Narrator: Towards Natural Control of Human-Scene Interaction Generation via Relationship Reasoning

Haibiao Xuan[1] Xiongzheng Li[1] Jinsong Zhang[1] Hongwen Zhang[2] Yebin Liu[2] Kun Li[1,*]
[1]Tianjin University [2] Tsinghua University
{hbxuan, lxz, jinszhang, lik}@tju.edu.cn {zhanghongwen, liuyebin}@mail.tsinghua.edu.cn
http://cic.tju.edu.cn/faculty/likun/projects/Narrator

Figure 1: Given a textual description, our approach can naturally and controllably generate semantically consistent and physically plausible human-scene interactions for various cases: (a) interactions guided by spatial relationships, (b) interactions guided by multiple actions, (c) multi-human scene interactions, and (d) human-scene interactions combining the above interaction types, which cannot be generated using prior works.

Abstract

Naturally controllable human-scene interaction (HSI) generation has an important role in various fields, such as VR/AR content creation and human-centered AI. However, existing methods are unnatural and unintuitive in their controllability, which heavily limits their application in practice. Therefore, we focus on a challenging task of naturally and controllably generating realistic and diverse HSIs from textual descriptions. From human cognition, the ideal generative model should correctly reason about spatial relationships and interactive actions. To that end, we propose ***Narrator****, a novel relationship reasoning-based generative approach using a conditional variation autoencoder for naturally controllable generation given a 3D scene and a textual description. Also, we model global and local spatial relationships in a 3D scene and a textual description respectively based on the scene graph, and introduce a part-level action mechanism to represent interactions as atomic body part states. In particular, benefiting from our relationship reasoning, we further propose a simple yet effective multi-human generation strategy, which is the first exploration for controllable multi-human scene interaction generation. Our extensive experiments and perceptual studies show that Narrator can controllably generate diverse interactions and significantly outperform existing works.*

*Corresponding author

1. Introduction

Throughout daily life, humans constantly interact with their surroundings and these interactions establish their relationships with the scenes. Naturally controllable human-scene interaction (HSI) generation has significant value and numerous applications in areas such as VR/AR content creation, human-centered AI, and generating training data for other computer vision tasks. In this paper, we tackle a challenging task of generating realistic and plausible human-scene interactions from natural language textual descriptions, particularly exploring more liberal forms of HSIs with complex spatial relationships, multiple actions, and multiple persons, as shown in Fig. 1.

Prior HSI methods [36, 13, 31] mostly focus on the physical geometry between humans and scenes, but lacks the semantic control of generation. Some works [30] further incorporate generative controls, but always coarsely describe them as action labels, not sentences. A recent method, COINS [37], specialises semantic control of interactions as combinations of actions and objects. However, additional manual effort is required to explicitly specify object instances when faced with multiple objects of the same kind. Moreover, binding actions to objects by force is not intuitive or reasonable. For example, a natural case, "standing by the window", does not contain a direct and explicit interaction object, and COINS cannot deal with it. These unnatural and constrained control ways fall short of meeting the needs of users and limit their applicability.

Humans usually naturally describe people who have diverse interactions in different places through spatial perception and action recognition. Thus, an ideal generative model should correctly reason about *spatial relationships* to obtain the human position that respects textual descriptions while exploring degrees of freedom about *interactive actions* to generate natural interactions. Specifically, *spatial relationships* can be represented as the interrelationship among different objects in a scene or a local area, and *interactive actions* are specified by atomic body part states, such as a person's left feet treading, torso leaning, right hand tapping, and head bowing. How to reason about these relationships and utilize these powerful cues for naturally controllable generation is a pressing problem.

To address these issues, we propose *Narrator*, a novel generative approach that incorporates a transformer-based conditional variational auto-encoder (cVAE) framework and leverages relationship reasoning to naturally produce diverse and plausible HSIs given the scene and textual description. The diversity and complex interrelationship of objects in scenes can lead to misjudgements of human position and unnatural interactions. Therefore, instead of understanding scenes or specific objects in isolation as previous works, we employ the scene graph to represent spatial relationships and propose a Joint Global and Local Scene Graph (JGLSG) mechanism to provide global perception for subsequent localization, allowing for interaction generations guided by spatial relationships (in Fig. 1 (a)). As body part states are key for modeling realistic and text-faithful interactions, we introduce a Part-Level Action (PLA) mechanism to establish the correspondence between human body parts and actions, allowing for interaction generations guided by multiple actions (in Fig. 1 (b)). Ultimately, we feed the multi-modal features extracted by JGLSG, PLA and PointNet++ [24] as a joint conditional embedding into cVAE, thus obtaining a unified latent space of the human body. To train and evaluate our approach, we annotate multi-level text descriptions from coarse to fine for each frame of the PROX dataset [12].

In real-world scenes, there are more situations where multiple people are interacting independently or in a connected way. Unfortunately, there is no work that solves this problem in an automatic and controlled way, but rather requires certain expertise and manual effort [18, 36]. Also, a straightforward way by using a single-person method like COINS [37], *i.e.*, sequential per-person generation and optimization, does not properly understand multi-person text descriptions, leading to unreasonable spatial distributions and unnatural interactions of the generated results. In contrast, benefiting from the flexibility and reusability of our JGLSG and PLA mechanisms, we propose a simple yet effective multi-human generation strategy. We reason out each person's interaction information from the text and globally update each generation to establish their relationships, thus achieving a better spatial distribution than simple multiple generation. To our knowledge, this is the first naturally controllable and user-friendly generative model for multi-human scene interaction (MHSI) (in Fig. 1 (c)).

