HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes

Anonymous Author(s)
Affiliation
Address
email

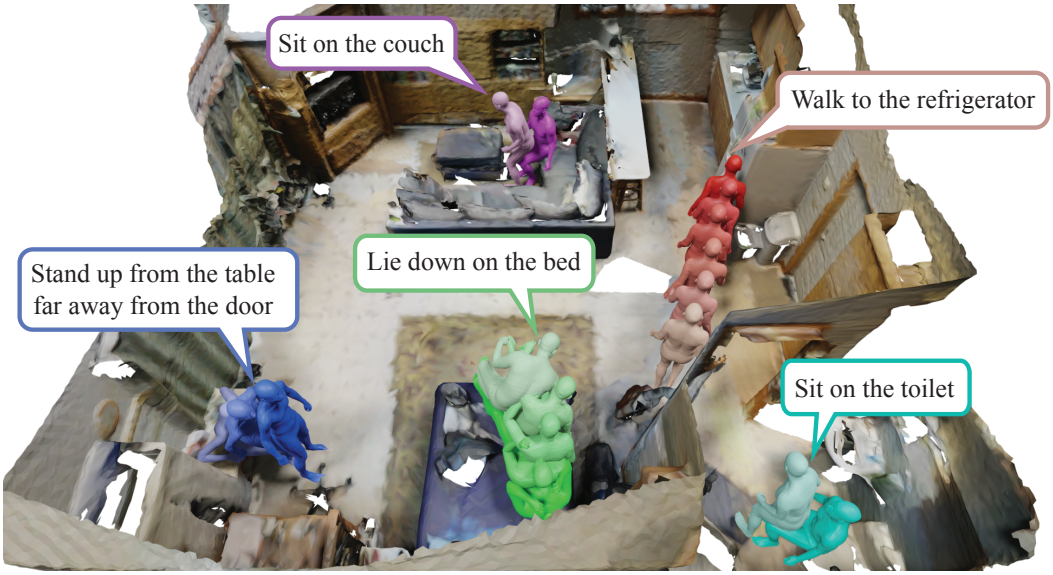


Figure 1: **Examples from our proposed Human-Scene Interaction (HSI) dataset—*HUMANISE*.** It contains 4,892 high-quality motion sequences in 512 indoor scenes. Each motion segment has rich semantic information about the action type and the singular interacting objects unambiguously, specified by the language description.

Abstract

Learning to generate diverse scene-aware and goal-oriented human motions in 3D scenes remains challenging due to the mediocre characters of the existing datasets on Human-Scene Interaction (HSI); they only have limited scale/quality and lack semantics. To fill in the gap, we propose a large-scale and semantic-rich synthetic HSI dataset, denoted as *HUMANISE*, by aligning the captured human motion sequences with various 3D indoor scenes. We automatically annotate the aligned motions with language descriptions that depict the action and the individual interacting objects; *e.g.*, sit on the armchair near the desk. *HUMANISE* thus enables a new generation task, *language-conditioned human motion generation in 3D scenes*. The proposed task is challenging as it requires joint modeling of the 3D scene, human motion, and natural language. To tackle this task, we present a novel scene-and-language conditioned generative model that can produce 3D human motions of the desirable action interacting with the specified objects. Our experiments demonstrate that our model generates diverse and semantically consistent human motions in 3D scenes. We believe our proposed dataset and model can facilitate the HSI research, and will release them for research purposes.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

1 Introduction

Imagine instructing a virtual human or robot to "sit on the armchair near the desk." Upon accepting the instruction, the agent needs to comprehend its semantic meaning (*i.e.*, *sit*) and understand the surrounding environment (*i.e.*, *the armchair near the desk*) to generate desired interactions with the 3D scene. The vision community has been trying to attain this goal by automating the realistic human motion generation but still struggles in connecting the instructions to the 3D human motions.

Specifically, existing works generate plausible human poses [Zhang et al., 2020b;a] or motions [Starke et al., 2019; Wang et al., 2021a; Cao et al., 2020; Wang et al., 2021b; Hassan et al., 2021a] in 3D scenes by adopting learning-based methods. Training these learning-based models exploits existing HSI datasets: PROX [Hassan et al., 2019] and GTA-IM [Cao et al., 2020]. Although these datasets contain 3D people moving in 3D indoor scenes, they suffer from two fundamental limitations that prevent existing methods from learning instruction-aware 3D human motions.

- **Limited scale and quality**. PROX [Hassan et al., 2019] only contains 12 indoor scenes with jittering human motion. The synthetic GTA-IM [Cao et al., 2020] only has 49 scenes and lacks 3D human shapes and diverse movement styles.
- **Absence of scene and language semantics**. PROX [Hassan et al., 2019] and GTA-IM [Cao et al., 2020] both suffer from incomplete 3D semantic segmentation annotations and the lack of human action labels/descriptions, making it infeasible to generate instruction-aware human motions.

