MAMMOS: MApping Multiple human MOtion with Scene understanding and natural interactions

Donggeun Lim[1], Cheongi Jeong[2], and Young Min Kim[1, 3]

[1] Dept. of Electrical and Computer Engineering, Seoul National University [2] Samsung Research

[3] Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

rms2836@snu.ac.kr, cbda1100@gmail.com, youngmin.kim@snu.ac.kr

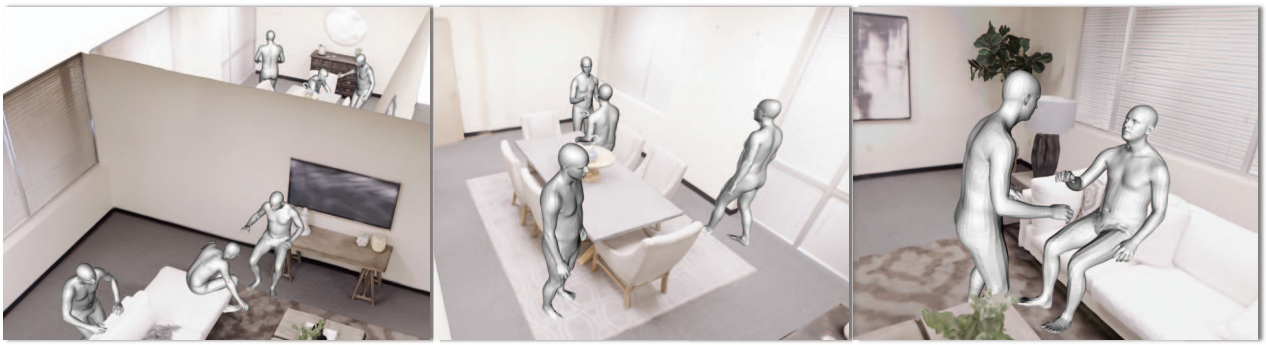


Figure 1: **MAMMOS** automatically generates multiple human motions sharing a given indoor space. Generated virtual humans correctly understand the scene context around them, and they are also able to interact with each other naturally.

Abstract

We present MAMMOS, an automated framework that generates the motions of multiple humans that naturally interact with each other in a given 3D scene. Many practical VR scenarios require creating dynamic human characters in harmony with the surrounding environment and other people. However, it is hard for an artist to manually generate multiple character motions tailored to the given 3D scene structure, or gather sufficient data to train an automated system that jointly considers the entangled requirements. MAMMOS is a hierarchical framework that successfully handles spatio-temporal constraints and generates high-quality motions. Given a simple tuple of action labels of the desired motion sequence, MAMMOS first places anchors in time and location for characters that avoid collisions yet enable necessary interactions. Then we generate the timelines of individual collision-free paths within the scene and connect them to perform diverse and natural motions. To the best of our knowledge, we are the first to generate long-horizon motion sequences of multiple humans with realistic interactions such that we can automatically populate the 3D scenes.

1. Introduction

Suppose we have a virtual asset of a 3D scene and we want to create character motions such that they naturally perform everyday activities (Figure 1). The 3D space can be a digital twin of a real-world environment or a purely virtual collaboration space. In addition to the spatial layout of furniture or nearby objects, the generated motions need to respect other people in the scene. When two people interact, they need to maintain proper social distance [13] with eye contact and synchronize their timing. If they are not interacting, their paths need to avoid collisions. Because all interpersonal relationships are dynamic, it is not trivial to consider them jointly with the scene context. Furthermore, it is infeasible to collect sufficient training data that exhaustively represent the diverse possible combinations of interpersonal relationships and 3D scene layouts. Nonetheless, humans are social beings, and it is necessary to include interaction between them when synthesizing many of practical scenarios. Although many prior works generate motions that consider the scene context, to the best of our knowledge, none of them systematically unravels the spatio-temporal constraints of multi-human scenarios.