In brief, our contributions can be summarized as:

- we present *Narrator*, a new generative method for naturally controllable human scene interaction generation given textual descriptions in natural language.
- we propose the JGLSG and PLA mechanisms for relationship reasoning considering narrator's perspective.
- we propose the first naturally controllable MHSI generation strategy to approximate the real world.

2. Related work

Human-Scene Interaction (HSI). Human-scene interaction, a challenging task in computer vision, recently has received increasing attention. Early HSI methods [16, 26, 28, 18, 7, 20] focused on scene affordance and function understanding, but the lack of relevant high-quality datasets and valid human representations lead to low-fidelity interaction and poor results. PiGraphs [26] learns the probability distribution of each verb-object category from real-world interactions to generate interaction snapshots, but the simple representation of the 3D human as a skeleton pre-

vents reasoning about contact details. Aided by the parametric body model SMPL-X [21] and the dataset PROX [12] recording human activities in 3D scenes, efforts in recent years continually iterate and refine towards realism and naturalism. Zhang *et al.* [36] trains a conditional variation autoencoder (cVAE) to predict semantically plausible 3D human poses with scene depth and semantics, and apply geometric constraints for physical plausibility. POSA [13] proposes a human-centric contact map that encodes contact probability and semantic information for each vertex on the body mesh, and uses these to guide the search for its most likely position in the scene. Although these methods [36, 35, 13, 15] model interactions with different representations, they cannot support controllable interaction synthesis. Given the action sequence, Wang *et al.* [30] propose a three-stage framework to place humans into scenes, produce feasible paths, and complete motion synthesis. COINS [37] represents semantic control as combinations of actions and objects, similar to PiGraph [26], and combine atomic interactions into compositional interactions.

COINS is the SOTA HSI method, and the most relevant work, but our work has the superiority in many aspects: 1) The control way of COINS is not intuitive and requires additional manual selections for object instances, while our approach is fully automatic conditioned on natural language description; 2) our approach has more flexible spatial localization ability and can handle non-direct interactions (*e.g.*, standing by the window), while COINS fails; 3) COINS has limited interaction types and combinations, whereas ours is more diverse and we can simultaneously support more constraints (*e.g.*, left/right hand lift, bend and crouch); 4) We are the first to explore and achieve controllable MHSI generations via our relationship reasoning.

Text-guided Action & Object Grounding. Grounding human actions and objects in scenes from textual descriptions are important and meaningful tasks that have received much exploration. For text-guided action grounding, recent works explore advances in natural language with many amazing results [2, 3, 34, 27, 14, 10, 22, 4, 11]. CLIP-Actor [34] utilizes multi-modal perception and semantic matching to synthesize the best matching action sequences from a text-visual coupling perspective. Guo *et al.* [11] propose a two-stage pipeline to implement the prediction from input text to visual action length and then to motion generation. On the other hand, 3D object grounding aims to locate the most relevant target object in 3D point cloud scenes conditioned on textual descriptions [6, 33, 9, 25, 19].

Different from the above-mentioned, our approach takes more account of possible human interactions in the scene and refines these into the body part states. Meanwhile, for better localization and grounding, we unite position features encoded from textual descriptions and 3D scenes into the conditional embedding of the cVAE.

3. Overview

Our goal is to naturally generate human-scene interactions that are semantically consistent with textual descriptions and physically plausible with scenes. Fig. 2 shows the framework of our approach. To this end, we propose a novel generative approach, *Narrator*, with a transformer-based conditional Variational Auto-Encoder (cVAE) network architecture (Sec. 4.1). Specifically, in contrast to existing works that consider scenes or objects in isolation, we design a Joint Global and Local Scene Graph (JGLSG) mechanism to reason about complex spatial relationships for global localization perception (Sec. 4.2). In addition, people simultaneously engage in diverse interaction activities with different body parts. This inspired us to introduce a Part-Level Action (PLA) mechanism for realistic and diverse interactions (Sec. 4.3). Meanwhile, we introduce an Interaction Bisector Surface (IBS) loss to obtain better generation results during scene-aware optimization (Sec. 5). We further broaden into multi-human fields and ultimately facilitate the first step to MHSI (Sec. 6).

Here we give the representation of the scene, the body mesh and the textual description. We denote the scene as $S = (V_s, S_s)$, where V_s and S_s stand for vertices and per-vertex semantic labels, respectively. We represent the 3D human body mesh using a SMPL-X model [21] and a POSA representation [13]. Specifically, the SMPL-X body mesh $M_{\text{SMPL-X}} = (V_b, F_b)$ with vertices $V_b \in \mathbb{R}^{10475\times 3}$ and triangles F_b, is parameterized by a differentiable function $F(t, r, \beta, p, h)$, where $t \in \mathbb{R}^3$ is the global translation, $r \in \mathbb{R}^6$ is a continuous representation [39] of the global orientation, $\beta \in \mathbb{R}^{10}$ is the body shape parameters, $p \in \mathbb{R}^{63}$ is the body pose parameters, and $h \in \mathbb{R}^{24}$ is the hand pose parameters. We also extract contact labels L_b using POSA. Overall, we define the body mesh as $M = (V_b, F_b, L_b)$. Besides, the textual description about HSI includes various levels of interaction detail, defined as a sequence of words $W_{1:N} = [w_1, \ldots, w_N]$ from the English vocabulary.