To tackles the issues above, we propose a large-scale and semantic-rich synthetic HSI dataset, *HUMANISE* (as shown in Fig. 1), by aligning the captured human motion sequences [Mahmood et al., 2019] with the scanned indoor scenes [Dai et al., 2017]. Specifically, for an action-specific motion sequence (*e.g.*, *sit*), we first sample an interacting target object (*e.g.*, *armchair*) and the possible interacting positions on the object surface. Next, we sample the valid translation and orientation parameters by employing the *collision* and *contact* constraints, such that the interactions between the transformed human and the scene are physically plausible and visually natural. We automatically annotate the synthesized motion sequences with template-based language descriptions; for instance, *sit on the armchair near the desk*. To specify the interacting objects unambiguously, we adopt the object referential utterances in Sr3D [Achlioptas et al., 2020]. In total, *HUMANISE* contains 4,892 high-quality motion sequences in 512 indoor scenes. Each motion segment has rich semantics about the action type and the corresponding interacting objects, specified by the language description.

Based on *HUMANISE*, we solve a new task: *language-conditioned human motion generation in 3D scenes*. It requires the model to generate diverse and physically plausible human motions with designated action types and interacting 3D objects that are specified by the language descriptions. This task is more challenging than previous motion generation tasks: (i) Both the 3D scene and the language description pose constraints on motion generation. (ii) The generated human motions should perform the correct action and be precisely grounded near the target location according to the language descriptions. (iii) The generated human motions should be realistic and physically plausible within the 3D scenes. Therefore, an ideal model should equip with capabilities of *generation*, *multi-modal understanding*, and *language grounding*.

Due to the demanding nature of this new task, we utilize the conditional variational auto-encoder (cVAE) [Sohn et al., 2015] framework to generate human motions conditioned on the given scene and language description. The scene and language conditions are learned with separate encoders and fused with the self-attention mechanism to generate the conditional embedding. The input motion is encoded with a recurrent module and decoded with a transformer decoder to generate human motions. We further design two auxiliary losses for target object grounding and action-specific generation. The qualitative and quantitative results show that our model generates diverse and semantically consistent human motions in 3D scenes; it outperforms the baselines on various evaluation metrics. By evaluating on downstream applications, further experiments demonstrate that *HUMANISE* alleviates the limitations of existing HSI datasets.

This paper makes three primary contributions: (i) We propose a large-scale and semantic-rich synthetic HSI dataset, *HUMANISE*, that contains human motions aligned with 3D scenes and corresponding language descriptions. (ii) We introduce a new task of *language-conditioned human motion generation in 3D scenes* that requires a holistic and joint understanding of scenes, humans, and language. (iii) We develop a generative model that produces diverse and semantically consistent human motions given the 3D scene and language description.

2 Related Work

Conditional Human Motion Synthesis Human motion synthesis can be conditioned on past motion [Li et al., 2018; Barsoum et al., 2018; Yuan and Kitani, 2020; Cao et al., 2020; Xie et al., 2021], 3D scenes [Starke et al., 2019; Wang et al., 2021a;b; Hassan et al., 2021a], objects [Chao et al., 2019; Taheri et al., 2020], action labels [Guo et al., 2020; Petrovich et al., 2021], and language descriptions [Ahuja and Morency, 2019; Ghosh et al., 2021; Lin and Amer, 2018]. Given a furnished 3D indoor scene, Zhang et al. [2020b;a]; Hassan et al. [2021b;a]; Wang et al. [2021a;b] generate physically plausible single-frame human pose or long-term human motions. Due to the lack of semantics constraints, these methods either generate random/uncontrollable human motions or need target positions as the input. Conditioned on informative textual descriptions, Lin and Amer [2018]; Ahuja and Morency [2019]; Ghosh et al. [2021] generate human motions that meet the input descriptions of how a human motion is performed. However, these works focus only on motions alone and neglect the scene context. In this work, we incorporate conditions on both scenes and language descriptions, generating human motion sequences consistent with the specified action and location.

Human-Scene Interaction (HSI) Datasets Real and synthetic datasets have been proposed to study the HSI related topics. PiGraphs [Savva et al., 2016] is captured with RGB-D videos and contains estimated human skeletons in 30 scanned scenes. However, it suffers from inaccurately reconstructed scenes, noisy human poses, and a low frame rate. PROX [Hassan et al., 2019] proposes new pipelines to reconstruct human motions from monocular RGB-D sequences captured in 12 indoor scenes. Nevertheless, the gap between their pose fitting results and high-quality, large-scale MoCap data (*e.g.*, AMASS [Mahmood et al., 2019]) is still non-trivial. GTA-IM [Cao et al., 2020] exploits the game engine to collect a synthetic dataset of human skeletons, but with limited human-scene interactions and scene diversities. In comparison, our *HUMANISE* dataset synthesizes the Human-Scene Interaction (HSI) by aligning the high-quality motions in AMASS [Mahmood et al., 2019] with real-world 3D scenes in ScanNet [Dai et al., 2017]. As a result, *HUMANISE* contains high-quality and realistic human-scene interactions with high diversity (*i.e.*, 512 scenes, $4,892$ motions) and rich semantics captured in the auto-generated language descriptions.

