaction Performance: To assess the quality of human-scene interaction, we employed several metrics including semantic consistency, physical plausibility, and diversity, following the evaluation methodology used in previous works [41], [18].

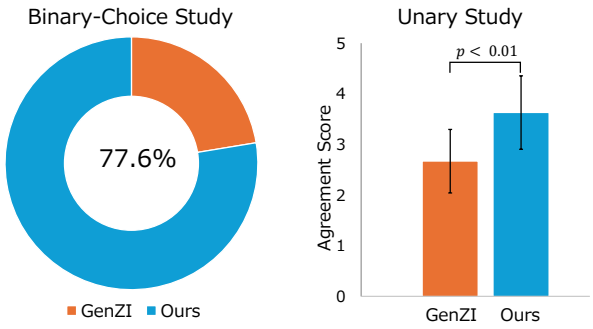

Fig. 5. Results of perceptual studies (Binary-Choice and Unary study). Participants evaluated synthesized human-scene interactions using our dataset. The results indicate that interactions generated by our method are preferred over those produced by GenZI.

Semantics: For semantic consistency, we use the CLIP-score (**CLIP**) as proposed by GenZI. This score is calculated by re-rendering the 3D interaction from multiple view angles and averaging the image-text cosine similarities from CLIP ViT-B/32 across all views. The view angles are determined following the method used by GenZI.

Physical Plausibility: To assess physical plausibility, we evaluate the non-collision score (**Collision**) and the contact score (**Contact**). The non-collision score is calculated as the ratio of the number of body vertices with non-negative scene SDF values to the total number of body vertices. The contact score is the ratio of the number of body meshes in contact with the scene to the total number of generated body meshes. A body mesh is considered in contact with the scene if any of its vertices have a non-positive SDF value.

Diversity: To measure diversity, K-Means clustering is applied to the SMPL-X parameters of the generated bodies, clustering them into 20 groups. We then evaluate diversity using the entropy of the cluster ID histogram (**Entropy**) and the mean distance to cluster centers (**Cluster Size**).

D. Implementation Details

Our method is implemented using PyTorch and executed on an Nvidia A100 GPU. The optimization process involves a learning rate of 0.001, with a total of 200 optimization steps. Additionally, the generation process is repeated $N = 30$ times to ensure robustness and consistency in the results. k, nearest points from the given interaction point location, is set to 50. The parameters $\alpha_{key}, \alpha_{scene}$, and α_{ground} are set to 2, 1, and 1.5, respectively. Additionally, the weights w_{reg} and w_{pen} are set to 50 and 10, respectively.

E. Comparison with Baseline

1) Quantitative Evaluation: In this evaluation, we conducted perceptual studies and evaluated interaction performance, following previous works [41], [18]. It is crucial to highlight, as discussed in [18], that results from perceptual studies are critical for evaluation because the metrics of interaction performance often fail to align with human perceptual