Our objectives are three-fold: input should be easy, and

(a) Anchor Placement **(b)** Path Generation **(c)** Motion Completion (+ Optimization)

Figure 2: **Overview of the pipeline.** Our system consists of four stages: anchor placement, path generation, motion completion, and optimization. (a) The anchor placement stage creates the characters' poses corresponding to the input action labels. For action pairs with specified interaction, the created anchors of the two characters are in proximity, facing each other. (b) In the path generation stage, we create collision-free paths between consecutive anchors. We also confirm the timeline of the waypoints such that interactions are synchronized. (c) In the motion completion stage, we synthesize smooth scene-aware motions that follow the created paths. The subsequent optimization stage refines the motion to be physically correct and natural.

the system should be scalable yet able to generate natural outputs. Instead of manually assigning the 3D configurations of body joints in the scene, we would like to receive a simple high-level description of motions as input, such as the number of people and their action labels (sit, stand, lie, and interact). All the subsequent complexities are automated, given the 3D scene. We achieve scalability to handle the large-scale scene with multiple people by first considering the global spatio-temporal context in a coarse level and subsequently generating fine motion with local information. Most of all, the motion has to be natural. We jointly consider the complex constraints, and avoid obvious violations of physical laws. At the same time, we enhance naturalness with eye contact and diverse sequences of interacting motions.

To tackle the challenging problem, we design a modular approach as shown in Figure 2. Given the text input describing action labels of a number of characters, our system first places anchors considering human-scene interaction or human-human interaction. The anchors are placed in the spatio-temporal domain within the scene to best show the start and end of a given label of action and yet avoid collisions. Then the path generation module can instantiate a set of plausible trajectories between the anchors from a diverse stochastic distribution. Following the trajectory, we generate detailed motions for individual characters, where we can only consider local scene or interaction contexts assigned for the intermediate waypoints of the assigned path. By detaching the holistic analysis from detailed motion generation, we can utilize the prior works on motion generation without extensively considering the context of multiple characters within the scene.

In summary, MAMMOS is the first to generate motions of multiple humans with mutual interactions within the scene context. The proposed pipeline is a practical system with easy input, scalability, and natural output trained with limited datasets and successfully creates multiple character motions adapted to new, large-scale scenes. Our framework presents scalable interaction generation that can create a realistic story in a shared virtual space.

2. Related Works

Motion Synthesis Creating the motion of the human body has been investigated in various contexts. Several works attempt to predict a future sequence of motions given a frame or a sequence of frames [26, 12, 5, 3, 40]. Many applications, on the other hand, require generating natural motion between keyframes of static poses, which are investigated in [7, 15]. Recent studies even synthesize motion without any explicit pose information, either from languages [1, 2, 10, 27, 28, 4, 9, 34, 38] or music [35, 19, 20, 21]. To allow more explicit control of desired motion, some methods take an input of specific action labels and generate a sequence of motion either with a recurrent unit [10] or a transformer [27]. Recent methods go beyond a small set of action categories and generate diverse motions from free-form text [28, 4, 9, 34, 38]. None of the previous works, however, tackle long-term motion sequences with action-label switches, scene interaction, and

multiple human interactions.

Human-Scene Interaction When generating human motions, the surrounding environment is another interesting context to consider. Starting from efforts to place a posed static human in a 2D image [33] or 3D skeleton in RGB, RGB-D, or depth image [11, 22], recent works provide means to place a full 3D mesh of a human in a 3D scene utilizing high-quality parametric models [23, 31, 25]. Recent works utilize the 3D body model with expressive hands and faces [25] and place them on 3D scenes [41, 39, 18, 43]. Zhang et al. [39] observes the local scene context using explicit basis point sets (BPS) [29], while Hassan et al. [18] utilize contact probability between human and motion. A more recent method [43] proposes compositional human-scene interactions given only a training set containing atomic interaction data. Our framework also creates static body poses of anchors tailored to the given action labels and the scene but further connects them to generate a smooth motion sequence. In part, the proposed method is similar to connecting end poses of the human body [37] or synthesizing motions of a given action label within the scene [36]. However, our approach further generates motions of multiple humans within the scene, such that they harmoniously interact with the environment and other humans.