Language Grounding in 3D Scenes Grounding 3D objects with language has been explored in 3D reference [Achlioptas et al., 2020; Thomason et al., 2022], 3D question answering (QA) [Das et al., 2018; Ye et al., 2021; Azuma et al., 2021], and embodied navigation [Anderson et al., 2018; Gordon et al., 2019; Gan et al., 2020]. With the language description condition, we aim to generate human motions that interact with the specified object in the scene. However, locating the interacting objects in the 3D scene with referential descriptions is still challenging [Roh et al., 2022; Chen et al., 2020] due to the noisy scanned scenes, diverse expressions, and complex spatial reasoning. To learn better conditional embedding that facilitates to locate and ground according to language, we introduce a multi-model conditional module that learns a joint embedding from textual descriptions and 3D point clouds. We further enhance the model's spatial and action grounding capabilities by introducing two auxiliary losses of locating the interacting 3D objects and action-specific generation.

3 The *HUMANISE* Dataset

A large-scale and high-quality dataset with rich semantic information is in need to facilitate the research in human motion synthesis and HSI. In this work, we propose a synthetic dataset by leveraging the existing datasets in human motion (*i.e.*, AMASS [Mahmood et al., 2019]), 3D indoor scene (*i.e.*, ScanNet [Dai et al., 2017], and additional annotations of these datasets (*i.e.*, motion action label in BABEL [Punnakkal et al., 2021] and 3D referential descriptions in Sr3D [Achlioptas et al., 2020]). Our novel automatic synthesis pipeline avoids expensive hardware and time-consuming labor for realistic capture. Fig. 2 shows more samples in *HUMANISE*. Of note, our dataset synthesis pipeline can generalize to other datasets and actions; please refer to *Sup. Mat.* for more details.

Below, we start by describing to align the motion with the 3D scene in Sec. 3.1, followed by how to automatically generate the language description in Sec. 3.2. Sec. 3.3 provides dataset statistics.

3.1 Motion Alignment

The idea of synthesizing *HUMANISE* is straightforward: Automatically aligning captured human motions with 3D indoor scenes. Given a human motion sequence and its action label, we first sample

Figure 2: **Samples from *HUMANISE* dataset**. *HUMANISE* includes representative interactive actions (*e.g.*, sit, stand-up, lie-down, walk) in diverse indoor scenes (*e.g.*, office, bedroom, living room, dining room).

an interacting target object and possible interacting positions on the object surface in the 3D scene. Next, we sample the valid global translation $\mathcal{T}$ and rotation $\mathcal{R}$ parameters by employing the *collision* and *contact* constraints. We align the human motion with the 3D scene through a transformation with $\mathcal{T}$ and $\mathcal{R}$, while all the local poses remain unchanged.

Collision Constraint We treat the point cloud of the 3D scene as a 3-dimensional k-d tree. Collision avoidance of human motions is equivalent to finding the closest point to the queried human vertex in the scene by searching through the k-d tree. Collision happens as the distance between the human vertex and the closet point in the scene is less than a threshold. A valid motion should avoid having collisions with the objects or walls for all the poses along the trajectory.

Contact Constraint We decompose this constraint into the foot-ground support constraint and the proximity between the human and the interacting object. The support constraint is employed with a distance threshold between foot and the ground plane. The human-object proximity differs among actions; we design the following proximity constraints to make the interaction realistic.

- **Sit** We sample a sittable object (*e.g.*, a chair) as interacting objects and uniformly sample 3D points on the object surface as potential contact points. We assume the final pose of the sitting motion should contact the sittable object on the hip area, such that the distances between the hip vertices and the contact point should be less than a threshold.
- **Stand-up** Since the stand-up motion interacts with the sittable object at the starting frame, we enforce the similar constraints as in *sit* on the starting pose, instead of the final pose.
- **Walk** The language description in *HUMANISE* instructs the walking motion to a target location in the scene, usually to an object; *e.g.*, *walk to the refrigerator*. We sample dense positions near the specified object on the floor as the potential targets for the final pose of walking motion. All the body poses along the walking trajectory should satisfy the feet-ground support and avoid collisions.
- **Lie-down** We select objects with a large flat surface (*e.g.*, bed) as the interacting objects. The translation $\mathcal{T}$ and rotation $\mathcal{R}$ are sampled such that the final pose of the lie-down motion is in contact with the object's top surface. Apart from the non-collision and contact constraint, we require the human body to lie within the object area.

