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

Zan Wang[1,2], **Yixin Chen**[2], **Tengyu Liu**[2]
Yixin Zhu[3]✉, **Wei Liang**[1,4]✉, **Siyuan Huang**[2]✉
✉ indicates corresponding authors
[1] School of Computer Science & Technology, Beijing Institute of Technology
[2] Beijing Institute for General Artificial Intelligence (BIGAI)
[3] Institute for Artificial Intelligence, Peking University
[4] Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing

https://silverster98.github.io/HUMANISE/

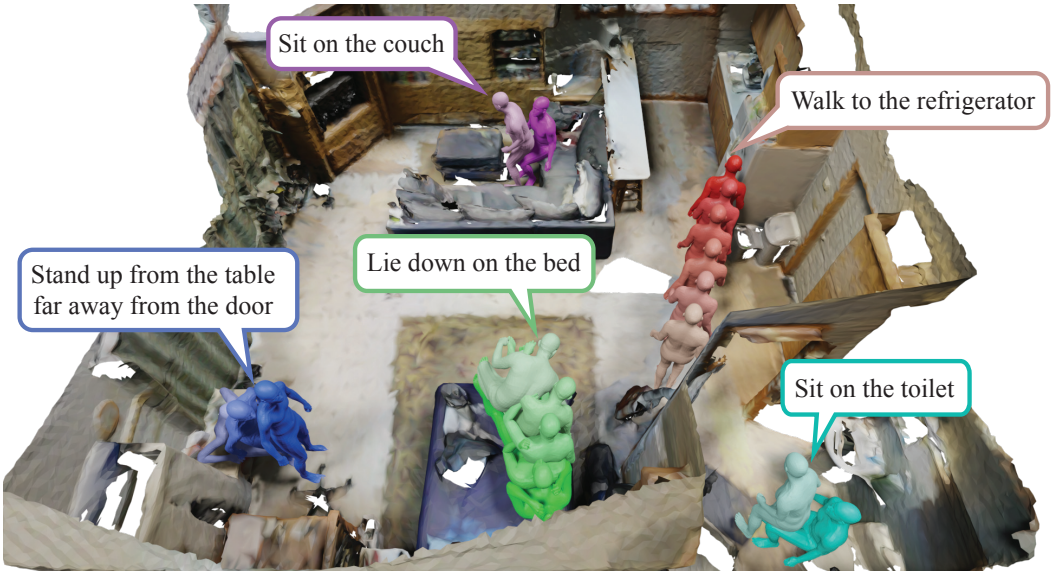


Figure 1: **Examples from our proposed Human-Scene Interaction (HSI) dataset—*HUMANISE*.** It contains 19.6k high-quality motion sequences in 643 indoor scenes. Each motion segment contains rich semantic information about the ground-truth action type and objects being interacted with, specified by language descriptions.

Abstract

Learning to generate diverse scene-aware and goal-oriented human motions in 3D scenes remains challenging due to the mediocre characteristics 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 unique interacting objects in the scene; *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.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1 Introduction

Imagine instructing an agent (*e.g.*, a virtual human or humanoid 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. Despite recent progress in human motion synthesis, generating 3D human motion conditioned on the instructions with proper scene context is still challenging.

Specifically, existing works generate plausible human poses [Zhang et al., 2020b;a] or motions [Starke et al., 2019; Cao et al., 2020; Wang et al., 2021b; Hassan et al., 2021a; Yi et al., 2022] by learning from existing HSI datasets (*e.g.*, PROX [Hassan et al., 2019], GTA-IM [Cao et al., 2020]). However, these datasets possess two fundamental issues that prevent existing methods from learning instruction-aware and generalizable 3D human motions.

- **Limited scale and quality**. PROX [Hassan et al., 2019] only contains 12 indoor scenes with jittering human motions. 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 interaction semantics**. PROX [Hassan et al., 2019] and GTA-IM [Cao et al., 2020] both suffer from incomplete 3D semantic segmentation annotations. Particularly, the lack of human action labels/descriptions hinders its ability to generate instruction-aware human motions.

To tackles the above issues, we propose a large-scale and semantic-rich synthetic HSI dataset, *HUMANISE* (see 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 agent and the scene are physically plausible and visually natural. Finally, 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, we adopt the object referential utterances in Sr3D [Achlioptas et al., 2020]. In total, *HUMANISE* contains 19.6k high-quality motion sequences in 643 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 propose 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 specified by the language descriptions. We argue that this task is more challenging than previous motion generation tasks in the following aspects: (i) The motion generation is conditioned on the multi-modal information including both the 3D scene and the language description. (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 *multi-modal understanding*, *language grounding*, and *physically plausible generation*.

Computationally, we leverage 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 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 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 tasks, further experiments demonstrate that *HUMANISE* alleviates the limits 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, human motions, and language. (iii) We develop a generative model that produces diverse and semantically consistent human motions conditioned on 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, algorithms [Zhang et al., 2020b;a; Hassan et al., 2021b;a; Wang et al., 2021a;b] can generate physically plausible single-frame human pose or long-term human motions. However, these methods either generate random/uncontrollable human motions or require target positions due to the lack of semantics constraints. Provided textual descriptions, existing motion generation algorithms [Lin and Amer, 2018; Ahuja and Morency, 2019; Ghosh et al., 2021] focus only on motions alone and neglect the scene context. In this work, we incorporate the context of both scenes and language descriptions, generating human motion sequences consistent with the specified action and location.

Human-Scene Interaction (HSI) datasets Data captured in both the physical and the virtual worlds have been adopted to study HSI related topics [Wang et al., 2019; Chen et al., 2019; Liang et al., 2019; Jia et al., 2020; Wang et al., 2020; Huang et al., 2022]. PiGraphs [Savva et al., 2016], captured with RGB-D sensors, suffers from inaccurately reconstructed scenes, noisy human poses, and a low frame rate. Even with new pipelines to reconstruct human motions from monocular RGB-D sequences, PROX's pose fitting results [Hassan et al., 2019] are still visibly worse compared to high-quality, large-scale MoCap data (*e.g.*, AMASS [Mahmood et al., 2019]). GTA-IM [Cao et al., 2020] exploits the game engine to collect a synthetic dataset of human skeletons, but with limited HSIs and scene diversities. In comparison, our *HUMANISE* dataset synthesizes large-scale HSIs by aligning high-quality motions in AMASS [Mahmood et al., 2019] with real-world 3D scenes in ScanNet [Dai et al., 2017], resulting in high-quality and realistic human-scene interactions with high diversity (*i.e.*, 643 scenes, $19.6k$ motions) and rich semantics provided by language descriptions.