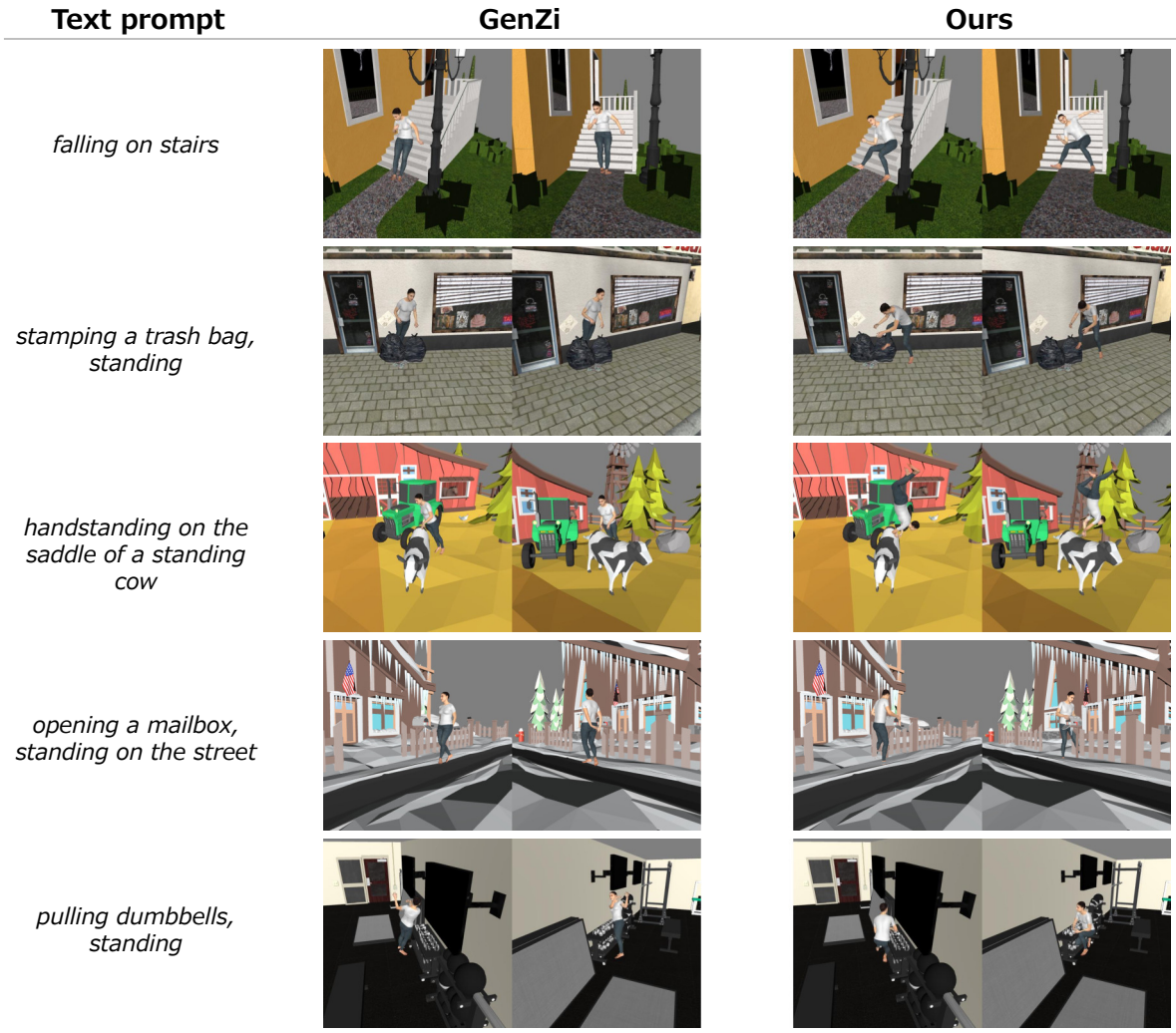

Fig. 6. Comparison of visualization results: GenZI vs. Ours. The synthesized interactions are captured from different angles. The interactions produced by our method demonstrate more plausible interactions compared to those generated by GenZI.

evaluations. For instance, despite achieving a Collision and Contact score of 1.0, this metric merely indicates any physical contact between the body and scene elements, irrespective of its relevance to the intended interaction. Similarly, the CLIP-score can be elevated when both human and object are present in the picture, without necessarily reflecting the interaction's realism.

Results of Perceptual Studies: 22 people answered Binary-Choice study and 18 people answered Unary study. We should note that some of the participants did not answer the Unary study. Fig. 5 shows the results of perceptual studies. As shown in Fig. 5, participants indicated that the interactions generated by our method matched the input text prompts better 77.6% of the time. In the Unary study, our method achieved an average score of 3.63, which is significantly higher ($p < 0.01$, Student's t-test) than GenZi's average score of 2.67. These results suggest that our method can generate more natural and realistic interactions, even in unusual scenarios, as perceived by human observers.

Results of Interaction Performance: As shown in Table I, our method achieves higher scores in CLIP (0.2597) and Collision (0.9890) compared to GenZI's scores of 0.2546 in CLIP and 0.9792 in Collision, and a perfect score of 1.0000 in Contact. These results suggest that our approach

TABLE I

COMPARISON OF QUANTITATIVE RESULTS FOR SEMANTIC CONSISTENCY, PHYSICAL PLAUSIBILITY, AND DIVERSITY ON OUR DATASET. OUR FULL VERSION ACHIEVES THE HIGHEST SCORES IN CLIP AND CONTACT METRICS, WHILE THE VERSION WITHOUT KEY CONTACT BODY PARTS SELECTION (OURS W/O KEY CONTACT) SCORES THE HIGHEST IN COLLISION.