Diverse Affordance We can sample motions representing diverse 3D affordance with the motion alignment algorithm. For instance, *stand up from the coffee table*, *lie down on the office table*, and *sit on the toilet*. *HUMANISE* contains these Human-Object Interactions (HOIs) that are not frequently captured in daily activity or previous HSI datasets. They encode meaningful and versatile object/scene affordances, which introduce both potentials and challenges for the model learning.

3.2 Language Description Generation

The motion descriptions in *HUMANISE* depict the action type and the singular interacting object/location unambiguously. To automatically generate these language descriptions, we design a compositional template following Sr3D [Achlioptas et al., 2020]:

$$< \text{action} >< \text{target-class} > [< \text{spatial-relation} >< \text{anchor-class(es)} >]. \quad (1)$$

For instance, sit on the armchair near the desk. The template contains four placeholders, of which *spatial-relation* and *anchor* are optional placeholders. The *action* is taken from the motion labels in BABEL [Punnakkal et al., 2021], whereas the *target* represents the interacting object. To uniquely refer to a target object, a *spatial-relation* is defined between the *target* and a surrounding object (*anchor*). We utilize the referential utterances in Sr3D [Achlioptas et al., 2020], which defines five types of spatial relations: horizontal proximity, vertical proximity, between, allocentric, and support.

3.3 Dataset Statistics

In total, *HUMANISE* includes 4,892 human motion sequences in 512 3D scenes. We choose the four most representative interactive indoor actions, *i.e.*, sit (1299), stand up (936), lie down (169), and walk (2488). Tab. 1 compares *HUMANISE* against existing HSI datasets, wherein HUMANISE shows advantages in clip number, scene variety, and human pose quality. It is the only dataset with rich semantic information of HSIs. Fig. 2 shows more examples from *HUMANISE*.

Table 1: **Comparison between HUAMNISE and existing HSI datasets.**

Dataset	#Scenes	#Clips	#Frames	Human Representation	Pose Jittering	Semantics
PiGraph [Savva et al., 2016]	30	63	0.1M	Skeleton	✓	✗
PROX-Q [Hassan et al., 2019]	12	60	0.1M	Shape	✓	✗
GTA-IM [Cao et al., 2020]	49	119	1.0M	Skeleton	✗	✗
HUMANISE	512	4,892	0.3M	Shape	✗	✓

4 Language-conditioned Human Motion Generation in 3D Scenes

4.1 Problem Definition

Given a 3D scene and a language description, our goal is to generate a realistic human motion sequence that is both semantically consistent with the language description and physically plausible in the scene. We denote the given scene as $S \in \mathbb{R}^{N\times 6}$, representing an RGB-colored point cloud of N points. The language description is a tokenized word sequence of length D, denoted as $L_{1:D} = [w_1, \cdots, w_D]$. We represent the human pose and shape in the motion sequence with the SMPL-X [Pavlakos et al., 2019] body model. Specifically, the SMPL-X body mesh $\mathcal{M} \in \mathbb{R}^{10475\times 3}$ is parameterized by $\mathcal{M} = F(t, r, \beta, p)$, where $t \in \mathbb{R}^3$ is the global translation, $r \in \mathbb{R}^6$ is a continuous representation [Zhou et al., 2019] of the global orientation, $\beta \in \mathbb{R}^{10}$ is the body shape parameters, $p \in \mathbb{R}^{J\times 3}$ is J joint rotations in axis-angle, and F is a differentiable linear blend skinning function. As we focus on producing the plausible human position and pose, we use the default mean body shape for β and denote the parameters of interest as $\Theta = \{t, r, p\}$.

4.2 Method

We propose a novel generative model to generate human motions conditioned on the given scene and language description. The overall architecture is illustrated in Fig. 3. Specifically, we employ a conditional variational auto-encoder (cVAE) framework to model the probability $p(\Theta_{1:T} \mid S, L_{1:D})$, where T is the length of the motion sequence. Below, we describe the details of each module.

Condition Module The condition module takes input from two modalities (*i.e.*, the given scene S and the language description L) and outputs a joint conditional embedding. To process the 3D scene, we employ the Point Transformer [Zhao et al., 2021] to produce point-level features $[u_1, \cdots, u_{N'}]$, where the transition-down module reduces the cardinality of the scene point cloud from N points to N' points. For the language description $L_{1:D}$, we use the pre-trained BERT [Kenton and Toutanova, 2019] to extract the word-level embedding $[v_1, \cdots, v_T]$.