Language grounding in 3D scenes Grounding 3D objects w.r.t. the language input has been explored in 3D reference [Achlioptas et al., 2020; Chen et al., 2021; Hong et al., 2021; Li et al., 2022; 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]. To date, locating the interacting objects in 3D scenes with referential descriptions is still challenging [Roh et al., 2022; Chen et al., 2020; 2021] due to the noisy scanned scenes, diverse expressions, and complex spatial relation reasoning. To better learn conditional embedding that facilitates locating and grounding according to the provided language descriptions with proper scene context, 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 grounding and action grounding capabilities by introducing two auxiliary tasks of locating interacting 3D objects and action comprehension.

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 synthesis pipeline can generalize to other 3D scene datasets and actions; please refer to Appendix C for more details.

Below, we start by describing how to align the motion with the 3D scene in Sec. 3.1, followed by automatic language description generation 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 an interacting target object by the predefined category labels 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.

Figure 2: **Samples from the *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).

Collision constraint We treat the point cloud of the 3D scene as a 3-dimensional k-d tree [Bentley, 1975]. 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 closest point in the scene is less than a threshold. A valid motion should avoid collisions with the objects or walls for all the poses along the trajectory.

Contact constraint We further decompose this constraint into the foot-ground support constraint and the proximity between the human and the interacting object. The support constraint is a distance threshold between the foot and the ground plane. The human-object proximity differs among actions; we design the following action-specific proximity constraints for realistic interaction:

- **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* provides instructions to walk to a target location, 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 sample motions representing diverse 3D affordance with the motion alignment algorithm, *e.g.*, *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 to the model building.

3.2 Language description generation

The motion descriptions in *HUMANISE* depict the action type and the objects/locations being interacted with. To automatically generate these language descriptions, we design a compositional

template following Sr3D [Achlioptas et al., 2020]:

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

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

HUMANISE includes 19.6k human motion sequences in 643 3D scenes. The four most representative actions are sit (5578), stand up (3463), lie down (2343), and walk (8264). Tab. 1 compares *HUMANISE* with existing HSI datasets; *HUMANISE* shows advantages in scene variety, clip number, frame number, and human pose quality. Of note, *HUMANISE* is the only dataset with rich semantic information of HSIs. Fig. 2 shows more examples from *HUMANISE*. Please refer to Appendices A and B for more statistics and quantitative comparison between *HUMANISE* and PROX, respectively.

Table 1: **Comparison between *HUMANISE* 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	643	19.6k	1.2M	Shape	✗	✓

4 Language-conditioned human motion generation in 3D scenes

4.1 Problem definition and notations

Our goal is to generate a human motion sequence that is both semantically consistent with the language description and physically plausible in interacting with 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 using 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, $r \in \mathbb{R}^6$ a continuous representation [Zhou et al., 2019] of the global orientation, $\beta \in \mathbb{R}^{10}$ the body shape parameters, $p \in \mathbb{R}^{J\times 3}$ J joint rotations in axis-angle, and F a differentiable linear blend skinning function. To model the effect of body shape in motion generation, we treat β as an additional condition 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. As a result, both the point-level and the word-level 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.

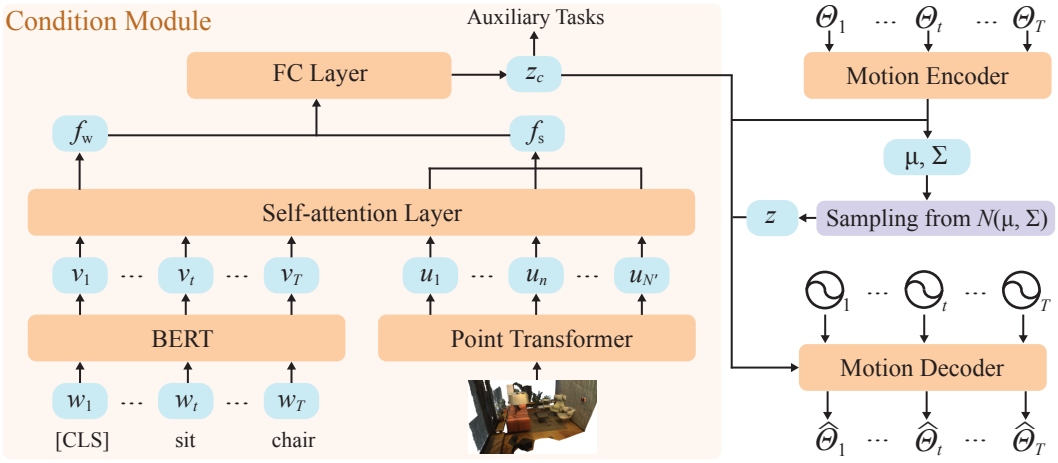


Figure 3: **Illustration of the proposed generative model.** It adopts the cVAE framework and incorporates a motion encoder and decoder to generate human motions. The condition is learned from the joint embedding of the 3D scene and language description with auxiliary tasks of 3D location grounding and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

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 $\widehat{\Theta}_{1:T}$ for 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 differentiable SMPL-X model. Of note, although the duration T equals the length of the input motion sequence in the training phase, it can be arbitrary values during inference.

4.3 Training loss

We devise the training loss with three components: reconstruction loss, KL divergence loss, and two auxiliary losses for object grounding and action-specific generation. Formally, the training loss is:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha_{kl}\mathcal{L}_{kl} + \alpha_o\mathcal{L}_o + \alpha_a\mathcal{L}_a, \tag{1}$$

where α denotes the loss weight of each term. We detail each component below.