4. Method

4.1. Network Architecture

For naturally controllable HSI generation, we employ a transformer-based cVAE architecture that can handle multi-modal information including scenes and textual descriptions, and model the probability $p(M \mid S, W_{1:N})$. We describe the details of each part as follows.

Condition Module. The condition module takes a 3D scene and a textual description as input, and outputs a joint conditional embedding. First, we employ PointNet++ [24] to extract the scene S into 256-dimension scene features f_s. Then, to reason the spatial and structural relationships, the scene and textual description are simultaneously input to the JGLSG to obtain the scene graph including human nodes, and the scene graph feature f_{sg} is obtained by encoding it

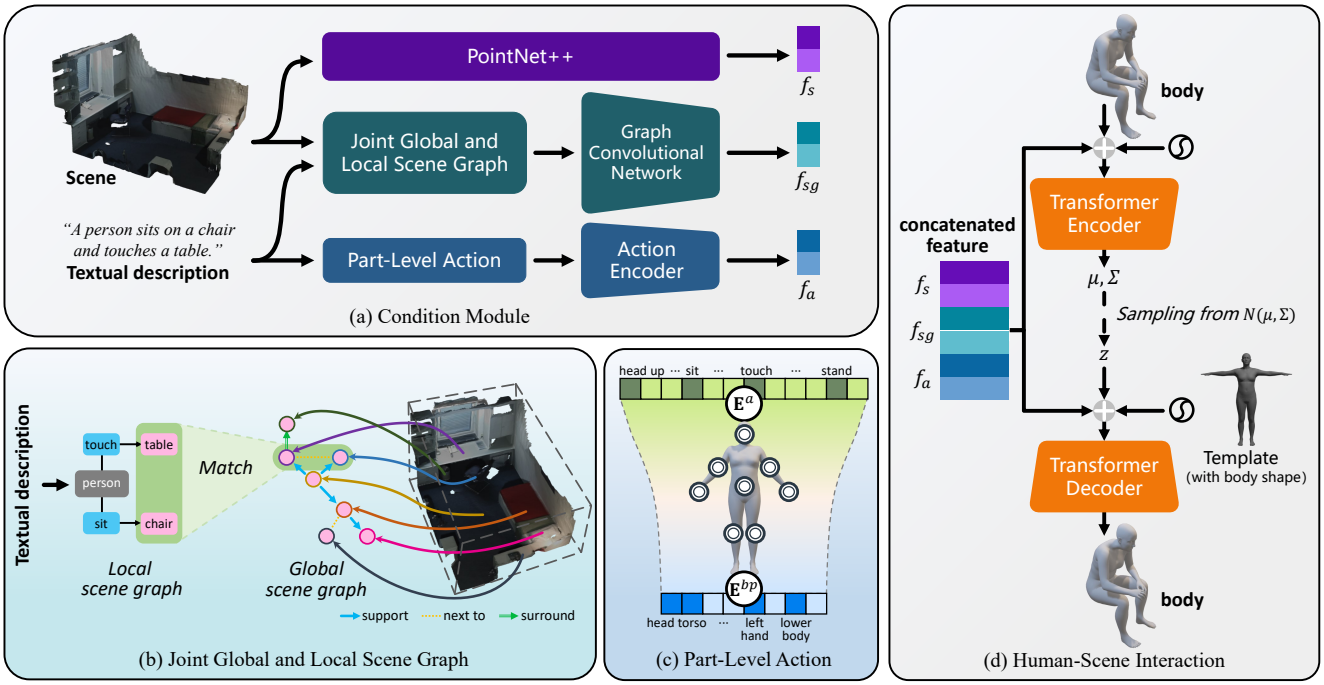


Figure 2: Overview of the proposed *Narrator* framework. Given a scene and a textual description, multi-modal features including scene features, scene graph features, and action features are extracted (a), where the latter two are reasoned through our Joint Global and Local Scene Graph (b) and Part-Level Action (c), respectively. These features are then concatenated as a joint conditional embedding and fed into the transformer-based cVAE framework for human-scene interaction (d).

using a Graph Convolutional Network [17]. In addition, to reason about the human-action relationship, the action combination parsed from the textual description is fed into the PLA for mapping to atomic body part states, and then the Action Encoder encodes them as the action feature f_a. Finally, these three features are concatenated as a joint conditional embedding f_{ce}.

Transformer Encoder. We first utilize a fully-connected layer (FC) to encode human body mesh M as a high-level embedding and concatenate it with joint conditional embedding f_{ce} as input. On top of the encoder, we apply average pooling to the output, followed by another FC to predict the Gaussian distribution $\mathcal{Q}(z \mid S, W_{1:N}, M)$. Finally, we sample the latent code z from the distribution using the reparameterization trick, as one of the decoder inputs.

Transformer Decoder. In the decoder, we use a SMPL-X template body with person-dependent shape parameters as the body token to improve the generalization of our model and achieve finer-grained control. We concatenate the latent code z with the joint conditional embedding f_{ce} as another input for the decoder. The output body mesh is fed into the SMPL-X regressor [37] to regress with consistent SMPL-X body parameters for loss supervision.

4.2. Joint Global and Local Scene Graph

Reasoning about spatial relationships can provide scene-specific clues to the model, which plays an important role in achieving natural controllability for HSI. Therefore, we design a Joint Global and Local Scene Graph mechanism, which is implemented through the following three steps.