3. Method

Given a scene $\mathcal{S}$ and a sequence of desired action labels with interaction $\mathcal{A}$, MAMMOS generates the sequence of human motion $\mathcal{M}$: $(\mathcal{S}, \mathcal{A}) \rightarrow \mathcal{M}$. The scene can be provided as a CAD model, a triangular mesh, or a point cloud scan as long as we can estimate the distances. If the user wants to generate the motion of N humans within the scene, the desired action sequences are provided as $\mathcal{A} = \{A^1, ..., A^N\}$, where A^i indicates the sequence of discrete action labels that i^{th} person needs to perform. The actions consider both the geometric context and the mutual interaction between people in the scene. The action sequences are composed of variable numbers of action labels paired with interaction indicators, $A^i = \{(a_1^i, c_1^i), (a_2^i, c_2^i)..., (a_{M_i}^i, c_{M_i}^i)\}$. Specifically, the action label $a_j^i \in \{\text{stand}, \text{sit}, \text{lie}\}$ indicates the category of the body pose in relation to the scene context. The interaction indicator $c_j^i \in \{0, \ldots, K\}$ represents whether the paired action a_j^i is an independent action of the character ($c_j^i = 0$) or has to interact with another human doing action paired with the same indicator value. We only allow two-person interaction, meaning there are exactly two identical interaction indicator values ($c_j^i = k, k = 1, \ldots, K$) within the set of action sequences. Then MAMMOS automatically assigns the exact locations and time windows for the actions to take place and creates smooth and natural motion trajectories. The generated motion $\mathcal{M} = (M^1, \ldots, M^N)$ is composed of a sequence of motion parameters $M^i = \{(r_0^i, \phi_0^i, \theta_0^i), \ldots, (r_T^i, \phi_T^i, \theta_T^i)\}$ derived from the 3D parametric model of human body, where $r_t^i \in \mathbb{R}^3$ is the global translation of the root position, $\phi_t^i \in \mathbb{R}^6$ is the global orientation in the 6D continuous representation [44], and $\theta_t^i \in \mathbb{R}^{32}$ represents the body pose in the form of VPoser [25].

The overall pipeline is depicted in Figure 2. We resolve the complex spatio-temporal constraints for motion generation by decomposing the problem into four stages, and progressively generate finer motions. The first anchor placement stage places N humans in their anchor poses corresponding to the sparse input action labels within the scene (Section 3.1). The next stage is path generation, which finds collision-free paths on grid locations of the discretized scene to coarsely connect the synthesized anchors (Section 3.2). The third stage of motion completion then interpolates the paths on the grid to find full motion parameters of dense motion (Section 3.3). The last optimization stage improves the motion quality and returns a realistic and physically plausible sequence of motions (Section 3.4). At each stage, we jointly consider the other humans and the spatial context at an appropriate resolution.

3.1. Anchor Placement

Given the scene $\mathcal{S}$ and the sequence of action labels $\mathcal{A}$, the anchor synthesis finds the location and pose $(r_j^i, \phi_j^i, \theta_j^i)$ of N people in the scene that corresponds to the action labels a_j^i. If there is no interaction associated with the action ($c_j^i = 0$), we can individually generate a pose θ_j^i given the action label a_j^i using a conditional variational autoencoder (CVAE) architecture [36] and place the posed character in appropriate translation r_j^i and rotation ϕ_j^i considering the scene layout as in [18]. To briefly elaborate, we choose from a set of discrete candidate translations and rotations: candidate translations are the cells of the grid uniformly dividing the scene, and the candidate rotations are eight discrete orientations around the vertical axis. We exhaustively test all the combinations and select the top ten best-scoring positions in terms of affordance while avoiding penetration. The top ten candidates are individually optimized to find the best final anchor. However, existing methods do not consider multi-human scenarios within the scene. MAMMOS additionally considers constraints provided as interaction labels and the inter-human distances to avoid a collision.

Interaction Anchor Placement Because the interaction anchors abide by additional inter-human constraints, we first place interaction anchors ($c_j^i \neq 0$) and place the remaining ones using existing methods. In addition to the spatial context considered for normal anchors, the pair of action anchors with the same interaction indicator number

Figure 3: **An example of interaction anchor placement.** Interaction anchors are placed so that the direction of the body does not deviate more than a threshold angle from the virtual line connecting two humans (dotted line).

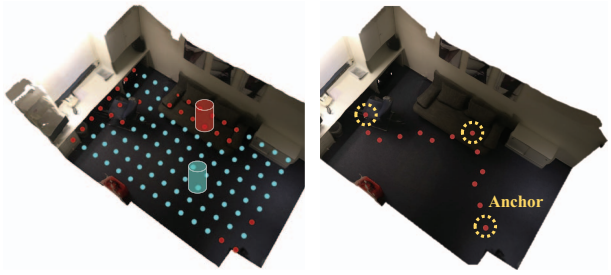
(c_j^i) should be in an appropriate distance and angle to face each other as shown in Figure 3. We simplify the problem and find the anchor positions within the 2D overhead map of the scene. The constraints are defined in terms of the root positions and face orientations, such that they are visible within the near and middle peripheral vision of the human eye (around ± 30 degrees) [8].

