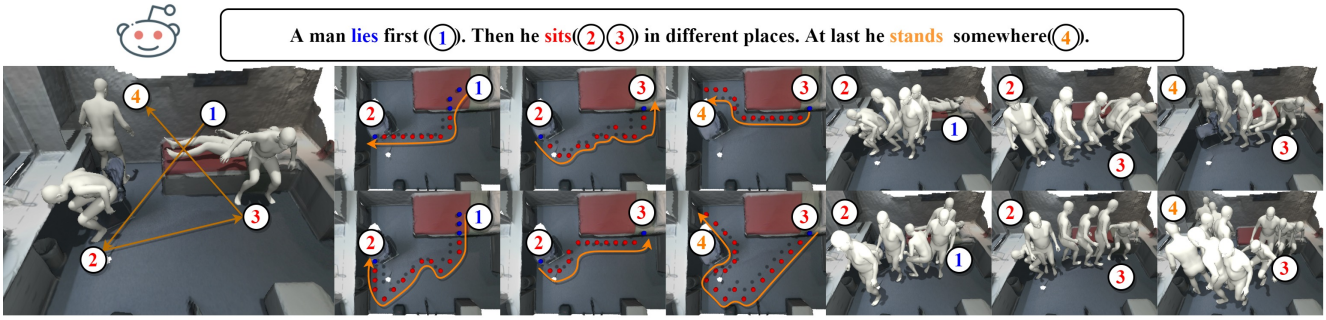
Towards Diverse and Natural Scene-aware 3D Human Motion Synthesis

Jingbo Wang[1] Yu Rong[1] Jingyuan Liu[2] Sijie Yan[1] Dahua Lin[1] Bo Dai[3]
[1] The Chinese University of Hong Kong [2] Hong Kong University of Science and Technology
[3] S-Lab, Nanyang Technology University
{wj020,ry017,dhlin}@ie.cuhk.edu.hk, jliucb@connect.ust.hk, yysijie@gmail.com, bo.dai@ntu.edu.sg



(a) Human Scene Interaction Anchors (b) Diverse Planning for Interaction Anchors (c) Diverse Motion Synthesis

Figure 1. We decompose scene-aware human motions into three aspects, namely human-scene interaction anchor, path planning, and body movements. Given the scene and the target action sequence, our framework first adopts specified schemes to generate diverse intermediate results for each aspect. These results are then integrated into miscellaneous yet coherent human motions.

Abstract

*The ability to synthesize long-term human motion sequences in real-world scenes can facilitate numerous applications. Previous approaches for scene-aware motion synthesis are constrained by pre-defined target objects or positions and thus limit the diversity of human-scene interactions for synthesized motions. In this paper, we focus on the problem of synthesizing diverse scene-aware human motions under the guidance of target action sequences. To achieve this, we first decompose the diversity of scene-aware human motions into three aspects, namely interaction diversity (*e.g. *sitting on different objects with different poses in the given scenes), path diversity (*e.g. *moving to the target locations following different paths), and the motion diversity (*e.g. *having various body movements during moving). Based on this factorized scheme, a hierarchical framework is proposed, with each sub-module responsible for modeling one aspect. We assess the effectiveness of our framework on two challenging datasets for scene-aware human motion synthesis. The experiment results show that the proposed framework remarkably outperforms previous methods in terms of diversity and naturalness.*

1. Introduction

The capability of synthesizing long human motion sequences is essential for a number of real-world applications, such as virtual reality and robotics. Beyond early attempts that consider body movement synthesis in isolation [1, 2, 33, 35, 38], recent works [4, 11, 30, 31] begin to explore the influences of surrounding scenes on human motion synthesis for different actions. Limited by the 2D representation of scene context [4, 31] or the reliance on manually assigned interacting targets [11, 30], these approaches mainly focus on modeling the body movements and fail to comprehensively investigate the inherent diversity of scene-aware human motions. In order to synthesize long-term human motions guided by the scene context and the target action sequence, we propose to model the inherent motion diversity across different granularities, each contributing to different aspects of human motion.

As shown in Figure 1, the diversity of scene-aware human motions can be factorized into three levels, given the target action sequence (*e.g.* A man lies first. Then he sits in different places. At last, he stands somewhere.). Firstly, given the surrounding scene context and the target action sequence, there exists a distribution of valid locations to re-

alize the actual human-scene interactions for each of these actions (*e.g*. We can sit on any chairs or beds and stand on the ground). Different locations can be sampled from the distribution and serve as the anchors of the whole synthesized motion sequence. Based on those anchors, we can then follow various paths to bridge them one by one. Finally, our body poses also differ from case to case when we move along the paths to connect all anchors. We demonstrate these three levels of diversity in Figure 1. Existing attempts for scene-aware human motion synthesis [11, 30] only emphasize the last level of diversity (*e.g*. walking to the pre-defined object or position in the scene) via manually assigning the interaction locations and motion paths. Consequently, the importance of the scene semantics is substantially muted, as it mainly affects the distribution of valid interaction anchors and the distribution of valid motion paths. To faithfully capture the diversity of scene-aware human motions, we propose a novel three-stage motion synthesis framework, each stage of which is responsible for modeling one level of the aforementioned diversity.