Global Scene Graph Generation: Given the scene, we use a model [40] pre-trained on 3DSGG [29] to generate a global scene graph, *i.e.*, $\mathcal{GSG} = (\mathcal{V}, \mathcal{E}), \mathcal{V} = \{o_i\}_{i=1}^{n}, \mathcal{E} = \{(o_i, r_{ij}, o_j)_k\}_{k=1}^{m}$, where o_i, o_j are the objects with category labels, r_{ij} is the relationship between o_i and o_j, n is the number of objects, and m is the number of relationships.

Local Scene Graph Generation: Our model adopts an off-the-shelf semantic parsing toolkit [32] to recognise the syntactic structure of the textual description and extract a set of triplets $\{T_{ij}\}$, where $T_{ij} = (s_i, p_{ij}, o_j)$ defines a triplet of subject-predicate-object. The output of the syntactic parser is not sufficient to represent the human spatial location, especially the number of objects in the scene. Hence, we build quantity checker for detecting its quantifier expression, and duplicating object nodes for the local scene graph $\mathcal{LSG}$ (*e.g.*, "three plants" in textual descriptions $\rightarrow$ three "plant" nodes in $\mathcal{LSG}$).

Scene Graph Matching: Then, we correspond the local scene graph to the nodes in the global scene graph based on same object semantic labels. During this process, two object category concepts can be matched if there is an overlap between their synsets, lemmas, or hypernyms (*e.g.*, "armchair" $\rightarrow$ "chair"). According to the corresponding result of each object, we add a virtual human node by extending the edge relationships for providing the generated position, so we obtain a final scene graph $\mathcal{SG}$ that is consistent with both the scene and the textual description.

Part	Interaction
Head	head up, head down, head left, head right
Torso	sit, sit down, lean, lie, lie down
(L/R) Arm	stretch, bend, straight, supported, raise, put
(L/R) Hand	touch, use, hold, support, supported, type, write, open
(L/R) Lower body	stand, stand up, step, step up, step down, step back, walk, run, move, crouch, turn around, raise, put

Table 1: List of body part actions, where (L/R) indicates the left and right of the part.

4.3. Part-Level Action

Human interactions in the scene are composed of atomic body part states, and hence we propose a part-level action mechanism to select the important parts and disregard the irrelevant parts from the given interactions. Specifically, we explore richer interactive actions than existing works [30, 37] and correspond these possible actions to the five main human body parts: head, torso, left/right arm, left/right hand, and left/right lower body, as shown in Tab. 1. Also, we use the one hot vectors E^a and E^{bp} to represent these actions and body parts, respectively. Then we concatenate them based on our proposed correspondence for subsequent encoding.

For interaction generation guided by multiple actions, the attention mechanism of the transformer network is employed to learn the different part states of the SMPL-X body mesh. Given a combination of interactive actions, the attention between its corresponding body part and all other actions for each action, is automatically masked for each action. Taking the example of "a person crouches on the floor using a cabinet", crouching corresponds to the state of the lower body, and hence the attention of other parts tokens will be masked to zero.

5. Scene-aware Optimization

We perform scene-aware optimization with geometric and physics constraints to improve the generation results, following [36, 13, 37]. Throughout the optimization process, it ensures that the generated poses do not deviate while encouraging contact with the scene and constraining the body mesh to avoid interpenetration with the scene surface. Given the scene mesh S and generated SMPL-X parameters, the optimization loss is given by:

$$\mathcal{L}_{opt} = \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{coll}} + \mathcal{L}_{\text{IBS}} + \mathcal{L}_{\text{reg}}, \tag{1}$$

where $\mathcal{L}_{cont}$ encourages body vertices to contact with the scene mesh, $\mathcal{L}_{coll}$ is the signed-distance-based collision term defined in [36], and the $\mathcal{L}_{reg}$ is a regularizer that penalizes SMPL-X parameters deviating from the initialization.

A further addition we make over existing HSI methods is adopting the Interaction Bisector Surface (IBS) [38], which is the set of points equidistant from two sets of points sampled on the scene and the human, respectively. For our task, we modify it as additional loss supervision $\mathcal{L}_{\text{IBS}}$:

$$\mathcal{L}_{\text{IBS}} = \sum_{v^p \in V} d_s^p, \tag{2}$$

where V denotes the set of all points in the IBS point set that satisfies either penetration or corresponds to the body vertices with contact labels, and d_s^p indicates the distance from point v^p to the scene.

For more details regarding the training and optimization, please refer to the Supp.Mat.

6. Multi-Human Scene Interaction

In real-world scenes, many situations are not just one person interacting with the scene, but multiple people interacting independently or in an associated way. However, due to the lack of MHSI datasets, existing methods fail to handle this task in a controlled and automatic manner, but require additional manual effort. To this end, we propose a simple but effective strategy for MHSI, using only existing single human datasets.

Given a textual description about MHSIs, our model first parses it into multiple local scene graphs $\mathcal{LSG}_i$ and human interactive actions $\mathcal{IA}_i$. We define the candidate set as $\mathcal{S}_c = \{(\mathcal{LSG}_i, \mathcal{IA}_i)\}_{i=1}^{l}$, where l is the number of people. For each element of the candidate set $\mathcal{S}_c$, we first feed it into *Narrator* together with the scene $\mathcal{S}$ and the corresponding global scene graph $\mathcal{GSG}$, subsequently performing the optimization process. To handle collisions between humans, we additionally introduce a loss $\mathcal{L}_{\text{HH}}$ during the optimization process as follows:

$$\mathcal{L}_{\text{HH}} = \sum_i \Psi_{\text{H}}(v_i), \tag{3}$$

where $\Psi_{\text{H}}(v_i)$ denotes the signed distance of vertex v_i of the generated body mesh to other persons.