Reconstruction loss Our reconstruction loss is formulated as

$$\mathcal{L}_{rec} = \mathcal{L}_t + \alpha_r\mathcal{L}_r + \alpha_p\mathcal{L}_p + \alpha_v\mathcal{L}_v, \tag{2}$$

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 the 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 Learning 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. 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 an 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 $\mathcal{L}_o$ in the target object center regression and the cross-entropy loss term $\mathcal{L}_a$ in the action category classification.

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 bidirectional 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 32 on a V100 GPU. It takes about 50 hours to converge when training the action-agnostic model on the entire *HUMANISE* dataset.

Point transformer Our point transformer module that processes the scene point cloud adopts from 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*.

BERT We employ the official pre-trained BERT [Kenton and Toutanova, 2019] model to process the language descriptions with a maximum length of 32. We map the obtained 768-D word-level features into 512-D before concatenating them with the point features.

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.

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 entire dataset with auxiliary tasks.

Dataset splitting We split motions in *HUMANISE* according to the original scene IDs and split in ScanNet [Dai et al., 2017], resulting in 16.5k motions in 543 scenes for training and 3.1k motions in 100 scenes for testing.

Ablative baselines We compare our model with three types of ablative baselines to verify the significance of the auxiliary tasks, the feature fusing module, and the scene encoder, respectively:

- We remove the auxiliary tasks for locating the target interacting object and recognizing the action category, denoted as *w/o* $\mathcal{L}_o$ (only removing $\mathcal{L}_o$), *w/o* $\mathcal{L}_a$ (only removing $\mathcal{L}_a$), and *w/o aux. loss* (removing both $\mathcal{L}_o$ and $\mathcal{L}_a$). We train these models on the entire *HUMANISE* dataset.
- 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*. We train and test on the *walk* subset.
- 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 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 over the sequence due to different motion duration.

Generation metrics As we hope our generated motion interacts 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

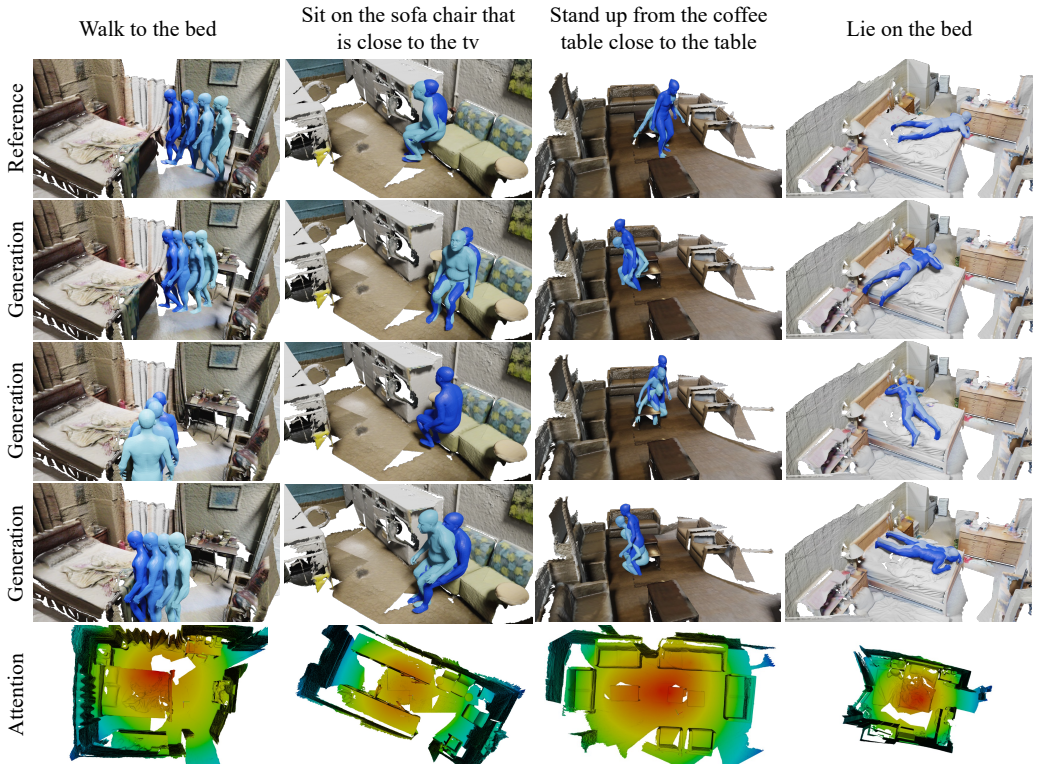


Figure 4: **Qualitative generation results of action-specific models on the *HUMANISE* dataset.** We visualize one reference motion and three generation motions for each scenario. The attention map visualizes the attention weights the [*CLS*] token attended to the scene point cloud in the self-attention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

$\mathcal{O} \subset S$ the target object point set. Of note, we use the human body in the first frame to compute 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. To measure the diversity of the generated motion, we report the Average Pairwise Distance (APD) [Yuan and Kitani, 2020; Zhang et al., 2021]: The average ℓ_2 distance between all pairs of motion samples within scenarios. APD is computed as $\frac{1}{K(K-1)T} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \sum_{t=1}^{T} \|\mathbf{x_{i,t}} - \mathbf{x_{j,t}}\|$, where K is sample size (we set $K = 20$), and $\mathbf{x_{i,t}}$ the sampled marker-based representation [Zhang et al., 2021].

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 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 *goal dist.* in *lie down* is much smaller than other actions. We suppose the model finds it easier to ground the target object in the *lie down* action, as the interacting objects are mostly large furniture with a flat surface such as "bed" and "table." We provide more clarifications with ablative experiments on the grounding loss weight and data efficiency in Appendix F.