The point-level features and the word-level features are concatenated and fed into the feature fusion module, a single-layer self-attention module. In this way, both the point-level and the word-level

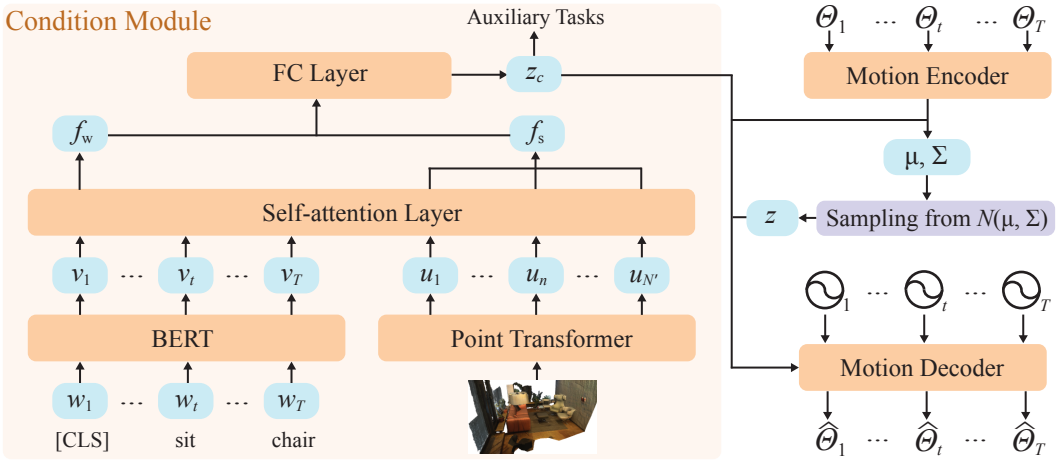


Figure 3: **Illustration of the proposed generative model.** It adopts the cVAE and incorporates a motion encoder and decoder to generate diverse human motions. The condition is learned from the joint embedding of 3D scene and language description with auxiliary tasks of 3D location grounding and action recognition.

features attend to the features from two modalities by self-attention. The updated point features are concatenated and forwarded to a fully-connected (FC) layer to obtain the scene feature f_s. We take the updated $[CLS]$ feature as the language feature f_w. The scene and language features are finally concatenated and mapped to a conditional latent embedding z_c with FC layers.

Motion Encoder We first use a bidirectional GRU to obtain a sequence-level feature of the input motion $\Theta_{1:T}$. Next, the output is concatenated with the conditional embedding z_c, followed by an MLP layer to predict the Gaussian distribution parameters (*i.e.*, μ and Σ). Finally, we sample a latent vector z from the distribution using the reparameterization trick [Kingma and Welling, 2013].

Motion Decoder Following Petrovich et al. [2021], we use a transformer decoder to generate a sequence of parameters $\widehat{\Theta}_{1:T}$ in a given duration T. Specifically, we use T sinusoidal positional embeddings to query the concatenation of the sampled latent z and the conditional embedding z_c. The outputs of the transformer decoder are mapped into body meshes with the followed differentiable SMPL-X model. Of note, although the duration T equals to the length of the input motion sequence in the training phase, it can be arbitrary values during inference.

4.3 Training Loss

In our work, the training loss contains reconstruction loss, KL divergence loss, and two auxiliary losses for target object grounding and action-specific generation, formulated as

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha_{kl}\mathcal{L}_{kl} + \alpha_o\mathcal{L}_o + \alpha_a\mathcal{L}_a, \tag{2}$$

where α denotes the loss weight of each term.

Reconstruction Loss Our reconstruction loss is formulated as

$$\mathcal{L}_{rec} = \mathcal{L}_t + \alpha_r\mathcal{L}_r + \alpha_p\mathcal{L}_p + \alpha_v\mathcal{L}_v, \tag{3}$$

where $\mathcal{L}_t$, $\mathcal{L}_r$, and $\mathcal{L}_p$ are ℓ_1 distance between the predicted body parameters and the ground truth (*i.e.*, global translation t, global rotation r, and joint rotations p). $\mathcal{L}_v$ is an additional term for regressing the positions of the vertices on the SMPL-X body mesh. In practice, the vertex set can be all 10475 vertices on SMPL-X body mesh or selected sparse markers [Zhang et al., 2021].

KL Divergence Loss Denoting the learned distribution of latent z as $\psi = \mathcal{N}(\mu, \Sigma)$, we regularize ψ to be similar to the normal distribution $\phi = \mathcal{N}(0, \mathbf{I})$, *i.e.*, $\mathcal{L}_{kl} = D_{KL}(\psi \| \phi)$.

Auxiliary Loss Notably, the learning of the conditional latent embedding z_c is crucial to our generative model, as it represents compound information of the action, the interacting object, and the 3D scene context from two different modalities. As such, rather than simply maximizing the negative evidenced lower bound (ELBO), we introduce two auxiliary tasks for locating the interacting target object and generating action-specific motions.

For the first auxiliary task, we directly use a FC layer on top of the latent z_c to regress the target object's center. For the second auxiliary task, we use another FC layer followed by softmax to map the latent z_c to the probability of each action. Specifically, we use the MSE loss $\mathcal{L}_o$ to regress the target object center and the cross-entropy loss term $\mathcal{L}_a$ to classify the action category.