Then, when the optimization loss is below a threshold determined by experimental experience, we accept this generation and simultaneously update $\mathcal{GSG}$ by adding the human node. Otherwise, we consider the generation result implausible and update $\mathcal{GSG}$ by pruning the corresponding object node. It is worth noting that this update way establishes the relationship between each generation and the previous results and avoids a certain degree of crowding, allowing for a better spatial distribution and more realistic interaction than simple multiple generation.

These procedures can be formulated as:

$$\begin{aligned} I_g &= \text{Narrator}(S, \mathcal{GSG}_i, \mathcal{LSG}_i, \mathcal{IA}_i), \\ I_g &= \text{Opt}(I_g), \\ \mathcal{GSG}_{i+1} &= \text{Update}(I_g). \end{aligned} \tag{4}$$

With this design, our model can deal with multi-person interaction generation trained on existing datasets.

7. Experiments

7.1. Datasets

PROX [12] includes 12 different 3D room scans, and captures natural actions of 20 subjects represented by SMPL-X body meshes. For fair comparison, we follow the same split of the train and test set as [13, 30, 37] and generate HSIs on the unseen scenes during training.

As there is no existing dataset suitable for our task, we annotate all video frames from PROX [12] with their corresponding textual descriptions about interactions to evaluate our natural and controllable HSI generation. To accurately describe human interaction in natural language, we design a combinatorial template following Sr3D [1]:

$$\begin{aligned} &< \text{subject} >< \text{action} >< \text{object-class} > \\ &< \text{spatial-relationship} >< \text{anchor-class(es)} >, \end{aligned} \tag{5}$$

where the action is taken from the motion labels in BABEL [23] and can be described as multiple actions. The *object* and the *anchor* are the interacting object and the other objects used for associative localization, respectively. The *spatial-relationships* are "support", "surrounding", "next-to", "between" and so on, which together provide spatial location and object distribution for generation.

To demonstrate the generalisation capability of the proposed framework, we further evaluate it on Matterport3D [5] and ScanNet [8], which both provide large-scale reconstructed 3D scenes. Please note that our framework does not utilize Matterport3D and ScanNet for training.

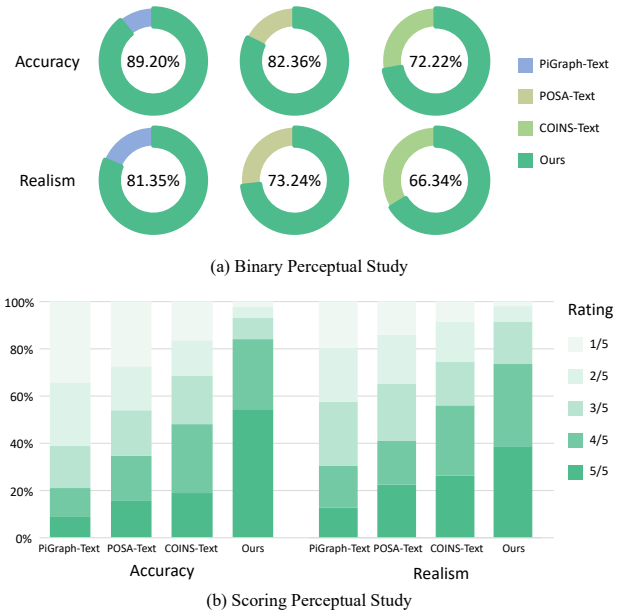
7.2. Baselines

Currently available methods do not allow for naturally controllable HSI generations directly from textual descriptions. Thus, we modify three state-of-the-art HSI methods and train their official models using the same dataset.

PiGraph-Text. PiGraph [26] generates scene placement and human skeletons from interaction category specifications. We remove the scene placement step, represent the body with SMPL-X model and replace verb-noun pairs with textual descriptions. We denote this modified PiGraph variant as PiGraph-Text.

POSA-Text. POSA [13] populates 3D scenes with humans guided by per-vertex contact features, but lacks effective control. To incorporate semantic guidance, we first generate body meshes that match textual descriptions and then place them in the appropriate positions using POSA. We denote this modified POSA variant as POSA-Text.

COINS-Text. COINS [37] synthesizes HSIs given interaction semantics as action-object pairs. However, COINS only works for relatively limited interaction combinations and cannot handle complex spatial relationships. Therefore, we extend more interactions and modify it to a two-stage process: first find the area of possible interactions based on BERT and then run COINS as is within that area. We denote this modified COINS variant as COINS-Text and present fairness experiments for COINS in the Supp.Mat.



(a) Binary Perceptual Study

(b) Scoring Perceptual Study

Figure 3: Perceptual study results (anonymized and order-randomized) comparing our approach against three baselines. In the binary perception study (a), the percentage numbers indicate the proportion of respondents who preferred our approach compared to another baseline. In the scoring perception study (b), the different colored bars indicate the percentage of corresponding ratings.