Anchor Occupation Rule After we place the interaction anchors incrementally with the increasing indicator numbers, we fill the remaining anchors to avoid obvious collisions. There are two simple rules: we maintain sufficient distances between (1) temporally adjacent anchors from the same human id; and (2) first or last anchors of different people such that all starting positions are nicely spread from each other as well as the ending positions. Basically, we sequentially generate non-interacting anchors and avoid collisions against the aforementioned anchor positions if they are already assigned. Other intermediate positions can be adjusted in the next stage when we generate paths and regulate temporal precedence.

3.2. Path Generation

Path generation assigns the sequence of grid locations for multiple people that satisfy the spatio-temporal context scene and interaction Given the anchor places with action labels, we generate paths between the subsequent anchors such that there are no collisions, and the interaction pairs are in sync in time. This is a very high-dimensional optimization compared to recent approaches that only consider the spatial constraints [37, 36, 16], and it is challenging to manually create natural motion considering the complex constraints. The paths are searched over discretized slices of time intervals and a grid of the 2D projection of the input scene. First, individual path planning generates spatial paths for individual humans. The subsequent timeline integration adjusts the temporal alignment.

Individual Path Planning We generate diverse individual paths avoiding collision from the paths of other humans on the 2D grid that is used in Section 3.1. The path planning process is also illustrated in Figure 4. For each grid,



(a) Navigating 2D Grid Map **(b)** Path Planning Example

Figure 4: **An example of the path planning process.** (a) Navigating a 2D grid map with the proxy cylinder. Cyan points are passable grid points and red points are non-passable grid points. (b) A sample created path. Anchor positions are circled with yellow dotted lines.

we find the intersection between a proxy cylinder and the scene to assess whether the grid cell is free for a human to pass by. Starting from an anchor position, the path to the subsequent anchor is incrementally generated by a modified A^* algorithm [14]. We assign the cost of a free grid point q as $f(q) =$

$$\underbrace{g(q) + h(q)}_{A^*} + \underbrace{(1 - m(p,q))}_{\text{Neural Mapper [36]}} + \underbrace{C \cdot \mathbb{1}_{\text{collision}}(q, t+1)}_{\text{Collision Avoidance (Ours)}}, \quad (1)$$

where $q \in \mathcal{N}(p)$ is a neighboring cell of the current position p. The cost function is composed of three groups. The first group is the terms from the original A^* algorithm. $g(q)$ measures the cost from the start point to q, and $h(q)$ is a heuristic function that estimates cost from q to the goal position, which indicates the next anchor in our case.

Because the standard A^* algorithm always creates a deterministic path, Wang et al. [36] suggested generating diverse and realistic paths with additional stochasticity referred to as the Neural Mapper. $m(p, q)$ is a probabilistic feasibility score of moving from grid point p to q, estimated by a trained neural network.

The third group of the terms in Equation (1) is our modification to filter out paths that incur collisions due to the temporal occupancy of other human trajectories. Basically, we add an indicator function $\mathbb{1}_{\text{collision}}(q, t+1)$ that returns 1 if a collision occurs at q at the next timestep $t+1$ else 0. C is a very large constant and effectively adds an unacceptably high cost to f when a collision is expected at the next timestep $t+1$ at q. Note that only the collision indicator function is dependent on t. If the cost function g or h also depends on t, the search space of the algorithm becomes prohibitive, and the path cannot be created within a reasonable time.

Timeline Integration Timeline integration adjusts temporal windows such that the interaction anchors with the