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

Model	Reconstruction					Generation			
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	APD↑	quality score↑	action score↑
sit	5.17	3.19	1.77	113.28	112.43	0.903	10.12	2.37±0.85	3.79±1.17
stand up	5.63	3.43	1.69	126.05	124.84	0.802	9.57	2.83±1.23	4.20±0.94
lie down	6.46	3.09	0.76	136.87	136.20	0.196	9.18	2.31±1.08	2.85±1.31
walk	5.84	2.80	1.85	125.05	123.88	1.370	12.83	2.91±1.27	3.88±1.26
w/o self-att.	5.72	2.65	1.85	122.19	120.81	1.500	13.28	2.88±1.14	3.80±1.09
PointNet++ Enc.	5.81	2.64	1.81	124.67	123.69	1.444	12.61	2.80±1.35	3.75±1.27
all actions	4.20	2.91	1.96	98.01	96.53	1.008	11.83	2.57±1.20	3.59±1.38
w/o $\mathcal{L}_o$	4.20	2.89	1.93	98.15	96.69	1.383	15.09	2.42±1.21	3.57±1.38
w/o $\mathcal{L}_a$	4.23	2.91	1.95	98.67	97.11	1.135	12.66	2.17±1.04	2.29±1.43
w/o aux. loss	4.28	2.99	1.92	99.30	97.80	1.361	15.18	1.97±0.98	2.44±1.38

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

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

- 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. As shown in Tab. 2, removing $\mathcal{L}_o$ leads to a higher *goal dist.*, which means $\mathcal{L}_o$ can help our generative model to locate the target 3D object. $\mathcal{L}_a$ will help the model generate motions with the correct action label as can be seen from the *action score*. The significant differences in *goal dist.*, *quality score*, and *action score* indicate that the two proposed auxiliary tasks help provide better generation results while maintaining the reconstruction quality. Qualitative results in Fig. 5 further verify 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 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.

Discussion of failure cases We provide some typical examples of failure cases in Appendix E.

Our model sometimes fails to ground the correct object for interaction as locating objects in the 3D scene with referential descriptions is still challenging [Roh et al., 2022; Chen et al., 2020]. This calls for more efforts in joint learning of different tasks and modalities in language modeling, scene understanding, and motion generation.

Another typical failure case is the incorrect HOI relation with physical implausibility and collision. We suspect this is primarily due to the lack of explicit HOI modeling. Our current architecture is an end-to-end generative model; the scene information is represented by the point-level and scene-level features rather than object-level geometry and semantics. These point out that goal-oriented motion generation may benefit from future work on 4D HOI modeling and object affordance understanding.

Qualitative results of different duration T Following Petrovich et al. [2021], we use a transformer decoder as the motion decoder to generate a sequence of body parameters $\widehat{\Theta}_{1:T}$ in a given duration T. Fig. 6 shows some generated motions of different duration (*i.e.*, 30 frames, 60 frames, 90 frames, and 120 frames). We render a human body every 15 frames for fair comparisons.

Please refer to Appendix F for additional experiments and Appendix D for more qualitative results.

5.4 *HUMANISE* for downstream tasks

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. More specifically, this task takes a scene point cloud, the start and the end locations, and the start and

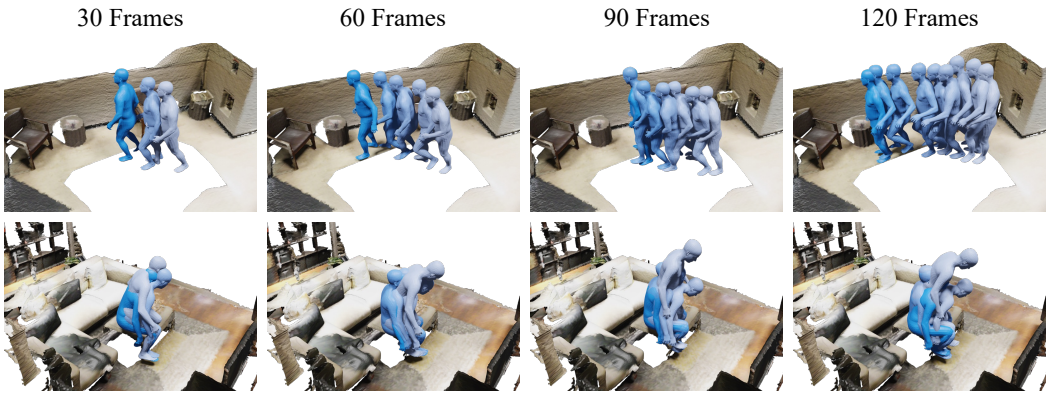


Figure 6: **Motions generated by sampling with different duration** T. The language descriptions in these two cases are *walk to the end table that is farthest from the door* and *sit on the coffee table*. Our model is capable of generating motions with various duration.

the end human body poses as input; the goal is to predict all human poses in between. For model architecture, we follow Wang et al. [2021a] to include a RouteNet and a PoseNet. The RouteNet takes the scene point cloud, the start and the end locations and orientations as inputs; it generates a possible route between the start and end location. The PoseNet takes the same scene point cloud, the start and the end human body poses, and the route predicted by RouteNet as inputs; it outputs the parameters for all the human poses along the route.

We train the RouteNet and PoseNet using *HUMANISE* for ten epochs and fine-tune them using PROX [Hassan et al., 2019] for 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. This result 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	**6.54**	**8.80**	**42.54**	**222.65**	**201.92**

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 scenes*. 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 to 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 still 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.

Acknowledgments and disclosure of funding We sincerely thank Chao Xu and Nan Jiang for their valuable advice on this project. We would like to thank the anonymous reviewers for their constructive comments. Z. Wang, Y. Chen, T. Liu, and S. Huang were supported in part by the National Key R&D Program of China (2021ZD0150200); Z. Wang and W. Liang were supported in part by the National Natural Science Foundation of China (NSFC) (No.62172043).