For **diverse human-scene interaction anchors**, we design our pose placing framework for the given action sequence. Different from [37, 39] which only consider the influence of scene context, we first synthesize scene-agnostic poses according to the target action via a conditional VAE (CVAE) [25]. Then we follow the practice of POSA [13] to place these poses into the scene. To be specific, the 3D scene is uniformly split into a set of non-overlapping grids, each of which is associated with a validity score that measures its compatibility as a candidate for placing the poses. We make two modifications to the original placing method used by POSA First, we introduce the position relationship between poses with the same action label to enhance the placing diversity by avoiding them being placed to the nearby positions. Furthermore, we leverage another CVAE model as the placing refiner to produce diverse offsets for each discrete grid. Examples of generated anchors are depicted in Figure 1 (a).

To produce **diverse obstacle-free motion paths** following the sampled anchors, we employ an adapted A* algorithm over the discrete 3D grids as the path planner. The standard A* algorithm used by previous works [11] only generates deterministic paths as they only consider collision between objects and distances to the target locations. To model the inherent diversity of motion paths, we amend the original algorithm with a trainable stochastic module learned in a data-driven manner. The new module, named Neural Mapper, can provide dynamic scene-conditioned probabilistic guidance to the A* algorithm, so that the algorithm can automatically produce diverse yet natural paths given the deterministic scenes and location anchors. We show several examples of generated diverse paths given the same start and end locations in Figure 1 (b).

Lastly, we propose a novel Transformer-based CVAE, called motion completion network, to synthesize **diverse body movements** guided by the paths generated in the previous step. Inspired by [23], we leverage Transformer as the basic architecture for synthesizing continuous and smooth motions. Differently, we focus on diverse motion completion of poses with long-term distance and different actions, rather than synthesize motions for the single action [23]. Therefore, this motion completion network first generates diverse moving trajectories, loosely following the paths sampled by the aforementioned A* algorithm. The body poses are then produced by taking the scene contexts, action labels, human-scene interaction anchors, and synthesized trajectories as inputs.

To summarize our contributions: 1) We analyze the **inherent diversity** of the human motion and decompose it into three components, namely the diversity on human-scene interaction anchors, paths, and body poses. 2) We propose a novel three-stage framework to **faithfully capture the diversities** of scene-aware human motions. This framework can automatically synthesize human motions following these diversities with the condition action labels. Qualitative and quantitative results on datasets such as PROX [12]demonstrate that our method significantly surpasses previous approaches in terms of diversity and naturalness. 3) In the proposed framework, we make several technique contributions for this task, including the action conditioned pose placing framework for generating diverse human-scene interaction anchors, Neural Mapper for planning diverse paths, and motion completion network for producing diverse and continuous motions. With our decomposition on motion diversity, these technique contributions can achieve our goal efficiently and effectively.

2. Related Works

Motion Synthesis. Early works [1, 2, 10, 22, 32, 33, 35] focus on synthesizing natural body poses and neglect the influences of other factors such as action and environments. Recent studies begin to explore the relationship between human motions with actions and scene contexts. Recent works [3, 8] generates human pose sequences with a CVAE model [25] based on the given action labels. ACTOR [23] builds up a transformer based on CAVE to synthesize human motion sequence directly from the given action label. Cao *et al.* [4] propose a three-stage motion prediction method that can predict different human motions with different destinations. Wang *et al.* [31] extend CSGN [33] to explore the influence of 2D scene contexts on human motion synthesis. Wang *et al.* [30] build up a framework to synthesize human motions in the 3D scene controlled by the given pairs of begin-end points. SAMP [11] extends [26] to use 3D oriented objects to facilitate the synthesis of human motions with specific action labels. Besides, a plan-

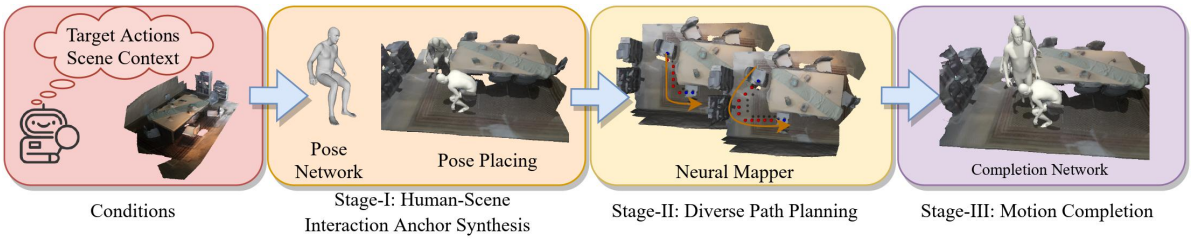


Figure 2. **Overview of the framework.** Our framework is composed of three stages. Given the target actions and the scene contexts, our framework first generates human-scene interaction anchors by firstly synthesizing scene-agnostic poses via the pose network and then placing the poses into the scene guided by the scene contexts and Pose Refiner. Then the framework produces diverse planning paths through the adapted A$*$ algorithm that is amended with the novel Neural Mapper. At last, the motion completion network is leveraged to synthesize natural human motions guided by the anchors and following the planned paths.

ning module is incorporated into their framework to find obstacle-free paths.