Method	Semantics	Physical Plausibility		Diversity	
	CLIP	Collision	Contact	Entropy	Cluster Size
GenZI	0.2546	0.9792	**1.0000**	**2.7388**	**1.0502**
Ours w/o Key Contact	0.2574	**0.9932**	0.7666	2.6821	0.9475
Ours w/o Auxiliary Contact	0.2590	0.9905	0.8333	2.7192	0.9414
Ours w/o Ground Contact	0.2590	0.9893	0.8333	2.6350	0.9665
Ours	**0.2597**	0.9890	**1.000**	2.6844	0.9787

generates high-quality human-scene interaction synthesis, as CLIP score is considered a more reliable assessment of human-scene interaction quality compared to other metrics, as indicated in [18].

In terms of diversity scores, our method scores lower than GenZi. This is due to GenZi occasionally producing significantly different and unusual postures when prompted with instructions for uncommon and infrequent interactions. Despite this, our method's performance in producing realistic and plausible interactions, as evidenced by the high CLIP and Collision scores, demonstrates its effectiveness in generating high-quality human-scene interactions.

2) Qualitative Evaluation: Fig. 6 presents qualitative comparisons between GenZI and our method. GenZI struggles with unusual scenarios like "stamping a trash bag, standing" (second row) and "handstanding on the saddle of a standing cow" (third row), where the postures do not align with the text prompts. In contrast, our method successfully synthesizes plausible human-scene interactions with natural postures that accurately reflect the prompts in these challenging situations.

F. Ablation Studies

We conducted several ablation studies to validate the effectiveness of our approach for customizing loss functions, as described in Section 3.3.

1) The effect of loss functions: To assess the effectiveness of Key Contact, Auxiliary Contact, and Ground Contact Body Parts, we compared the performance of our method with variants where d_{key}, $d_{\text{auxiliary}}$, d_{ground} were individually omitted. The results for physical plausibility, diversity, and semantic scores are presented in Table I. Our full version archives higher scores in CLIP, Contact, and Cluster Size compared with other variants. Ours w/o Key Contact marks a higher Collision than our full version, however, the Contact score (0.76666) is significantly lower than that of our method (1.0000). Additionally, without Key Contact, the model struggles to find a reasonable location and orientation. Consequently, the gap in CLIP score between our method with and without Key Contact is larger than in the other ablation study conditions.

2) The effect of Initial Posture Generation: In our method, Initial Posture Generation, as described in Sec. III-B, plays a crucial role in determining natural postures and synthesizing humans within scenes. To assess its importance, we conducted an experiment by omitting Initial Posture Generation, setting all "body_pose" parameters of SMPL-X to zero (T-pose), and directly proceeding to the Optimization phase (Sec. III-D).

We assessed the outcomes through additional perceptual studies and interaction performance evaluations. The perceptual studies, using Binary-Choice and Unary methods as described in Sec.IV-C.1, involved 12 and 10 participants respectively. The results, displayed in Fig.7, reveal that 86.7% of interactions generated with Initial Posture Generation aligned more naturally compared to those using T-pose initialization in the Binary-Choice study. In the Unary study, the average score for Initial Posture Generation (3.86) was significantly higher than that for T-pose initialization (2.13), with $p < 0.01$ according to the Student's t-test.

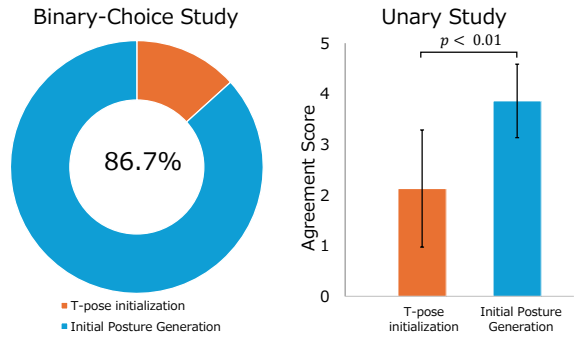

Fig. 7. Results from the experiment on the effect of Initial Posture Generation. Perceptual studies, including Binary-Choice and Unary, show that interactions generated with Initial Posture Generation are preferred over those produced with T-pose initialization.