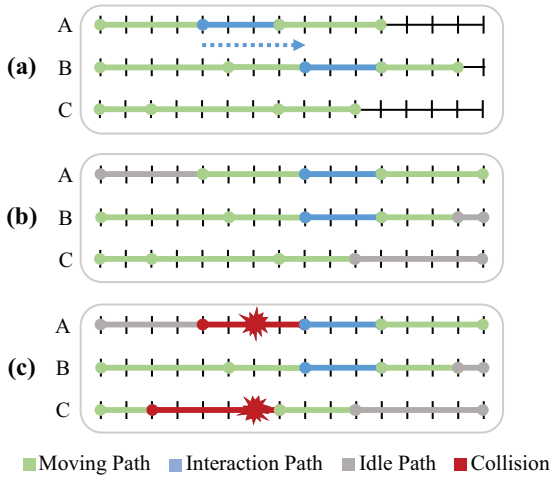


Figure 5: **Steps iterated for the timeline integration.** (a) In the initial timelines, the interactions (blue) are not synched after independent path generation. To synchronize, the timeline of Human A should be shifted by the timestep indicated by the blue dotted-arrow. (b) We adjust the timelines so that all interactions are synced with their counterparts. We add an idle time (gray) to fill the temporal window after shifting. (c) After each modification for the timeline integration, we check possible collisions for paths. We recreate the subpaths with collision (red).

same indicator number are temporally aligned for a duration of time, and we avoid possible collisions. The number of action labels is different to start with, and the grid paths connecting generated anchors are in different lengths. We iterate to add necessary idle times to match the times for interaction and to check possible collisions as illustrated in Figure 5. During idle time, the character stays at the same spatial grid. When a collision is unavoidable, we regenerate the problematic subpaths subject to collisions and iterate the process.

3.3. Motion Completion

Motion completion creates frames of smooth motion that follow the discrete paths in the grid. As the spatial-temporal context is already considered from the path generation, the motion completion can focus on local generation as suggested by hierarchical frameworks [37, 42]. There are three types of motions to generate in our framework: the moving motion of the paths, the interaction motion of the interaction anchors, and the idle motion derived from the timeline integration.

Moving Motion For motions moving between grid points, we define keyframes $(r_k^i, \phi_k^i, \theta_k^i)$ and interpolate consecutive motion keyframes using neural networks. For anchors, we already have target keyframe poses (Section 3.1). For intermediate grid points in the paths between anchors, we assign the grid locations as the x, y locations of the pelvis of the motion keyframes. We set the rotation ϕ_k to face the next grid point. The pose parameters θ_k are derived from the predefined set of walking poses. We place alternating stable feet on the path and update the z position to minimize the penetration loss and the contact loss for optimization (Section 3.4).

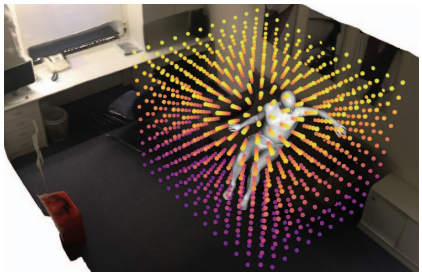


Figure 6: **An example of grid points where local SDF is retrieved.** Cube-shaped and human-centered grid points are used to encode local scene context (fewer grid points are depicted than are actually used).

Then we interpolate the keyframes by modifying the motion generation of [37] to encode local scene context. While previous works [37, 36] generate plausible motion and adapt to the scene context, their original works place one person at a time and are constrained to small rooms. On the other hand, we handle larger scenes to place multiple people with interaction. We noticed that the performance does not generalize to larger scenes when the neural network processes the entire scene as an input. Instead, we train a framework only with human-centered local information encoded as the signed distance field (SDF), achieving much more stable performance in scenes of various scales. As shown in Figure 6, we extract a grid position centered at the human pelvis and calculate the local SDF information of neighboring positions. The motion interpolation simply transfers the generated local motions into the global coordinate frame. The moving motion is trained with the PROX dataset [17] which contains rich interaction between human and scene.

Interaction Motion When two people interact with each other, we enrich the poses with natural interaction for a fixed duration. The initial anchor poses only consider the action labels, either sitting or standing. While the moving motion interpolates the intermediate poses of fixed intervals along the generated path, the interaction motion needs to be a variable length of vibrant motions. To fit the purpose, we use the CVAE architecture conditioned only on motion history and enable different duration by employing an RNN structure. Specifically, we employ GRU-based CVAE architecture used in the marker predictor of the GAMMA from [42].

Before **After**

Figure 7: **An example of eye contact optimization.** Red arrows are the estimated current eye directions and the blue ones are the target eye directions to be matched to make eye contact.

It is trained with the TCD Hands dataset [24], which contains the SMPL-X parameters [25] on hand gestures and finger motions in our daily life, such as talking, waving, signing, etc. We extract the motion of the upper body from the dataset with interacting motion and train the CVAE to generate diverse and natural interaction motion that seamlessly continues from the given anchor.