References

Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., and Guibas, L. (2020). Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*. 2, 3, 5, 10, 19

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

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Azuma, D., Miyanishi, T., Kurita, S., and Kawanabe, M. (2021). Scanqa: 3d question answering for spatial scene understanding. *arXiv preprint arXiv:2112.10482*. 3

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

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517. 4

Cao, Z., Gao, H., Mangalam, K., Cai, Q.-Z., Vo, M., and Malik, J. (2020). Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*. 2, 3, 5

Chao, Y.-W., Yang, J., Chen, W., and Deng, J. (2019). Learning to sit: Synthesizing human-chair interactions via hierarchical control. *arXiv preprint arXiv:1908.07423*. 3

Chen, D. Z., Chang, A. X., and Nießner, M. (2020). Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*. 3, 9, 19

Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., and Zhu, S.-C. (2019). Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*. 3

Chen, Y., Li, Q., Kong, D., Kei, Y. L., Gao, T., Zhu, Y., and Huang, S. (2021). Yourefit: Embodied reference understanding with language and gesture. In *International Conference on Computer Vision (ICCV)*. 3

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3, 7, 19

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., and Batra, D. (2018). Embodied question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Gan, C., Zhang, Y., Wu, J., Gong, B., and Tenenbaum, J. B. (2020). Look, listen, and act: Towards audio-visual embodied navigation. In *International Conference on Robotics and Automation (ICRA)*. 3

Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., and Slusallek, P. (2021). Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*. 3

Gordon, D., Kadian, A., Parikh, D., Hoffman, J., and Batra, D. (2019). Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., and Cheng, L. (2020). Action2motion: Conditioned generation of 3d human motions. In *International Conference on Multimedia*. 3

Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., and Black, M. J. (2021a). Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*. 2, 3

Hassan, M., Choutas, V., Tzionas, D., and Black, M. J. (2019). Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*. 2, 3, 5, 10, 15

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

Hong, Y., Li, Q., Zhu, S.-C., and Huang, S. (2021). Vlgrammar: Grounded grammar induction of vision and language. In *International Conference on Computer Vision (ICCV)*. 3

Huang, C.-H. P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., and Black, M. J. (2022). Capturing and inferring dense full-body human-scene contact. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2013). 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. 7

Jia, B., Chen, Y., Huang, S., Zhu, Y., and Zhu, S.-c. (2020). Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision (ECCV)*. 3

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). 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)*. 5, 7

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 7

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 6

Li, C., Zhang, Z., Lee, W. S., and Lee, G. H. (2018). Convolutional sequence to sequence model for human dynamics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Li, Y., Chen, X., Zhao, H., Gong, J., Zhou, G., Rossano, F., and Zhu, Y. (2022). Understanding embodied reference with touch-line transformer. *arXiv preprint arXiv:2210.0566*. 3

Liang, W., Liu, J., Lang, Y., Ning, B., and Yu, L.-F. (2019). Functional workspace optimization via learning personal preferences from virtual experiences. *IEEE Transactions on Visualization and Computer Graph (TVCG)*, 25(5):1836–1845. 3

Lin, X. and Amer, M. R. (2018). Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*. 3

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*. 2, 3, 19

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

Petrovich, M., Black, M. J., and Varol, G. (2021). Action-conditioned 3d human motion synthesis with transformer vae. In *International Conference on Computer Vision (ICCV)*. 3, 6, 9

Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J. (2021). Babel: Bodies, action and behavior with english labels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3, 5, 19

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NIPS)*. 7

Roh, J., Desingh, K., Farhadi, A., and Fox, D. (2022). Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning (CoRL)*. 3, 9

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

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

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

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., et al. (2019). The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*. 16

Taheri, O., Ghorbani, N., Black, M. J., and Tzionas, D. (2020). Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*. 3

Thomason, J., Shridhar, M., Bisk, Y., Paxton, C., and Zettlemoyer, L. (2022). Language grounding with 3d objects. In *Conference on Robot Learning (CoRL)*. 3, 10

Wang, H., Liang, W., and Yu, L.-F. (2020). Scene mover: automatic move planning for scene arrangement by deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 39(6):1–15. 3

Wang, J., Xu, H., Xu, J., Liu, S., and Wang, X. (2021a). Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3, 7, 9, 10

Wang, J., Yan, S., Dai, B., and Lin, D. (2021b). Scene-aware generative network for human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3

Wang, Z., Chen, L., Rathore, S., Shin, D., and Fowlkes, C. (2019). Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*. 3

Xie, K., Wang, T., Iqbal, U., Guo, Y., Fidler, S., and Shkurti, F. (2021). Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*. 3

Ye, S., Chen, D., Han, S., and Liao, J. (2021). 3d question answering. *arXiv preprint arXiv:2112.08359*. 3

Yi, H., Huang, C.-H. P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., and Black, M. J. (2022). Human-aware object placement for visual environment reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Yuan, Y. and Kitani, K. (2020). Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*. 3, 8

Zhang, S., Zhang, Y., Ma, Q., Black, M. J., and Tang, S. (2020a). Generating person-scene interactions in 3d scenes. In *International Conference on 3D Vision (3DV)*. 2, 3

Zhang, Y., Black, M. J., and Tang, S. (2021). We are more than our joints: Predicting how 3d bodies move. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 6, 8

Zhang, Y., Hassan, M., Neumann, H., Black, M. J., and Tang, S. (2020b). Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3

Zhao, H., Jiang, L., Jia, J., Torr, P. H., and Koltun, V. (2021). Point transformer. In *International Conference on Computer Vision (ICCV)*. 5, 7

Zhou, Y., Barnes, C., Lu, J., Yang, J., and Li, H. (2019). On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 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 the Appendix G.3.
   (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 have released the code, data, and trained checkpoint in the project page. This will reproduce the qualitative results from Fig. 4 and will reproduce the quantitative result from Tab. 2.
   (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 have released the code and data for reproduction.
   (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 Appendix G.1.
   (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 Appendix G.2.
   (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]