The limitations of previous works [11, 30] mainly lie in their reliance on predefined objects or positions, which constrain their ability to explore the inherent interaction diversity of synthesized scene-aware human motions. In this work, we aim to overcome the limitations of the previous works and synthesize diverse motions guided by target action sequences in the given scenes. To achieve this, we first synthesize diverse human-interaction anchors, which interacts with different objects in the scene. Then we plan diverse paths and complete diverse body movements between these anchors.

Motion Prediction. Motion prediction is closely related to our problem. Different from the motion synthesis, the goal of this task is to predict human dynamics in the future with the given moving orientations or previous motions. Martinez *et al.* [20] and ERD [7] proposed motion prediction framework based on the Seq2Seq model [27]. Ac-LSTM [18] mixes synthesized frames and observed frames to enhance the capability of LSTM [14] during the training stage. The graph convolution network [16, 34]is widely used in recent motion prediction [6, 17, 19]. These methods model dynamic spatial and temporal relationships between the obvious frames and the future frames. Different from these works, our goal is to synthesize motions without prior knowledge of the previous motions.

3. Methodology

3.1. Overview

We first formally define the task of scene-aware 3D human motion synthesis. We use triangular mesh $S = (v^s, f^s)$ to represent the scene context, where v^s and f^s stand for vertices and faces. Our task is to synthesize diverse 3D human motions in the given scene context S, driven by a sequence of target action labels $A = (a_1, a_2, ..., a_N)$. Each label stands for one scene-related human action, such as sitting or laying. The synthesized 3D human motions are represented as a sequence of SMPL-X models [21] described by their parameters $\{P_0, ..., P_T\}$, where P_i is composed of (t_i, ϕ_i, θ_i), $t_i \in \mathbb{R}^3$ is the global translation, $\phi_i \in \mathbb{R}^6$ is the global orientation represented in 6D continuous rotation [40]. $\theta_i \in \mathbb{R}^{32}$ is the body pose parameters, represented in the form of VPoser [21]. We use mean values for remaining SMPL-X parameters, including shape parameters, facial parameters, and hand poses.

The overview of our framework is depicted in Figure 2. We aim to solve this challenging problem in a hierarchical manner via exploiting the inherent properties of the scene-aware human motions. Our framework first generates diverse human-scene interaction anchors for the given actions. In this step, the framework first produces scene-agnostic poses corresponding to the action labels and then places these poses into the scene considering the compatibility between the synthesized poses and the scene. In the next step, we leverage a path planning module to produce diverse obstacle-free paths under the guidance of the synthesized anchors from the first step. Finally, a motion completion module is adopted to synthesize diverse body movements that fill in the missing motions between consecutive anchors while roughly following the planned paths from the second step. In the following, we introduce our modules in detail.

3.2. Human-Scene Interaction Anchor Synthesis

We first synthesize human-scene interaction anchors. Unlike previous works [37, 39] that only condition human motion synthesis on the scene context, we use action labels describing interaction types as an additional condition. To be specific, we first synthesize scene-agnostic poses corresponding to the action labels. Then we follow the practice of POSA [13] with several modifications to diversely place the synthesized poses into the scene. This design affords us more control over the final synthesized motions.

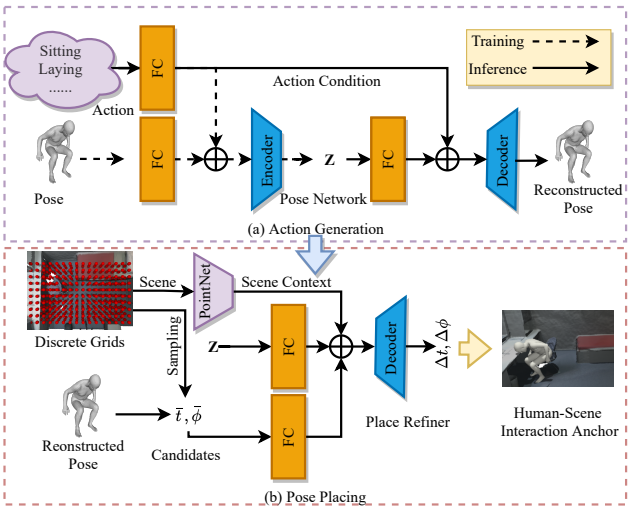


Figure 3. **Human-scene interaction anchor generation.** There are two steps to synthesize human-scene interaction anchors. The first step is generating diverse scene-agnostic poses conditioned on the target action, as shown in (a). The second step is placing synthesized poses in the given scene, as depicted in (b).