TABLE II
COMPARISON OF QUANTITATIVE RESULTS WITH T-POSE INITIALIZATION AND INITIAL POSTURE GENERATION

Method	Semantics	Physical Plausibility		Diversity	
	CLIP	Collision	Contact	Entropy	Cluster Size
T-pose initialization	0.2571	0.9756	0.9766	**2.7368**	0.9534
Initial Posture Generation (Proposed)	**0.2597**	**0.9890**	**1.000**	2.6844	**0.9787**

Interaction performance was evaluated using the metrics outlined in Sec. IV-C.2, with results displayed in Table II. The data reveal that interactions with Initial Posture Generation scored higher across all evaluation indices, except for Entropy. In addition, for visualization results, we kindly refer the reader to the supplementary material.

These findings underscore that Initial Posture Generation is vital for creating natural and realistic human-scene interactions.

V. CONCLUSION

In this paper, we introduce a novel zero-shot method for generating both common and uncommon human-scene interactions. This approach separates posture specification from the scene and defines interaction conditions individually. Initially, we generate interaction images using T2I based on text prompts, allowing for a diverse range of postures without training data bias. To specify positions and orientations, we leverage VLMs and LLMs to customize conditions for each interaction. By defining and optimizing loss functions based on the conditions, our method ensures natural and contextually appropriate interactions. Validation on a new dataset confirms the efficacy of our approach. Future work will involve expanding the types of conditions for customizing loss functions.

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] J. P. Araújo, J. Li, K. Vetrivel, R. Agarwal, J. Wu, D. Gopinath, A. W. Clegg, and K. Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023.

[5] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9995, 2023.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Z. Cen, H. Pi, S. Peng, Z. Shen, M. Yang, S. Zhu, H. Bao, and X. 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.

[8] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[9] G. Delmas, P. Weinzaepfel, T. Lucas, F. Moreno-Noguer, and G. Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022.

[10] G. Delmas, P. Weinzaepfel, T. Lucas, F. Moreno-Noguer, and G. Rogez. Posescript: Linking 3d human poses and natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] G. Delmas, P. Weinzaepfel, F. Moreno-Noguer, and G. Rogez. Posefix: correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023.

[12] M. Hassan, V. Choutas, D. Tzionas, and M. 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.

[13] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021.

[14] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. 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.

[15] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. 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.

[16] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. Sasha Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[17] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[18] L. Li and A. Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[21] A. Monszpart, P. Guerrero, D. Ceylan, E. Yumer, and N. J. Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019.

[22] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[23] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019.

[24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[25] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[28] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions On Graphics (TOG)*, 35(4):1–12, 2016.

[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[30] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[31] X. Wang, G. Li, Y.-L. Kuo, M. Kocabas, E. Aksan, and O. Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *2022 International Conference on 3D Vision (3DV)*, pages 353–362. IEEE, 2022.

[32] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

[33] Z. Wang, Y. Chen, T. Liu, et al. HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes. In *NeurIPS*, 2022.

[34] Z. Xiao, T. Wang, J. Wang, J. Cao, W. Zhang, B. Dai, D. Lin, and J. Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023.

[35] H. Xuan, X. Li, J. Zhang, H. Zhang, Y. Liu, and K. Li. Narrator: Towards natural control of human-scene interaction generation via relationship reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22268–22278, 2023.

[36] Y. Ye, X. Li, A. Gupta, S. De Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023.

[37] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu. Pymafx: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[38] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang. Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651. IEEE, 2020.

[39] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022.

[40] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. 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.

[41] K. Zhao, S. Wang, Y. Zhang, T. Beeler, , and S. Tang. Compositional human-scene interaction synthesis with semantic control. In *European conference on computer vision (ECCV)*, 2022.

[42] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[43] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4:

Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[44] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.