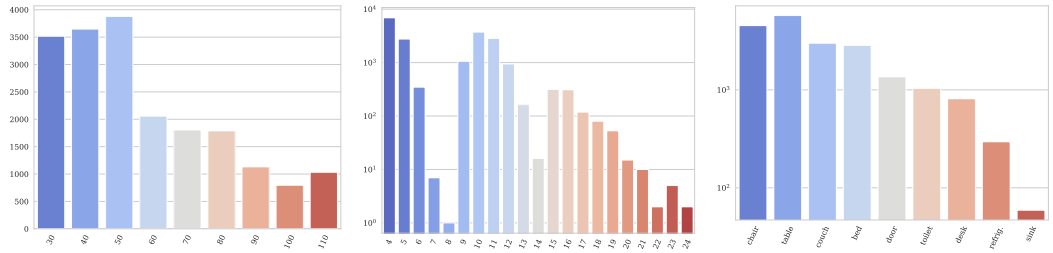
Appendix

We present additional statistics of *HUMANISE* and detailed quantitative comparisons between *HUMANISE* and PROX in Appendices A and B, respectively. In Appendix C, we show that our motion alignment pipeline can be easily generalized to various 3D scene datasets and actions, resulting in a vast amount of synthetic data to be used for HSI tasks. Appendix D provides additional qualitative results for both reconstruction and generation, and Appendix E presents some examples of failure cases. In Appendix F, we describe additional ablative and proof-of-concept experiments. Appendix G includes the checklist required by the *NeurIPS* submission.

A Statistics of *HUMANISE*

We report additional statistics of the *HUMANISE* dataset. Fig. A1a shows the distribution of frame length. *HUMANISE* contains motions ranging from 30 to 120 frames. Fig. A1b shows the sentence lengths distribution of the language description. The average length of language descriptions in *HUMANISE* is 7.63. Fig. A1c shows the frequencies of the 9 interactive object categories. Since we focus on indoor scene activities, the most frequently interactive objects are "chair," "table," and "couch."

We also report the motion diversity when synthesizing the *HUMANISE* dataset. 1305 different motions from AMASS are selected to synthesize 19.6k motion sequences in 643 3D scenes. Among all the synthesized motion sequences 19.6k [1305], the number of the synthesized motion sequences and the number of selected motions from AMASS for each action are: 8264 [717] for walk, 5578[365] for sit, 3463 [190] for stand-up, and 2343 [33] for lie-down.



(a) Frame length distribution. (b) Sentence length distribution (c) Interactive object frequency.

Figure A1: **Statistics of the *HUMANISE* dataset.**

B Quantitative comparison between *HUMANISE* and PROX

In this section, we present quantitative metrics to compare the dataset quality between *HUMANISE* and PROX [Hassan et al., 2019]. We first introduce a collision distance to evaluate the quality of physical implausibility between the human body and the 3D scene. More specifically, it is calculated as the average distance (in meters) between the human body surface and the scene points that penetrate the human body. We also perform human perceptual studies to evaluate the dataset quality in terms of the overall quality, motion smoothness, and quality of human-scene interaction. A worker is asked to score from 1 to 5 for (i) the overall *quality*, (ii) the *smoothness* of the motion, and (iii) whether the human bodies have physically plausible *HSI* with the 3D scenes. A higher value indicates higher quality. We randomly select 100 motion segments each from PROX and *HUMANISE*, and three workers score each sample. As can be seen from Tab. A1, motions in *HUMANISE* have higher quality in terms of collision, smoothness, HSI and overall quality, which further validates the significance of our proposed dataset in HSI research.

Table A1: **Comparison between HUAMNISE and existing HSI datasets.**

Dataset	Collision Distance↓	Smoothness Score↑	HSI Score↑	Quality Score↑
PROX-Q [Hassan et al., 2019]	0.024	3.13±1.29	3.54±1.35	3.32±1.20
HUMANISE	0.012	4.49±0.84	4.21±1.01	4.14±0.93

C Generalization of motion alignment pipeline

Our dataset synthesis pipeline can generalize to other 3D scene datasets and more action types, which means we can generate a huge amount of synthetic data by aligning the motions with 3D scenes. Below we show some qualitative examples.

Generalize to other 3D scene datasets We employ our motion alignment pipeline to generate human motions in completely hand-craft 3D scenes, *i.e.*, Replica [Straub et al., 2019]. Some aligned results are shown in Fig. A2.

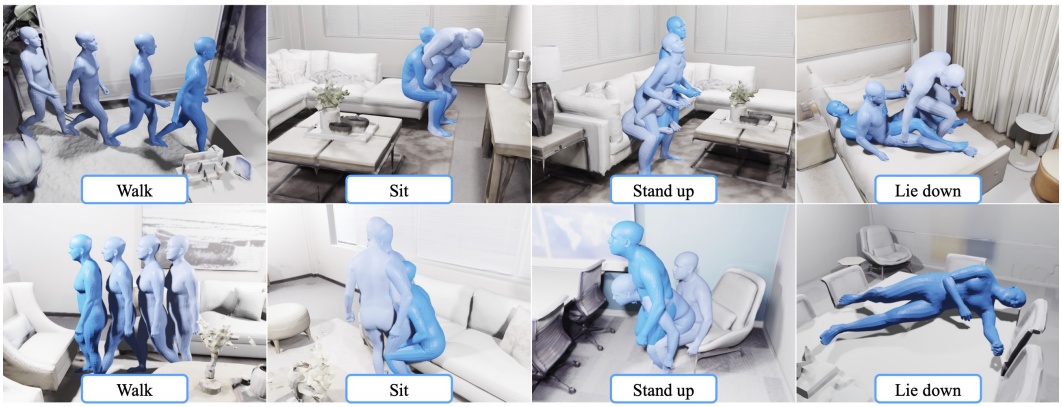


Figure A2: **Motions aligned with scenes in Replica [Straub et al., 2019].**

Generalize to other actions We additionally align motions of extra action types (*i.e.*, *jump up*, *turn to*, *open*, and *place something*) with scanned scenes. Some selected results are shown in Fig. A3.