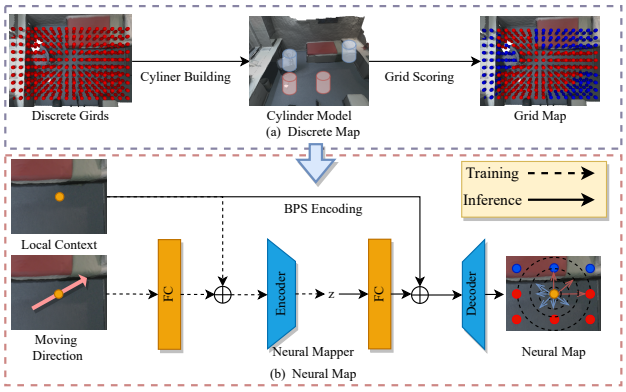


Figure 4. **Map building.** The map for our planning algorithm is built based on the collision detection (a) and Neural Mapper (b). Neural Mapper provides diverse moving probability for each neighbor gird to plan diverse paths.

Scene-Agnostic Pose Synthesis. As shown in Figure 3 (a), we follow the standard CVAE framework to synthesize scene-agnostic poses θ_i with the target action a_i. To be specific, we first sample noises from the prior Gaussian distribution and encode them with a fully-connected layer. Then, we use the one-hot vector a_i to represent the action condition and encode it with another fully-connected layer. These two features are added up and then served as additional input besides the noise. The model outputs the synthesized pose θ_i, which is directly used as the body pose for the anchor P_i for i-th anchor.

Scene-Conditioned Anchor Placing. In this step, we place the scene-agnostic poses $(\theta_1, \theta_2, ..., \theta_N)$ into the given scene. There are two aspects to be taken into consideration in this step. The first one is how to place poses to locations with compatible scene structure and interaction semantics. The other one is how to efficiently find multiple reasonable locations given a pose.

Therefore, we first select our placing candidates following the practice of POSA [13]. Specifically, each candidate consists of a translation parameter $\bar{t}_i$ and an orientation parameter $\bar{o}_i$ for the anchor P_i. We split the given scene into uniform non-overlapping discrete girds as translation candidates. For each discrete gird, we then uniformly sample eight different orientations that are parallel with the ground plane to build orientation candidates. Each translation candidate is paired with one of its associated orientations to form one placing candidate. For each scene-agnostic pose θ_i, we then rank all the placing candidates by their compatibility scores with the pose, which is proposed by [13] that considers both the affordance and penetration. An intuitive idea is to select the candidate with the best score. However, our empirical study shows that candidates with the same action labels tend to be located close to each other since the same action usually shares similar physical and semantic structures, as shown in the first row of Figure 7. To increase the placing diversity, we introduce an additional penalty on the locations that have been occupied by anchors with the same action labels. As shown in Figure 7, this new penalty helps produce more diverse placing candidates for similar poses. In this way, we can sample an initial placing candidate $(\bar{t}_i, \bar{\phi}_i)$ for each pose θ_i. The initial anchor $\bar{P}_i = (\bar{t}_i, \bar{\phi}_i, \theta_i)$ is then constructed subsequently.

In practice, we further adopt another sub-module called Place Refiner to improve the micro diversity of the placing candidates. Place Refiner is implemented as a CVAE model that takes the noise of θ_i, the scene context encoded by the PointNet [24] and the initial anchor $\bar{P}_i$ as the input. It outputs the offset $(\Delta t_i, \Delta \phi_i)$ to the sampled position and orientation $(\bar{t}_i, \bar{\phi}_i)$. The final position and orientation are obtained as $t_i = \bar{t}_i + \Delta t_i$ and $\phi_i = \bar{\phi}_i + \Delta \phi_i$. The framework of Place Refiner is depicted in Figure 3 (b).

3.3. Diverse Path Planning

In this step, we discuss how to generate diverse obstacle-free paths from human-scene interaction anchors. Previous works such as SMAP [11] often use standard A^* searching [9] for this purpose. The A^* algorithm tends to generate deterministic shortest path for practice. However, humans usually move stochastically in the given scene. To reflect the diversity of human path planning, we incorporate the standard A^* algorithm with scene-aware random information concerning the diversity of human motion.

To begin with, we first discuss how to apply the standard A^* algorithm into our scenario. We first divide the whole 3D scene into the same set of non-overlapping discrete grids as in Section 3.2. We then define and calculate the cost function f for each grid in the A^* algorithm [9] as:

$$f(q) = g(q) + h(q); q \in \mathcal{N}(p), \tag{1}$$

where $g(q)$ measures the cost for moving from the beginning point to grid q, and $h(q)$ measures the cost between grid q and the target grid, during searching points as the next step for p in the neigbourhood $\mathcal{N}(p)$. To ensure obstacle-free paths, we further filter out inaccessible grids that might have collisions with the human body. The collisions are detected via placing a cylinder model that approximates the volume of a human at each grid. We show an example in the right of the Figure 4 (a), where red stands for valid and blue stands for invalid. After calculating f for each grid and excluding invalid grids, an obstacle-free path connecting two human-scene interaction anchors can thus be obtained using the standard A^* algorithm. It is worth noting that the path obtained in this manner is deterministic and fixed for the same pair of two human-scene interaction anchors.