Idle Motion Lastly, we add subtle movement even when people stay still due to the idle moments introduced by the timeline integration. Without additional movement, the idle people freeze in the anchor positions, which appears awkward and unnatural. As a simple yet effective remedy, we introduce subtle posture variations by adding a small Gaussian noise $\sim N(0, \sigma)$ to the anchor pose in the VPoser embedding space. The random variation is further smoothed out using the smoothness loss during optimization (Section 3.4).

3.4. Optimization

The last optimization stage refines the created motions to be physically valid and natural. Given the sequence of generated motion for each person M^i, motion parameters $(r^i_{0:T}, \phi^i_{0:T}, \theta^i_{0:T})$ are optimized based on physical constraints. We adopt constraints from previous works to penalize physically impossible artifacts: foot location, penetration, contact, and smoothness constraints from [37] and self-penetration constraints from [6]. The full loss is presented in the supplementary material. Notably, we introduce novel eye contact optimization, which plays a significant role in realism for natural interaction.

Eye Contact While two people are interacting with each other, they need to make eye contact. We add eye contact optimization on the frames performing the interacting motion. We define the energy function using the estimated eye direction derived from the pre-fixed vertex topology of the SMPL-X body mesh as depicted in Figure 7. Let's denote a vertex on the center of the forehead as v^k_f and the one on the back of the head as v^k_b for a human k. Assuming that human i and j are interacting with each other, our eye contact loss can be defined as below,

$$E_{\text{eye}} = \underbrace{\arccos \frac{e^i \cdot w^i}{\|e^i\| \|w^i\|}}_{\text{angle between } e^i \text{ and } w^i} + \underbrace{\arccos \frac{e^j \cdot w^j}{\|e^j\| \|w^j\|}}_{\text{angle between } e^j \text{ and } w^j} \quad (2)$$

where $e^i = v^i_f - v^i_b$ is the current eye direction of human i and $w^i = v^j_f - v^i_b$ is the target eye direction of human i which is towards the eye of human j. e^j and w^j are obtained similarly. Minimizing E_{eye} ultimately makes human i and j look into each other's eyes.

We encourage eye contact with minimal head rotation and avoid significantly changing the previously generated anchor pose. Instead of optimizing the embedded pose representing the entire body $\theta \in \mathbb{R}^{32}$, we only alter the relative rotation between neck and head extracted from full body pose $\Theta \in \mathbb{R}^{63}$. However, directly optimizing a subset of the body pose Θ can occur undesired deterioration of the body shape. We add the pose prior loss [25] to the optimization constraint to maintain valid body shapes.

Ablation Method	Winning Percentage of 'all'(%) ↑	Relative Score of 'all' ↑
all vs all - I	93.4	2.84 (1.21)
all vs all - E	86.8	2.77 (1.20)
all vs all - C	96.7	3.78 (1.13)
all vs all - I - E - C	95.6	3.64 (1.11)

Table 1: **Ablation study results of multi-human motions.** We provide the ratio of users who chose that the complete pipeline with all components is more natural and the mean and standard deviation of the scores that user provided in a scale from 1 to 5 (5 is more natural).

4. Experiments

We evaluate the quality of the generated motions using MAMMOS. When we employ network architecture from previous studies, we mostly use the same training settings as in their original papers. For moving motion completion described in Section 3.3, we replace the original scene context feature by the local SDF information. Correspondingly, we replaced the PointNet [30] in the original architecture with fully-connected layers. They are trained and tested with the same split of PROX dataset [17] as in [37, 39, 41, 36]. We also evaluate the generalization to larger scale 3D scenes with Replica dataset [32]. For interaction motion of the upper body, we reduced the overall dimensions of the original CVAE architecture [42] by half to adapt to the smaller dataset size. The exact architectures and dimensions are explained in the supplementary material.

Figure 8: **Qualitative results for multi-human motion.** We show sample results of multi-human motions with diverse interactions in different scenes.