Figure A3: **Examples of motions with action types *jump up*, *turn to*, *open*, and *place something*.**

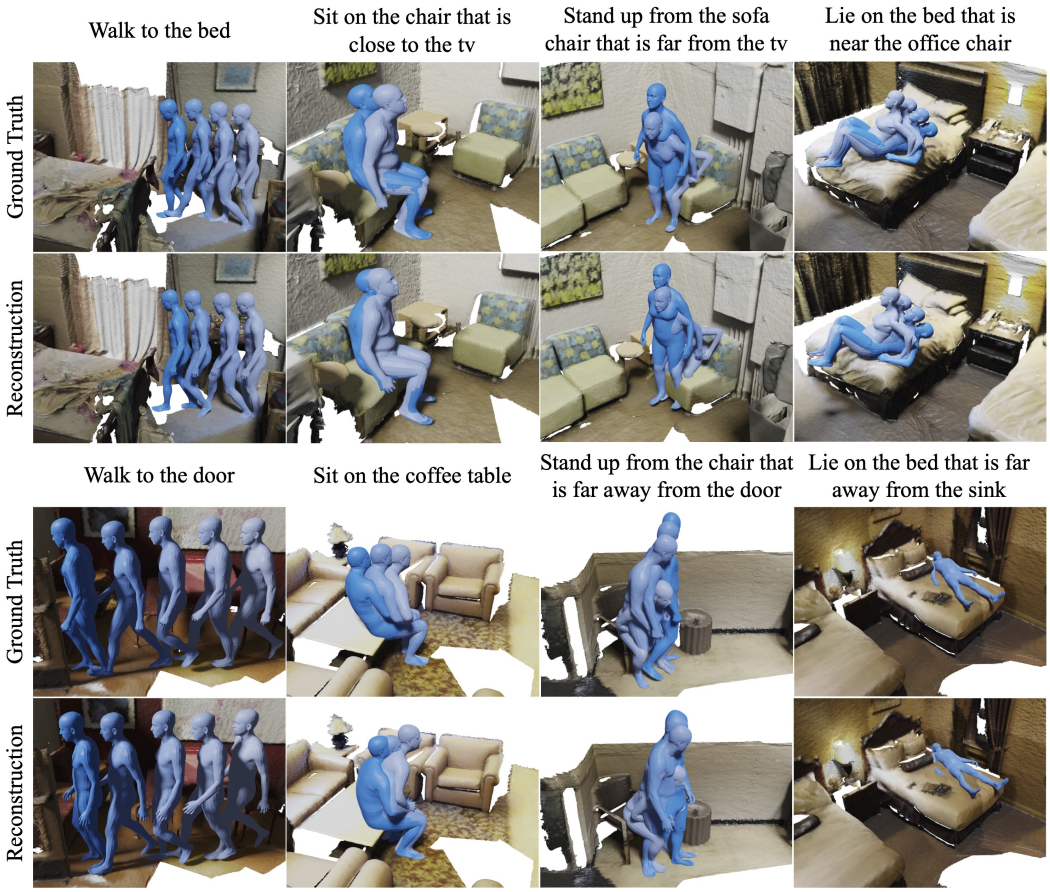


Figure A4: **Qualitative reconstruction results of the action-specific models.**

D Additional qualitative results

Qualitative results of reconstruction More qualitative reconstruction results are shown in Fig. A4.

More qualitative results of generation More qualitative generation results are shown in Fig. A5.

E Typical examples of failure cases

Fig. A6 and Fig. A7 shows that our model sometimes fails to ground the correct target object for interaction and fails to generate correct HOI relation.

F Additional ablations

F.1 Ablations of the *lie-down* action

To diagnose why the *lie down* action has smaller *goal dist.* in the action-specific setting, we conduct additional ablative experiments by using different grounding loss weight, *i.e.*, α_o, and varying the amount of data during training. Specifically, we try $\alpha_o = 0.01$, $\alpha_o = 0.001$, and $\alpha_o = 0$, while the original $\alpha_o = 0.1$. We additionally use 10% and 50% of the original dataset to train the action-specific model of *lie down* action. The quantitative results are shown in Tab. A2.

From the table, we can find that (1) the *goal dist.* increases as α_o decreases, but the values are still smaller than other actions (see Tab. 2). (2) the models trained with fewer data reach approximately the same performance in *goal dist.* as the model trained with all *lie down* data, but the reconstruction metrics drop significantly. These two observations indicate that the grounding task for *lie down* action is easier than other actions. We hypothesize this is because most *lie down* actions interact with *bed* or *table*, which has a flat surface and occupies a much larger area compared to other types of furniture.

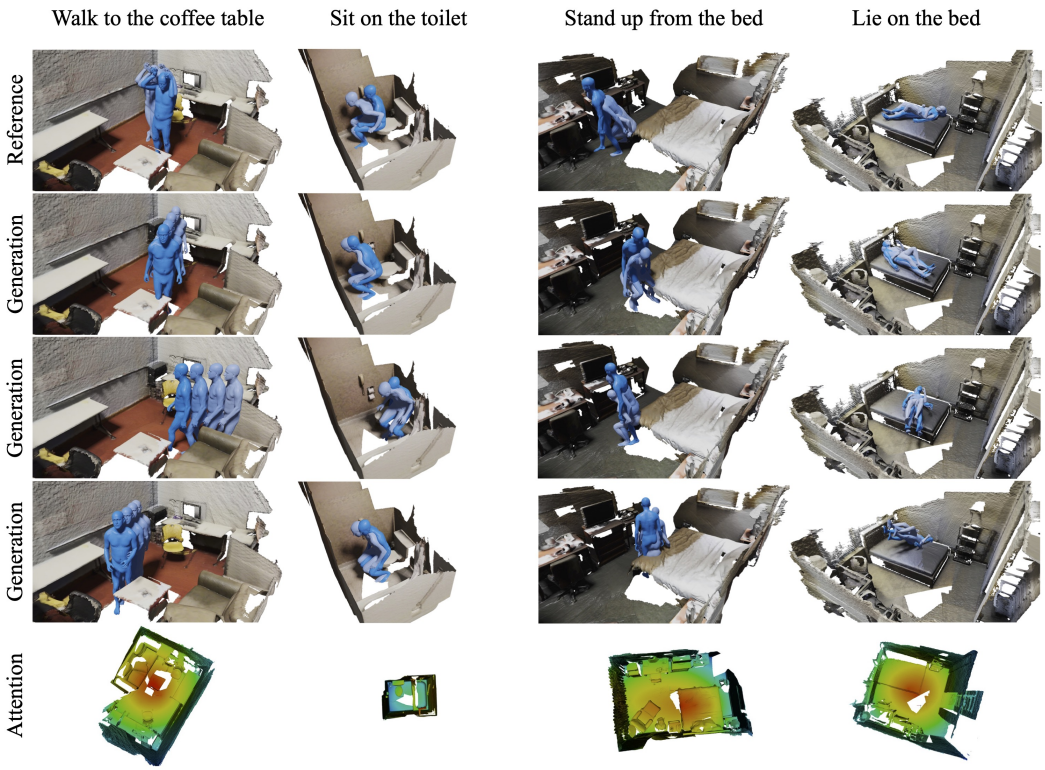
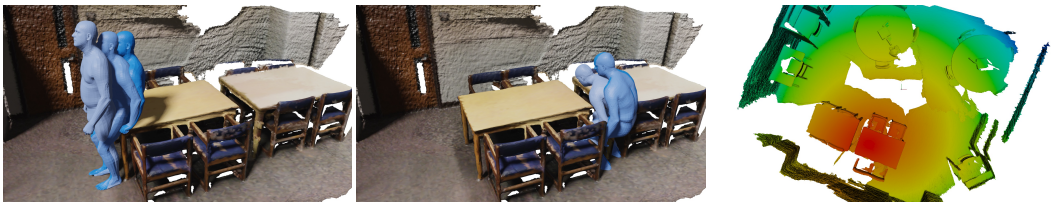