An intuitive solution to incorporate diversity in path planning is appending the cost function f defined in Equation (2) with a random noise term. This strategy sounds feasible but fails to generate reasonable paths, which is demonstrated by the examples shown in the top two rows of Figure 5. To this end, we replace the random noise term with a controllable signal m produced by another CVAE, referred as Neural Mapper. For each grid p, Neural Mapper takes sampled latent code and the local scene context feature obtained via BPS [28, 37] as the input and outputs the feasibility score for each neighbor grid $q \in \mathcal{N}(p)$. Based on the Neural Mapper, the cost function is updated as:

$$f(p, q) = g(q) + h(q) + (1 - m(p, q)); q \in \mathcal{N}(p). \tag{2}$$

The score of m indicates the feasibility of moving from the current grid to this adjacent one so that we can build the cost as $1-m$ to reflect the moving guidance by our Neural Mapper. The Neural Mapper is trained in a data-driven manner thus it can help the A^* algorithm to generate diverse and reasonable paths. We show several examples produced by Neural Mapper in the bottom row of Figure 5.

Without complex manually designed conditions and constraints, the proposed Neural Mapper equips the A^* algorithm with the ability to find diverse obstacle-free paths in a flexible and generalizable way. In Neural Mapper, we can easily change the characteristics of sampled paths by restricting the latent codes, without hurting their naturalness and coherency.

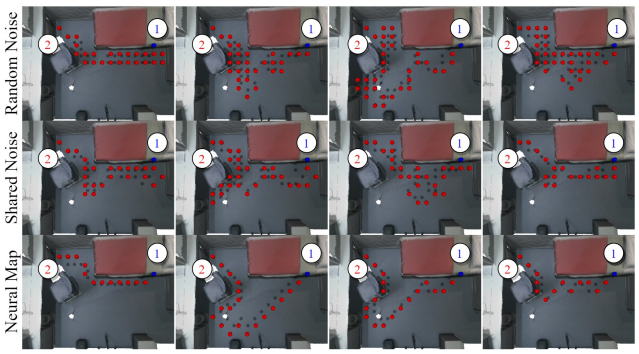


Figure 5. **Examples for diverse planning.** We sample different paths from ① to ② with different strategies. **Random Noise** means sampling different weights for each discrete gird. **Shared Noise** means all discrete girds share the same noise vector. **Neural Map** refers to probability generated by our Neural Mapper. The results demonstrate that our method is more effective in generating diverse and natural paths than simply adding randomly noise does.

3.4. Motion Completion

With the obstacle-free path obtained from path planning, we are now ready to complete the missing motions between consecutive human-scene interaction anchors. As shown in Figure 6, our motion completion network consists of two components, namely Path Refiner and Motion Synthesizer. Although paths for human-scene interaction anchors are planned in Section 3.3, this Path Refiner accounts for the gap between diverse real human motions and the path formed by straight lines between the discrete grids. Both the Path Refiner and the Motion Synthesizer follow the CVAE framework. Specially, we apply Transformer [29] as the basic architecture for both the encoder and decoder of these two networks to synthesize continuous and smooth motions. Our motion completion network simultaneously synthesizes M frame paths and body poses as [30, 31], instead of one-by-one in an auto-regressive manner [8, 10, 11].

For Path Refiner, we take the scene context encoded by PointNet [24] to synthesize the refined path. The refined path is composed of pairs of the translation and orientation sequence$\{(t_1, \phi_1), ..., (t_M, \phi_M)\}$. Following [23], we introduce the positional encoding formed from sinusoidal functions which take time steps $t \in [1, ..., M]$ as input to ensure the continuity and smoothness of the refined path. Moreover, we leverage one more positional encoding obtained from the planned path by encoding each step of the planned path (t_i, ϕ_i) in Section 3.3 by a fully connected layer, to ensure the refined path is still in the obstacle-free regions.The effectiveness of this additional positional encoding is illustrated in Section 4.2, where our Path Refiner further improves the diversity of synthesized motion.

The motion sequence with M body poses $\{\theta_1, ..., \theta_M\}$ is completed by our Motion Synthesizer. Same as the Path Refiner, we take the scene context encoded by the PointNet as

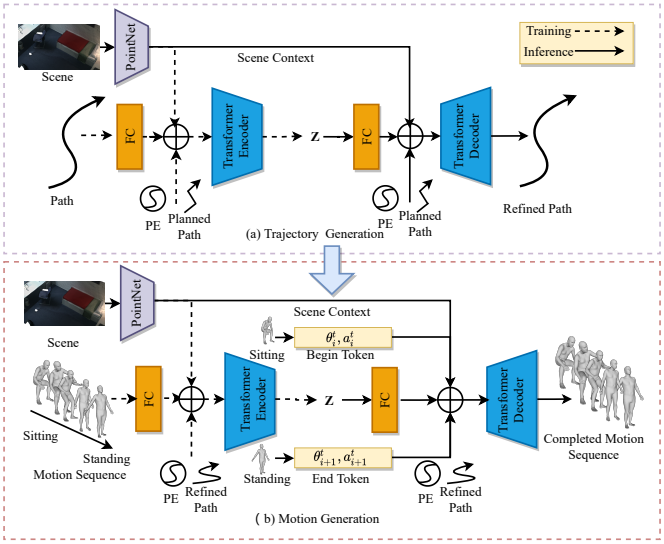


Figure 6. **Completion Network.** Motion completion network is composed of two modules, namely the Trajectory Refiner and the Motion Synthesizer. Trajectory Refiner reconstructs the motion trajectories under the guidance of planned path. The Motion Synthesizer takes the scene context, planned path, the pose and action label of human-scene interaction anchors as the inputs and generates body movements. The two-modules are trained altogether in an end-to-end manner.