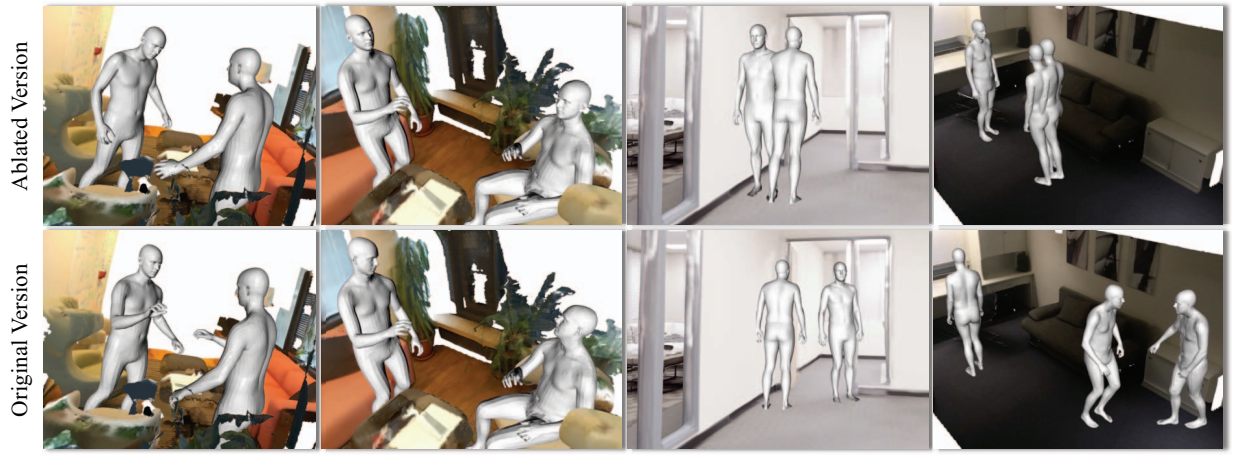


Figure 9: **The effect of each technical component.** MAMMOS implements critical components to create natural and physically plausible interactions between people.

4.1. Multi-Human Motion

MAMMOS generates multi-human motion including natural interactions, which has not been addressed in previous literature. Since there are no existing implementations for baseline comparison, we visually compare the effectiveness of our approach with users' assessment. Our approach is compared against ablated versions that eliminate the key components of MAMMOS: collision-free path generation (-C), interaction motion generation (-I), and eye contact optimization (-E). Users are asked to compare two versions of multi-human motions and choose a preferred one. They make a total of five binary comparisons. Two of them are to choose the more natural one between the result of applying all methods mentioned above (all) and the result of nothing applied (all-I-E-C). In the other three questions, we ask users the same question, except the comparison target is changed to an ablated version without one of the components.

The responses to the user study are summarized in Table 1. In all cases, our proposed method received the majority of choices, which clearly demonstrates that all of the proposed components are essential in realizing natural human-human interaction. While all of them are critical, the physical violation of collision-free paths appears especially noticeable. Sample frames of our multi-human motion are available in Figure 1 and 8, whereas the comparison against ablated versions is in Figure 9.

Scene	Anchor (s)	Path (s)	Motion (s)	Optimization (s)
N3OpenArea (PROX)	10.66 (0.50)	0.01 (0.01)	0.27 (0.02)	1.83 (0.14)
Apartment1 (Replica)	13.58 (1.10)	0.01 (0.01)	0.29 (0.04)	1.77 (0.23)

Table 2: **The average and standard deviation of runtime for 50 runs on an RTX 4090 GPU.** The anchor placement stage are presented in seconds per anchor, while other stages are measured in seconds per 30 frames.

Although different combinations of human-object and human-human relationships can significantly increase the required resources, our staged pipeline handles them in a scalable way. Table 2 compares the runtime of two different environments, whereas the Replica dataset is about four times larger than the PROX dataset. MAMMOS only experiences a slightly longer anchor placement stage in the Replica dataset to evaluate more candidate locations. Other stages mainly rely on the local context, and are barely af-

Method	Non-collision↑		Contact↑		Smoothness↑	
	PROX	Replica	PROX	Replica	PROX	Replica
Long-term [37]	94.58	99.39	95.04	90.89	96.41	93.28
Ours	**97.17**	**99.72**	**99.80**	**99.99**	**97.71**	**97.42**

Table 3: **Evaluation on the physical plausibility for single-human motions (without optimization).** All scores are high when using our method for both PROX and Replica scenes.

(a) Ours **(b)** *Long-term* [35]

Figure 10: **Comparison of single-human motion.** Our approach produces more natural and temporally smooth results without jittering artifacts. (The brighter the color of a human, the more time has passed.)

fected by the scene size. The total runtime increases almost linearly as the number of humans increases.