4.4 Implementation Details

We train our generative model on *HUMANISE* for 150 epochs using Adam [Kingma and Ba, 2014] and a fixed learning rate of 0.0001. For hyper-parameters, we empirically set $\alpha_{kl} = \alpha_o = 0.1$, $\alpha_a = 0.5$, $\alpha_r = 1.0$, and $\alpha_p = \alpha_v = 10.0$. We set the dimension of global condition latent z_c to 512 and latent z to 256. The hidden state size is set to 256 in the single-layer GRU motion encoder. The transformer motion decoder contains two standard layers with the 512 hidden state size. We train our model with a batch size of 24 on a V100 GPU. It takes about 16 hours to converge.

Point Transformer Our point transformer module that processes the scene point cloud adopts the original architecture [Zhao et al., 2021]. For each scene, we down-sample $N = 32768$ points from the original scanned scene and obtain $N' = 128$ points with a 512-D feature for each point after applying the point transformer. We pre-train the point transformer with a semantic segmentation task on ScanNet [Dai et al., 2017] dataset, whose train-test split is the same as *HUMANISE*.

5 Experiments

We benchmark our proposed task, *language-conditioned human motion generation in 3D scenes* on *HUMANISE*, and describe the detailed settings, baselines, analyses, and ablative studies.

Dataset Splitting We split motions in *HUMANISE* according to the original scene IDs and split in ScanNet [Dai et al., 2017], resulting in 4131 motions for training and 761 for testing.

5.1 Experimental Setting

To evaluate the model performance, we conduct experiments in both the action-specific and action-agnostic settings. In the action-specific setting, we train and evaluate our model on *HUMANISE* subsets of each action without the auxiliary task of classifying the action category. In the action-agnostic setting, we train our full model on the whole dataset with auxiliary tasks.

Ablative Baselines We compare our model with three ablative baselines to verify the significance of the auxiliary tasks, the feature fusing module, and the scene encoder. (i) We remove the auxiliary tasks for locating the target interacting object and recognizing the action category, denoted as *w/o aux. loss*. We train this model on the full *HUMANISE* dataset. (ii) Instead of feeding the scene and the language feature into the self-attention layer, we directly concatenate them as the global conditional feature. We denote this model as *w/o self-att*. (iii) We replace the point transformer with PointNet++ [Qi et al., 2017] (*i.e.*, *PointNet++ Enc.*), also pertained with the semantic segmentation task. We train and test the second and the third ablation models on the *walk* subset.

5.2 Evaluation Metrics

Reconstruction Metrics Following Wang et al. [2021a], we introduce the following metrics to evaluate the model's reconstruction capability. We report the ℓ_1 error ($\times 100$) between the predicted SMPL-X parameters and the ground truth, including global translation t, global orientation r, and body pose p. We also compute the Mean Per Joint Position Error (MPJPE) [Ionescu et al., 2013] and Mean Per Vertex Position Error (MPVPE) in millimeters between the reconstructed SMPL-X body mesh and the ground truth to evaluate the reconstruction in a more intuitive and fine-grained manner. We report these metrics by averaging the sequence due to the different motion durations.

Generation Metrics As we hope our generated motion to interact with the correct object specified by the language description, we introduce a body-to-goal distance metric (*goal dist.*). It is the shortest distance (in meters) from the target object to the interacting human body, computed as $\mathrm{Max}(\mathrm{Min}(\Psi^{+}_{\mathcal{M}}(\mathcal{O})), 0)$, where $\Psi^{+}_{\mathcal{M}}(\cdot)$ is the positive signed distance field of body mesh $\mathcal{M}$, and $\mathcal{O} \subset S$ is the target object point set. Of note, we use the human body in the first frame to compute