the condition to complete these scene-aware motions. The completed motions should fulfill two requirements, namely matching the paths produced by the Path Refiner and naturally transforming between the given human-scene interaction anchors. To achieve this, the Motion Synthesizer at first takes the refined path as additional position encoding to guide motion synthesis, similar to the practice of Path Refiner. For the motion transformation, we need to model the relationship between the given two human-scene interaction anchors and the potential motions that could be completed in our Motion Synthesizer. Inspired by the practice of action token in [23], which helps the transformer decoder to build up the relationship between synthesized motions and the given action, we encode the action labels and poses of human-scene interaction anchors by additional fully connected layers as learnable tokens and add them to the beginning and ending of the positional encoding respectively. With these tokens, our Motion Synthesizer can directly build up this relationship between and synthesize reasonable and smooth motions. Following these two steps, the motion completion network can generate natural motions for the given human-scene interaction anchors following the planned path.

4. Experiments

In this section, we first illustrate our experiment settings and metrics for evaluation. Then we discuss the effectiveness of the proposed framework. At last, we demonstrate the qualitative results in different scenes.

4.1. Experimental Setting

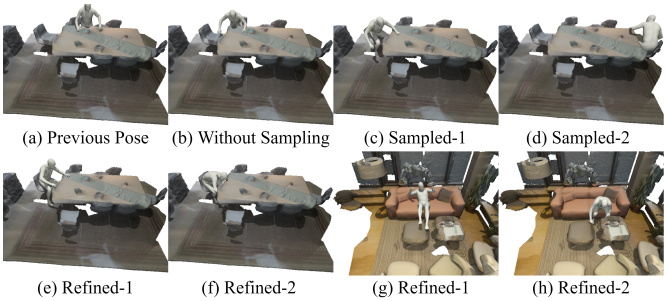
Implementation Details. All proposed CVAE models in the paper are optimized via ADAM [15] with learning rate set to $1e{-}4$ All models are trained for 40 epochs with batch size set to be 8. For better physical plausibility, we perform additional optimization used in [13] and [30] to refine human-scene interaction anchors described in Section 3.2 and the completed motions described in Section 3.4. More details for the training scheduler and the optimization are included in the supplementary material.

Dataset. Following [30, 37, 39], we train our framework on PROX dataset [12]. We manually label the motions in PROX with action labels (*i.e.* sit, lie, stand, walk, and squat) as the action condition. We do not conduct experiments on GTA-IM [4] and SAMP [11] as they do not provide reconstructed 3D real-world scenes. For the fair comparison, we follow the split of the train and test set as [30, 37, 39] and synthesize human motions on the unseen scenes during training. To demonstrate the generalization ability of the proposed framework, we further evaluate it on Matterport3D [5], which provides large-scale reconstructed 3D scenes. Please be noted that our framework does not leverage Matterport3D for training.

Diversity Metric. We measure the diversity on synthesized human motions in three aspects, namely human-scene interaction anchors, planned paths, and completed motions. To evaluate the diversity of the human-scene interacting anchors, we preform **K-Means** ($K = 20$) clustering on the synthesized human-scene interaction anchors, following [13]. To be specific, we consider two types of the clusters. The first one considers all parameters (θ, t, ϕ). The second one only considers translation t and orientation ϕ. The diversity is measured as the entropy of the cluster sizes and the average distances between the clusters center and the samples belonging to it. We evaluate path diversity by the standard deviation (**STD**) of distances between the paths from Neural Mapper and the ones from the standard A^*. To fairly compare with previous works that manually assign anchors or target objects [13, 39], we evaluate the diversity of the synthesized human motions with the fixed human-scene interaction anchors. To measure the ability of our motion completion network in generating diverse results, we do not introduce the diverse sampling strategies as [36, 38]. Following [11], we calculate the Average Pairwise Distance (**APD**) on the SMPL-X parameters of synthesized motions to measure its diversity.

Table 1. **Evaluation on human-scene interaction anchors.** We evaluate the diversity of the human-scene interaction anchors (Anchor, considering θ, t, and ϕ) and the placing (Position, considering only t and ϕ) with/without optimization post-process. S means the sampling strategy based on pose relationship in Section 3.2, and R means our Placing Refiner.

Method	Anchor		Position	
	Entropy ↑	Cluster ↑	Entropy ↑	Cluster ↑
Baseline [13]	2.62 / 2.60	2.44 / 2.40	2.63 / 2.61	0.68 / 0.67
Baseline [13] + S	2.74 / 2.73	2.55 / 2.53	2.69 / 2.68	0.79 / 0.78
Baseline [13]+ S + R	**2.77 / 2.73**	**2.57 / 2.53**	**2.72 / 2.70**	**0.83 / 0.80**



(a) Previous Pose (b) Without Sampling (c) Sampled-1 (d) Sampled-2

(e) Refined-1 (f) Refined-2 (g) Refined-1 (h) Refined-2

Figure 7. **Placing Results.** The first row shows the results with and without the pose related sampling strategy. The second row shows the results where our Place Refiner and the optimization post-processing in [13] works together.