Figure A5: **Qualitative generation results of the action-specific models.**

(a) Reference (b) Failure Sample (c) Attention

Figure A6: **Failure case with incorrect grounding.** The language description, in this case, is *walk to the table that is in the middle of the box and the trash can*.



(a) Reference (b) Failure Sample (c) Attention

Figure A7: **Failure case with incorrect HOI relation .** The language description, in this case, is *sit on the table that is close to the door*.

Table A2: Ablations about *lie down* action.

Model	Reconstruction					Generation	
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	APD↑
$\alpha_o = 0$	6.67	3.06	0.79	142.94	142.08	0.444	11.72
$\alpha_o = 0.001$	6.76	2.78	0.82	142.30	141.77	0.306	10.97
$\alpha_o = 0.01$	6.77	3.31	0.77	144.15	143.43	0.211	9.50
50%	8.04	4.24	0.96	170.50	169.96	0.177	8.56
10%	10.44	9.15	5.27	248.82	243.99	0.207	8.05
Ours	6.46	3.09	0.76	136.87	136.20	0.196	9.18

F.2 Multi-stage pipeline and end-to-end pipeline

As an initial attempt to explore the difference between an end-to-end approach and a cascaded approach for our task, we conduct experiments to compare with baselines that directly use action or target object as conditions. More specifically, We modify our cVAE model by replacing the condition with the concatenation of the global scene feature, the target object center, and the action category. We use the point cloud feature extracted from PointNet++ as the global scene feature, and we adopt the one-hot embedding to represent the GT action categories. For the target object center, we use a 3D grounding model pre-trained on ScanNet, *i.e.*, ScanRefer [Chen et al., 2020], to estimate the target object center. We denote this baseline as $\mathrm{GT}_{\mathrm{action}}$. We also test another baseline that directly uses the ground truth position instead of the predicted target object position. We denote this baseline as $\mathrm{GT}_{\mathrm{action+target}}$. The quantitative results are shown in Tab. A3.

From the table, we can see that the baseline that directly uses GT action-conditioning reaches approximately the same performance as our model while utilizing the GT action and the GT target position can significantly improve the motion generation metrics. We hypothesize this is because the action can be easily parsed from the instruction and 3D object grounding is significantly more challenging than other subtasks, which affects the performance most. The results also justify the multi-stage pipeline could achieve similar performance as the end-to-end pipeline in the current setting.

Table A3: Comparison between multi-stage pipeline and end-to-end pipeline.

Model	Reconstruction					Generation			
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	APD↑	quality score↑	action score↑
$\mathrm{GT}_{\mathrm{action}}$	4.23	2.90	1.90	98.67	97.17	1.406	15.98	2.56±1.02	3.93±1.11
$\mathrm{GT}_{\mathrm{action+target}}$	4.13	2.84	1.92	96.02	94.57	0.207	9.37	2.78±1.33	3.86±1.31
Ours	4.20	2.91	1.96	98.01	96.53	1.008	11.83	2.57±1.20	3.59±1.38

G CheckList

G.1 Use of existing assets

Our dataset *HUMANISE* is build on data from AMASS [Mahmood et al., 2019], ScanNet [Dai et al., 2017], BABEL [Punnakkal et al., 2021] and Sr3D [Achlioptas et al., 2020]. AMASS is licensed under the terms of the Dataset Copyright License for non-commercial scientific research purposes. The ScanNet dataset is released under the ScanNet Terms of Use, and the code is licensed under the terms of the MIT license. BABEL is licensed under the terms of the Software Copyright License for non-commercial scientific research purposes. Sr3d is licensed under the terms of the MIT license.

G.2 Human study instructions

Here we include the instructions given to participants of the human perceptual studies for evaluating the generation results in *HUMANISE*.

instruction You are required to score the given human motion in the scene from the following two aspects: (i) overall quality and (ii) action consistency. For overall quality, you need to consider all the factors that may influence the quality of human motion, including the naturalness, smoothness, physical plausibility, and consistency with the corresponding language descriptions, *etc*. Score the motion from 1 to 5. A higher score means the motion has a higher quality. For action consistency, you should only consider if the action of the motion is consistent with the given language description. Score the motion from 1 to 5. A higher score means the action of the motion is more consistent with the language description.

G.3 Societal impacts

We firmly believe that our work will promote the understanding of HSI, which facilitates real-world applications including VR/AR, human avatars, video game animation, assistant robots, and metaverse. On the other hand, scene and HSI understanding can be extended to activity tracking, which might raise privacy issues to some extent.