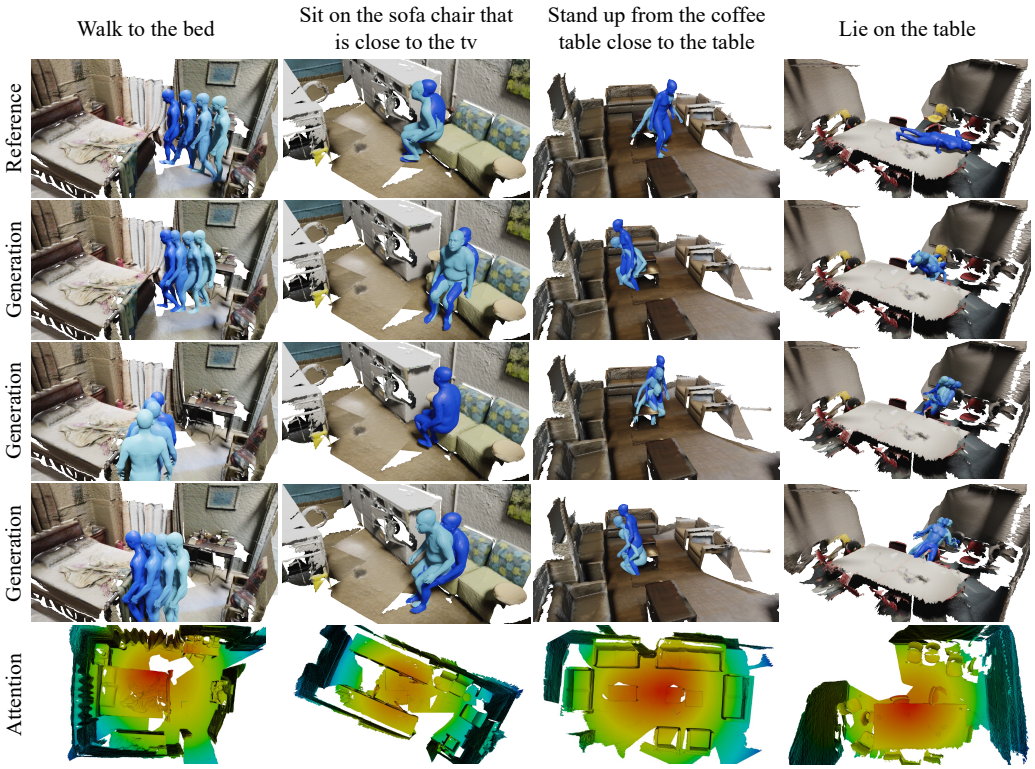


Figure 4: **Qualitative generation results of action-specific models on *HUMANISE* dataset.** We visualize one reference motion and three generation motions for each scenario. The attention map visualizes the attention weights the $[CLS]$ token attend to the scene point cloud in the self-attention layer. Red denotes higher weight.

the body-to-goal distance for *stand up* action and use the last frame for other actions. We generate ten randomly sampled motion sequences for each case in the test split and report the average distance.

Perceptual Study We perform human perceptual studies to evaluate the generation results in terms of the overall quality and action-semantic accuracy. Given the rendered motion and the language description, a worker is asked to score from 1 to 5 for (i) the overall generation *quality*, and (ii) whether the generated motions are performing the *action* specified by the description. A higher value indicates that the generated result is more plausible with the given scene and language description. We randomly generate samples in 20 scenarios for each model, and three workers score each sample.

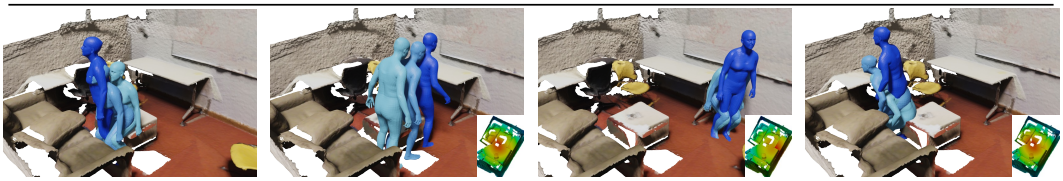
5.3 Results and Discussion

Tab. 2 tabulates the quantitative results of both reconstruction and generation. Fig. 4 and Fig. 5 show some qualitative results. Below we discuss some significant observations.

- As seen in Tab. 2, our model performs well in the action-specific setting, *e.g.*, *walk*, *sit*, and *stand up*. Fig. 4 also shows our model generates human motions that are visually appealing and semantically consistent with the language description. Furthermore, the attention visualizations show that the $[CLS]$ token attends to the point cloud features around the target interacting objects. Meanwhile, we also notice that the performance drops in the *lie down* action, which might be caused by the unbalanced data distribution; the *lie down* subset is only 10% of the *sit* subset.
- The action-agnostic setting is more complex and difficult than the action-specific setting, but the auxiliary tasks improve the model's capability in spatial and action grounding. From Tab. 2, our full model *all actions* outperforms the one *w/o aux. loss* by a large margin in both reconstruction and generation. The significant difference in *goal dist.* and *action score* show that the proposed auxiliary tasks can help our generative model to locate the target 3D object and generate action-specific motions. Qualitative results in Fig. 5 further verifies this observation.
- The ablative experiments on the feature fusing module and scene encoder validate our model designs. For the action *walk*, our full model outperforms the model *w/o self-att.* in all generation metrics. We conclude that the transformer layers learn better conditional embedding by attending to features

Table 2: **Quantitative results of reconstruction and generation on *HUMANISE* dataset.**

Model	Reconstruction					Generation		
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	quality score↑	action score↑
sit	6.35	5.73	6.52	158.42	154.40	0.853	2.88±1.31	4.39±0.82
stand up	10.06	6.77	5.67	235.68	230.22	0.906	3.10±1.13	4.38±0.94
lie down	19.63	41.78	10.83	805.47	759.77	0.995	2.17±1.24	2.76±1.27
walk	8.31	5.75	6.76	206.52	196.45	1.077	3.13±1.33	4.28±1.03
w/o self-att.	8.03	5.25	6.26	192.43	182.96	1.189	3.05±1.18	3.97±1.06
PointNet++ Enc.	8.45	5.55	6.71	209.85	199.30	1.192	3.03±1.39	4.31±1.00
all actions	8.61	6.43	5.71	210.96	205.42	1.081	2.31±1.10	3.17±1.52
w/o aux. loss	8.65	6.65	5.83	210.96	206.04	1.321	2.02±1.00	2.58±1.45



(a) Reference (b) *w/o aux. loss* (c) *w/o aux. loss* (d) Ours

Figure 5: **Ablation results of action-agnostic models.** The model *w/o aux. loss* struggles in (b) generating the action specified by the description or (c) locating the interacting object. (d) Instead, our full model generates motions semantically consistent with the language description.

across the scene and language modalities. The difference on *goal dist.* compared with *PointNet++ Enc.* shows our better grounding performance through finer-grained scene understanding.