Naturalness Metric. We evaluate the naturalness of synthesized motions via **user study** and the physical plausibility. We ask users to compare our results against other methods and score them from 1 to 5 (the higher the better) as the results. Besides, we involve the **non-collision** score and **contact** score [30, 37, 39] to measure the physical plausibility of synthesized motions between the 3D scenes.

Motion Metric. To evaluate the quality of the whole synthesized motions, we follow [11] to calculate the Frechét Distance (**FD**) between synthesized motions and ground-truth motions. This distance is computed using the parameters $P_i = (\theta_i, t_i, \phi_i)$ of each frame.

4.2. Experimental Results

In this section, we show quantitative results on PROX dataset. The quantitative results on Matterport3D dataset are included in our supplementary materials. We also show qualitative results on these two datasets in this section.

Human-Scene Interaction Anchors. We first show the diversity of synthesized human-scene interaction anchors in Table 1. For this evaluation, we sample 100 poses for each action and employ the placing strategy in POSA [13] as our baseline. As shown in the table, the position related sampling process and the Pose Refiner can both improve the diversity of interaction anchors. We further show the effectiveness of these two process in Figure 7. Examples shown in this figure are all generated from the action label "sit". (a) shows the first placed poses. Without the position related sampling, (a) and (b), which have the same action label, are placed close to each other. (c) and (d) are the generated anchors using our position related sampling. It is revealed that they interact with different objects in the scene. The second row demonstrates the result pairs ((d) V.S. (e) and (f) V.S. (g)), which are produced by our Place Refiner works and optimization post-process [13]. Using the diverse translations and orientations as initialization states, the optimization algorithm can produce diverse optimal solutions. In the supplemental material, we show comparison with previous works [37, 39] that are extended to synthesize specific actions using our action condition.

Table 2. **Evaluation on diverse planning module**. We compare against the standard A^* algorithm and methods with sampled random noises. The metrics show the diversity and naturalness of the planned paths.

Method	1/6 ↑	1/3 ↑	1/2 ↑	2/3 ↑	5/6 ↑	User Study ↑
Standard	0	0	0	0	0	4.31(0.48)
Random Noise	0.346	0.566	0.628	0.523	0.324	2.52(0.53)
Shared Noise	0.297	0.483	0.603	0.485	0.281	3.52(0.45)
Ours	0.286	0.446	0.508	0.415	0.233	4.27(0.52)

Planning. We compare the diversity and naturalness of the planned path. For the evaluation of diversity, We compute the standard deviation of the distances between sampled paths and the paths produced by standard A^* . In practice, we calculate distances between the discrete points on the paths, which are set as 1/6, 1/3, 1/2, 2/3, and 5/6 of the sample paths. We also show the results of the user study to reflect the naturalness of the planned paths. The number of samples is set to be 50 for each method. The evaluation results are shown in Table 2. Standard A^* algorithm, which is used in [11] only produces the deterministic path for practice while the proposed Neural Mapper can generate diverse paths with similar naturalness as the standard A$*$ does. On the other hand, two methods using random noises cannot produce natural results, although they generate more diverse paths than ours. Similar results are also demonstrated in Figure 5. Compared with the methods using random noises, Neural Mapper can provide consistent and reasonable guidance for the similar local scene context to avoid unnatural moving. Moreover, Neural Mapper can cope with other manual constraints such as avoiding passing a certain region. We will discuss it in our supplementary materials.

Motion Synthesis. In this subsection, we compare with other advances on scene-aware motion synthesis [11, 30, 31]. We use the official model of [30] trained on PROX