7.3. Evaluation Metrics

Physical Plausibility. From the physical perspective, we evaluate the contact and non-collision scores between the generated body and scene mesh. The former is calculated as the proportion of actual contact vertices for all body vertices with contact labels, while the latter is calculated as the ratio of the number of body vertices with non-negative scene SDF values and the number of all body vertices. In addition, we also evaluate the non-collision score between human bodies for the MHSI.

Diversity. We perform K-Means (K = 50) clustering on the generated human-scene interactions and report: (1) the entropy of the cluster ID histogram, and (2) the cluster size, *i.e.*, the average distance between the cluster center and the samples belonging in it.

Text-to-Generation Distance. For MHSI, we additionally calculate the difference between the text feature from the given description and the visual feature from the results which are randomly generated for 103 results in all scenes.

Perceptual Study. We evaluate the interaction realism and the accuracy of generated results by conducting perceptual studies, which consisted of two main parts: (1) we perform a binary-choice perceptual study in which samples gener-

Methods	Perceptual Accuracy (↑)	Physical Plausibility		Diversity	
		Contact (↑)	Non-Collision (↑)	Entropy (↑)	Cluster Size (↓)
PiGraph-Text	2.81±1.30	0.84	0.81	3.56	1.75
POSA-Text	3.14±1.43	0.72	0.96	3.73	1.96
COINS-Text	3.48±1.35	0.91	0.93	3.98	1.83
Ours	**4.02±0.97**	**0.94**	**0.98**	**4.16**	**1.54**

Table 2: Quantitative comparison with three baselines. Perceptual accuracy is used to evaluate the degree of consistency with textual descriptions. Contact score and non-collision score are used to evaluate interaction realism and plausibility. Entropy and cluster size are used to evaluate interaction diversity.

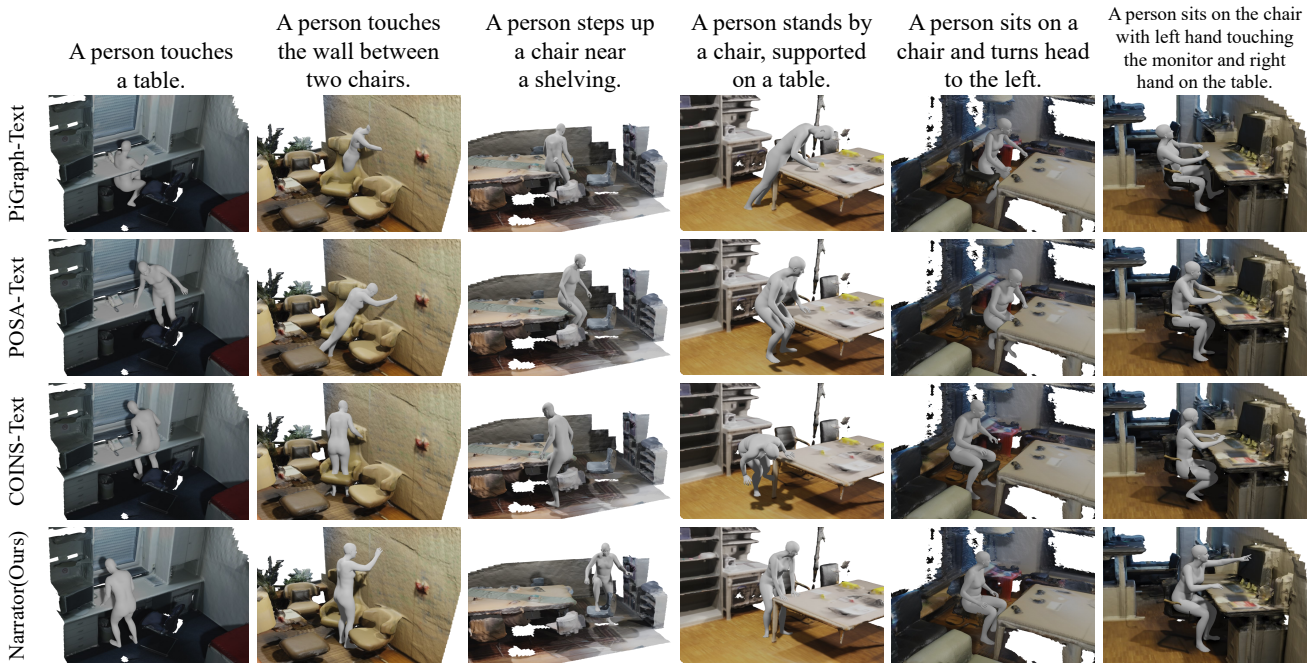


Figure 4: Qualitative comparison of interactions generated with our approach and three baselines. We present different textual queries in columns and different methods in rows. Overall, our interaction generations are semantically more consistent with textual descriptions and physically more realistic with scene interactions.

ated by different methods based on the same textual description are displayed, and the respondents are asked to select the more realistic and natural sample; (2) the respondents are also asked to rate the accuracy and the consistency between shown interaction samples and textual descriptions, from 1 (strongly disagree) to 5 (strongly agree). Please note that the order is randomly swapped in each display. For more details regarding the perceptual studies, please refer to the Supp.Mat.

7.4. Comparison

Perceptual study. Fig. 3 shows the results of the perception study in terms of accuracy (correspondence to the textual descriptions) and realism. Respondents perceive our generated results as better matching the textual descriptions compared to the three baselines, while our generations are also clearly preferred in terms of realism.