Please refer to *Sup. Mat.* for additional experiments, qualitative results, and failure case discussions.

5.4 *HUMANISE* for Downstream Task

We further demonstrate the quality of *HUMANISE* by another existing downstream task proposed in Wang et al. [2021a]: *human motion synthesis in 3D scenes*. This task only considers the scene-conditioned generation, and the language annotations in *HUMANISE* are not used in this setting. Specifically, we train the RouteNet and PoseNet in Wang et al. [2021a] using *HUMANISE* with ten epochs and fine-tune them using PROX [Hassan et al., 2019] with another ten epochs. For a fair comparison, we reproduce the results of the same model trained on PROX for 20 epochs under the same setting. We evaluate both models on the test split of PROX.

The reconstruction errors with metrics from Wang et al. [2021a] are reported in Tab. 3. The model pre-trained on *HUMANISE* outperforms the model trained only with PROX on all evaluation metrics by a large margin, except for the orientation. It shows our dataset can alleviate the limitations of existing HSI datasets and potentially support more applications in HSI related topics.

Table 3: **Results of human motion synthesis on PROX dataset.**

Method	translation	orientation	pose	MPJPE	MPVPE
[Wang et al., 2021a]	7.65	**9.38**	44.52	242.50	222.13
HUMANISE pre-trained	**6.74**	9.43	**41.53**	**222.48**	**203.39**

6 Conclusion

In this work, we propose a large-scale and semantic-rich HSI dataset, *HUMANISE*. It contains diverse and physically plausible interactions equipped with language descriptions. *HUMANISE* enables a new task: language-conditioned human motion generation in 3D scene. We further design a scene-and-language conditioned generative model that produces diverse and semantically consistent human motions based on *HUMANISE*.

Limitations Our work is primarily limited on two aspects. (i) The human motions are relatively short since action-specific motions usually last within 5 seconds. A large-scale, long-duration HSI dataset is in need for the scene understanding community. (ii) The problem of *language grounding in 3D scenes* is critical for our task but far from being solved, as also revealed in recent 3D grounding works [Achlioptas et al., 2020; Thomason et al., 2022]. A robust grounding module could help generate more meaningful human motions based on language description.

References

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 *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5, 9

Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, 2019. 3

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *arXiv preprint arXiv:2112.10482*, 2021. 3

Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5

Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. *arXiv preprint arXiv:1908.07423*, 2019. 3

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, 2020. 3

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 7

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *International Conference on Robotics and Automation (ICRA)*, 2020. 3

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, 2021. 3

Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *International Conference on Multimedia*, 2020. 3

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 5, 9

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021a. 2, 3

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b. 3

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 7

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 5

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 3

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 6

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 7

Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning (CoRL)*, 2022. 3

Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4): 1–12, 2016. 3, 5

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015. 2

Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics (TOG)*, 38(6):209–1, 2019. 2, 3

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 3

Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Conference on Robot Learning (CoRL)*, 2022. 3, 9

Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. 2, 3, 7, 9

Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b. 2, 3

Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*, 2021. 3

Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *arXiv preprint arXiv:2112.08359*, 2021. 3

Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020. 3

Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Generating person-scene interactions in 3d scenes. In *International Conference on 3D Vision (3DV)*, 2020a. 2, 3

Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b. 2, 3

Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *International Conference on Computer Vision (ICCV)*, 2021. 5, 7

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

7 Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The dataset contribution is illustrated in Sec. 3 and Tab. 1. The task contribution is shown in Fig. 4. The model contribution is analysed in Sec. 5 and Tab. 2.
   (b) Did you describe the limitations of your work? [Yes] Please refer to Sec. 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please refer to *Sup. Mat.*.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide the code, sample data, and trained checkpoint in the *Sup. Mat.*. This will reproduce the qualitative results from Fig. 4 and will roughly reproduce the quantitative result from Tab. 2. The full dataset is too large to include in the *Sup. Mat.*, but we will release the full dataset upon acceptance of this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please refer to Sec. 4.4 and Sec. 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We are aware of the requirements but cannot make it due to time limit. We include the code in *supp.* for reproduction and will add it in the later version.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please refer to Sec. 4.4

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] We mention it in the *supp.*
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We include the details of human evaluation in the *supp.*
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]