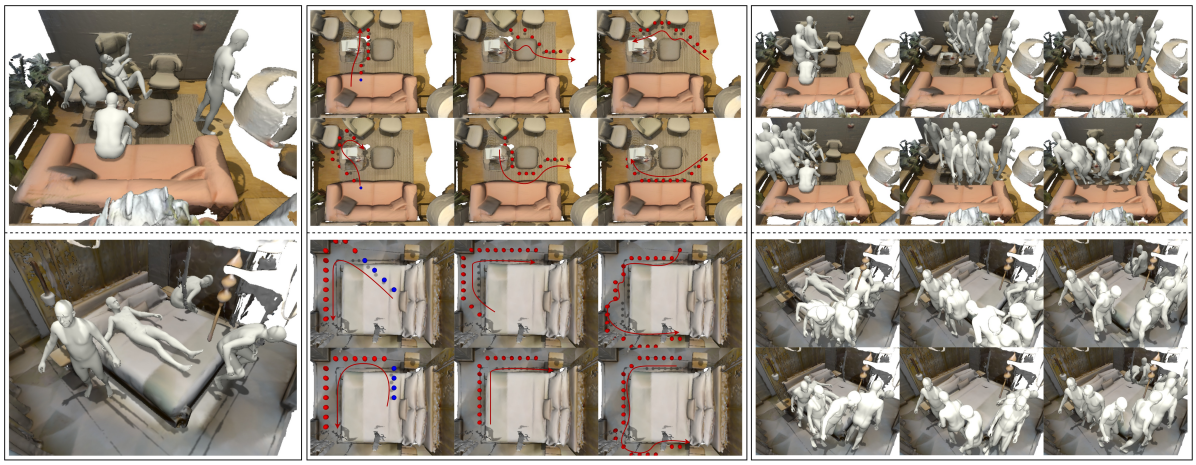


Figure 8. **Qualitative Results.** In this figure, we show the output of each stage in our framework. The first row is synthesized on PROX dataset, and the second one is on Matterport3D. Our framework can synthesize motions with diverse interactions to the given scene.

Table 3. **Evaluation on motion completion module.** We mainly evaluate the models in two aspects. FD is used to show the completion ability. APD is used to evaluate motion diversity. We compare our framework with several state-of-the-art methods. Specially, "w/ OPT" and "w/o OPT" refer to the results obtained with/without optimization post-process [30]. "Ours*" means our motion completion network without the Path Refiner.

Method	FD ↓		APD ↑	
	w/o OPT	w/ OPT	w/o OPT	w/ OPT
SA-CSGN [31]	176.20	175.28	2.13	2.15
Wang et.al. [30]	121.22	120.01	0.00	0.00
SAMP [11]	115.34	114.22	2.56	2.57
Ours*	126.46	124.52	2.46	2.46
Ours	**112.74**	**111.65**	**2.77**	**2.78**

dataset and extend [11] and [31] to PROX for fair comparison. In Table 3, we first compare against these methods using **FD** and **APD** for the motion quality and diversity. Firstly, for the comparison of **FD**, we sample 500 motion sequences which begin with the same action, as well as 500 motions which finish the same action. Besides, for the comparison of **APD**, we sample 100 pairs of human-scene interaction anchors and synthesize 10 motions for each pair. It is revealed that our method achieves the best results against other methods. All the comparison results show that our method can synthesize more diverse and natural motions than other methods do. In the supplementary material, we first compare more naturalness results between these methods, *e.g.* physical compatibility and user study. Then we further discuss our design choice on the motion completion network, including the effectiveness of the Path Refiner and the positional encoding based on planned paths.

Qualitative Results. We show more qualitative results of the proposed method on the PROX [12] and Matterport3D [5] in Figure 8. We show all three aspects of the synthesized scene-aware motions, including the human-scene interaction anchors, diverse planned paths, and completed motions. These results demonstrate that our framework can synthesize diverse human motions in the specific scene contexts for the given target action sequence. More qualitative results are included in the following video[1].

5. Conclusion

In this paper, we focus on synthesizing diverse and natural human motions in the given scene environment guided by target action sequence. We decompose the diversity of scene-aware human motions into three levels, namely the diversity of action-conditioned human-scene interactions, the diversity of obstacle-free paths, and the diversity of body movements. To comprehensively leverage the inherent diversity of human motions, we propose a novel hierarchy framework with each component accounting for each level of the diversity. Thanks to the effective decomposition of diversity and elaborated designed modules, our framework is able to produce various vivid human motions in the scene across all three levels with improved efficiency and generality. Furthermore, the factorized design of our framework make it can be easily incorporated into other human motion synthesizing frameworks.

6. Acknowledgement

This study is supported under the General Research Fund (GRF) of Hong Kong (No.,14205719), the RIE2020 Industry Alignment Fund–Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

[1]Please refer to http://wangjingbo.top/papers/CVPR2022_PoseGeneration/Posegeneration.html.

References

[1] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018. 1, 2

[2] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[3] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *International Conference on Computer Vision*, 2021. 2

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 1, 2, 6

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 6, 8

[6] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 3

[8] 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 *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 2020. 2, 5

[9] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 1968. 4, 5

[10] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 2020. 2, 5

[11] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, Oct. 2021. 1, 2, 3, 4, 5, 6, 7, 8

[12] 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*, 2019. 2, 6, 8

[13] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Conference Computer Vision and Pattern Recognition*, 2021. 2, 3, 4, 6, 7

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 3

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. 6

[16] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. 3

[17] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[18] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017. 3

[19] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[20] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition*, 2019. 3

[22] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2

[23] 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. 2, 5, 6

[24] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 4, 5

[25] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 2015. 2

[26] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics.*, 2019. 2

[27] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014. 3

[28] 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*, 2020. 5

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 5

[30] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 5, 6, 7, 8

[31] 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*, 2021. 1, 2, 5, 7, 8

[32] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, Xiaolong Wang, and Trevor Darrell. Hierarchical style-based networks for motion synthesis. In *European Conference on Computer Vision*, 2020. 2

[33] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2

[34] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 3

[35] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[36] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[37] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, Nov. 2020. 2, 3, 5, 6, 7

[38] 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*, 2021. 1, 6

[39] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3d people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 6, 7

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3