4.2. Single-Human Motion

While our main focus is generating multi-human motion, our pipeline also results in more natural single-human motion. Unlike multi-human cases, we can also compare against previous works that create motions of a single human within a scene context, namely *long-term*[37], and *towards* [36]. Both are trained with the PROX dataset as ours, but the implementation is available only for the *long-term*. We, therefore, include *towards* only for the visual comparison using the same scene.

We evaluate the quality of motion before optimization in Table 3. The evaluation is presented in three metrics: the non-collision score, the contact score, and the smoothness score, which are defined in the supplementary materials. The non-collision score and contact score [41, 37] evaluate the physical plausibility. We sample 200 anchor pairs and generate a motion sequence using these pairs to calculate the non-collision score and contact score. Our method exhibits better non-collision and contact scores, generating physically plausible human motion in various scenes. The smoothness score evaluates the smoothness of the synthesized motion by comparing the distances between the body vertices of consecutive frames. The smoothness score

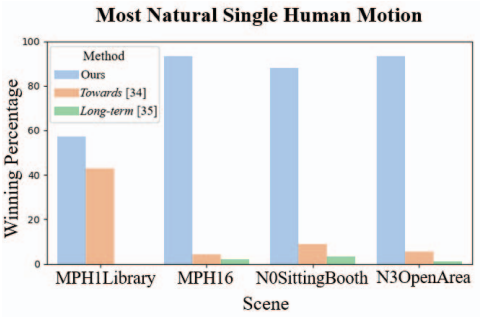


Figure 11: **User study on the naturalness of single-human motion.** Our method generates the most natural human motions.

does not differ significantly for the PROX dataset, but our method shows much better results for the Replica dataset, which has a much larger scale than the PROX dataset. It shows that our method is more scalable to other scenes. The highlighted region in Figure 10 shows that our framework generates more smooth human motions with less jittering in large-scale scenes compared to *long-term*.

We also provide the results of the user study on the visual comparisons in Figure 11. Similar to the multi-human motion, subjects are asked to choose the most natural result. Each subject evaluates four different test scenes from the PROX dataset. For each scene, three different motion sequences are presented (ours, *long-term*, and *towards*) with the same action sequence in the same scene. The results confirm that our method generates more natural motion than the other methods for all test scenes.

5. Conclusions

In this paper, we present MAMMOS, the multi-human motion generation framework that can populate natural and diverse motions of virtual humans. While it is challenging to simultaneously consider the dynamics of human-human interaction and human-scene interaction, we automate the motion generation only from a 3D scene and a simple list of action labels. MAMMOS gradually handles the complex spatio-temporal constraints in modularized stages, and effectively scales to large-scale 3D scenes with multiple people. The generated motions of the people not only improves the quality of individual people within the scene, but also allow them to naturally interact with collision-free path and appropriate interaction.

Acknowledgments This work was supported by Creative-Pioneering Researchers Program through Seoul National University and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023. Young Min Kim is the corresponding author.

References

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018.

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019.

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

[5] 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*, pages 1418–1427, 2018.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016.

[7] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

[8] Theodore P Grosvenor. *Primary care optometry*. Elsevier Health Sciences, 2007.

[9] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

[10] 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*, pages 2021–2029, 2020.

[11] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, pages 1961–1968. IEEE, 2011.

[12] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017.

[13] Edward T Hall. *The Hidden Dimension*. Bantam Doubleday Dell Publishing Group, New York, NY, July 1988.

[14] 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*, 4(2):100–107, 1968.

[15] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[16] 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.

[17] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

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

[19] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1272–1279, 2022.

[20] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020.

[21] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[22] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12368–12376, 2019.

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[25] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

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

[27] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

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

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

[30] 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*, pages 652–660, 2017.

[31] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[32] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[33] Fuwen Tan, Crispin Bernier, Benjamin Cohen, Vicente Ordonez, and Connelly Barnes. Where and who? automatic semantic-aware person composition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1519–1528. IEEE, 2018.

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[35] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.

[36] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022.

[37] 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*, pages 9401–9411, 2021.

[38] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[39] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651. IEEE, 2020.

[40] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021.

[41] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6194–6204, 2020.

[42] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022.

[43] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, , and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European conference on computer vision (ECCV)*, 2022.

[44] 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*, pages 5745–5753, 2019.