Quantitative comparison. Tab. 2 shows the quantitative results compared to the three baselines. It can be seen that our approach has the highest accuracy and the best match to the textual description. our approach achieves higher contact and non-collision scores in terms of physical plausibility, demonstrating our ability to to effectively alleviate scene-body interpenetration and maintain plausible contact relationships. As for the diversity metrics, our approach has greater cluster entropy and smaller cluster size, achieving diversity with guaranteed accuracy.

Qualitative comparison. We further provide qualitative comparisons in Fig. 4 with the three baselines. PiGraph-Text suffers from more severe penetration problems due to the limitations of its own representation. POSA-Text requires finding body placements and tends to fall into local minimums during optimization, thus generating poor interactive contacts. COINS-Text binds actions to specific ob-

Methods	Human-Scene		Human-Human	Text-to-Generation
	Contact (↑)	Non-Collision (↑)	Non-Collision (↑)	Distance(↓)
COINS (multi-gen.&opt.)	0.895	0.913	0.829	5.748±0.113
Ours	**0.932**	**0.967**	**0.954**	**3.058±0.003**

Table 3: Quantitative comparisons on MHSI with the sequential per-person generation and optimization method using COINS. Human-human non-collision score and text-to-generation distance are used to evaluate the physical plausibility and semantic consistency of multi-human generation, respectively.

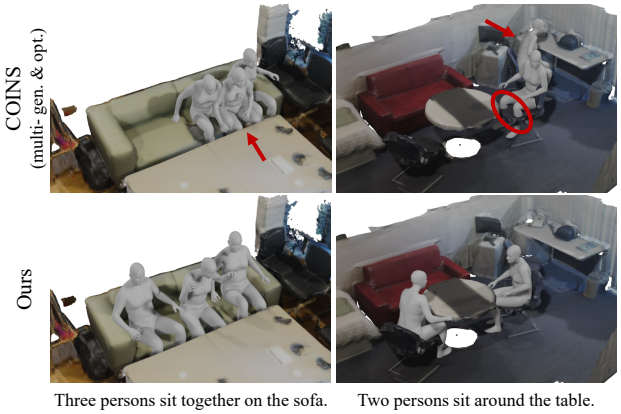


Figure 5: Qualitative comparisons with the sequential per-person generation and optimization method using COINS [37].

Figure 6: More generation results using our approach for MHSI on PROX [12] (a), Matterport3D [5] (b), and ScanNet [8] (c) datasets.

jects and lacks global awareness of the scene, which leads to penetration with unspecified objects and difficulty in handling complex spatial relationships. In contrast, our approach can produce better results by correctly reasoning about spatial relationships and profiling the human body under multiple actions, from different levels of textual descriptions. More results in the Supp.Mat.

7.5. Multi-Human Scene Interaction

Perceptual study. We conduct a perception study on MHSI and receive approval from the respondents: 84.69% of the respondents think that the results match textual descriptions, while 74.80% think that interactions are human-like and natural.

Methods	Physical Plausibility		Diversity	
	Contact	Non-Collision	Entropy	Cluster Size
Full(BERT)	0.92	0.95	3.82	1.91
w/o PLA	0.91	0.89	3.73	1.66
w/o IBS	0.90	0.94	3.95	1.87
Full	**0.94**	**0.98**	**4.16**	**1.54**

Table 4: Quantitative results of ablation study.

Figure 7: Qualitative results of ablation study on JGLSG.

Quantitative comparison. We provide quantitative comparisons and detailed analysis to better understand the contribution of our MHSI strategy. As mentioned earlier, considering that there is no effort for MHSI for the time being, we chose a straightforward way as a baseline, *i.e.*, sequential per-person generation and optimization method using COINS. For a fair comparison, we equally introduce the human collision loss. As can be seen in Tab. 3, in terms of the human-human non-collision score, our method achieves an improvement from the baseline of 0.829 to 0.954, which strongly demonstrates the effectiveness of our iterative generation strategy. Moreover, the feature distance between text and generation also indicates that our approach has more accurate semantic consistency.

Qualitative comparison. As can be seen in Fig. 5, our generation results are more natural, physically plausible, and semantically consistent, while the results of COINS are often crowded together or in the wrong position. Additionally, our more MHSI results on different scene datasets are provided in Fig. 6, including multi-person constraint, per-person constraint, and complex combination of spatial relationships and interactive actions.

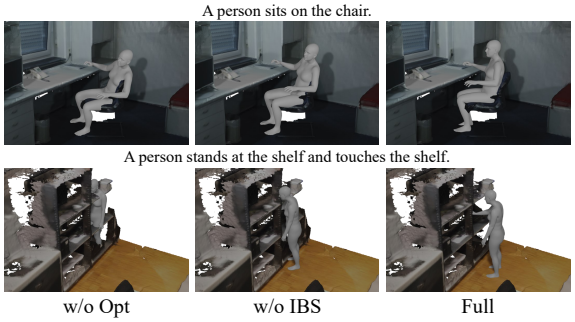


Figure 8: Qualitative results of ablation study on $\mathcal{L}_{\text{IBS}}$.

7.6. Ablation Study

In this section, we evaluate the effect of three components of our framework.

JGLSG. We study the effect of the JGLSG mechanism on interaction generation by replacing it with BERT, abbreviated as Full(BERT). Quantitative results in terms of physical plausibility and diversity are shown in Tab. 4, and demonstrate that our approach is more effective compared to the version using BERT. From the qualitative comparison in Fig. 7, we can see that the interaction results obtained with the help of the JGLSG mechanism for reasoning about spatial relationships, are more accurate and more consistent with the textual description.

PLA. Tab. 4 also shows the impact of the PLA mechanism on interaction generation. Our model can refine various types of interactive actions to produce more reasonable results that are consistent with textual veracity.

IBS loss. To better handle penetration and contact issues, we additionally introduce $\mathcal{L}_{\text{IBS}}$. Tab. 4 and Fig. 8 show comparative results using $\mathcal{L}_{\text{IBS}}$ and without it, demonstrating that it can improve interaction plausibility.

8. Conclusion and Discussion

Conclusion. In this paper, we observe from a narrator's perspective that humans describe human interactions in scenes through spatial perception and action recognition. We propose, *Narrator*, a novel relationship reasoning-based generative approach for naturally controllable generation of human scene interactions from textual descriptions. We design a JGLSG mechanism to reason about spatial relationships, and introduce a PLA mechanism for diverse interactive actions. In particular, benefiting from relationship reasoning, we further propose the first naturally controllable generation strategy for multi-human scene interaction. Experimental results demonstrate that our approach can controllably generate complex and diverse interactions and significantly outperform existing works.

Limitation. Our work is mainly limited to two aspects: (1) there is still room for expansion in our interaction categories, which requires large-scale datasets; (2) our approach focuses on static interactions, while dynamic interactions are also interesting and meaningful. In further work, we will explore richer datasets and extend the scene graph to address more diverse situations, such as those involving the properties of objects or human movements.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (62122058, 62171317 and 62125107), and Tianjin Science Fund for Distinguished Young Scholars (22JCJQJC00040).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *Eur. Conf. Comput. Vis.*, pages 422–440, 2020. 6

[2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *Int. Conf. Rob. Autom.*, pages 5915–5920, 2018. 3

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *Int. Conf. 3D Vis.*, pages 719–728, 2019. 3

[4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3D humans. *arXiv preprint arXiv:2209.04066*, 2022. 3

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 6, 8

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3D object localization in rgb-d scans using natural language. In *Eur. Conf. Comput. Vis.*, pages 202–221, 2020. 3

[7] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Int. Conf. Comput. Vis.*, pages 8648–8657, 2019. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 6, 8

[9] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3D visual graph network for object grounding in point cloud. In *Int. Conf. Comput. Vis.*, pages 3722–3731, 2021. 3

[10] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Int. Conf. Comput. Vis.*, pages 1396–1406, 2021. 3

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5152–5161, 2022. 3

[12] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Int. Conf. Comput. Vis.*, pages 2282–2292, 2019. 2, 3, 6, 8

[13] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3D scenes by learning human-scene interaction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14708–14718, 2021. 2, 3, 5, 6

[14] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3D avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3

[15] 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. *arXiv preprint arXiv:2301.06015*, 2023. 3

[16] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Trans. Graph.*, 33(4):1–12, 2014. 2

[17] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *Int. Conf. Comput. Vis.*, pages 9267–9276, 2019. 4

[18] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12368–12376, 2019. 2

[19] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3D-sps: Single-stage 3D visual grounding via referred point progressive selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16454–16463, 2022. 3

[20] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Trans. Graph.*, 38(4):1–15, 2019. 2

[21] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10975–10985, 2019. 3

[22] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109*, 2022. 3

[23] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 722–731, 2021. 6

[24] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inform. Process. Syst.*, 2017. 2, 3

[25] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3D visual grounding. In *Conf. Rob. Learn.*, pages 1046–1056, 2022. 3

[26] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Trans. Graph.*, 35(4):1–12, 2016. 2, 3, 6

[27] Ziyang Song, Dongliang Wang, Nan Jiang, Zhicheng Fang, Chenjing Ding, Weihao Gan, and Wei Wu. Actformer: A gan transformer framework towards general action-conditioned 3D human motion generation. *arXiv preprint arXiv:2203.07706*, 2022. 3

[28] Fuwen Tan, Crispin Bernier, Benjamin Cohen, Vicente Ordonez, and Connelly Barnes. Where and who? automatic

semantic-aware person composition. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 1519–1528, 2018. 2

[29] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3D semantic scene graphs from 3D indoor reconstructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3961–3970, 2020. 4

[30] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3D human motion synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20460–20469, 2022. 2, 3, 5, 6

[31] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3D human motion and interaction in 3D scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9401–9411, 2021. 2

[32] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6609–6618, 2019. 4

[33] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2D semantics assisted training for 3D visual grounding. In *Int. Conf. Comput. Vis.*, pages 1856–1866, 2021. 3

[34] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. *arXiv preprint arXiv:2206.04382*, 2022. 3

[35] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3D environments. In *Int. Conf. 3D Vis.*, pages 642–651, 2020. 3

[36] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3D people in scenes without people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6194–6204, 2020. 2, 3, 5

[37] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. *arXiv preprint arXiv:2207.12824*, 2022. 2, 3, 4, 5, 6, 8

[38] Xi Zhao, He Wang, and Taku Komura. Indexing 3D scenes using the interaction bisector surface. *ACM Trans. Graph.*, 33(3):1–14, 2014. 5

[39] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5745–5753, 2019. 3

[40] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3D indoor scene augmentation. In *Int. Conf. Comput. Vis.*, pages 7384–7392, 2019. 4