M.S. THESIS

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

3차원 장면 이해를 통한 다중의 3차원 사람 모션 생성

BY

정천기

FEBRUARY 2023

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

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

3차원 장면 이해를 통한 다중의 3차원 사람 모션 생성

지도교수 김 영 민

이 논문을 공학석사 학위논문으로 제출함

2023 년 1 월

서울대학교 대학원

전기정보공학부

정 천 기

정천기의 공학석사 학위논문을 인준함

2023 년 1 월

위 원 장	이 정 우
부위원장	김 영 민
위 원	주 한 별

Abstract

We present MAMMOS, which maps the motions of multiple humans that naturally interact with each other in 3D scene structure of choice. For the ultimate metaverse, one needs to create multiple characters interacting in response to the environment or other people. However, it is hard for an artist to generate multiple characters in diverse 3D scenes or gather training data to train an automated system that understands the entangled spatio-temporal contexts. MAMMOS is a modular approach that successfully handles complex constraints for realistic motion with nuances and intention. MAMMOS consumes a simple text input and first places anchors in time and location for individual characters that avoid collisions yet enable necessary interactions. Then we generate the spatio-temporal paths of multiple people 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 with multiple humans with rich interactions such that we can automatically populate the 3D scenes with realistic character motions.

Keywords: Motion Synthesis, Human-Scene Interaction, Metaverse

Student Number: 2021-24285

Contents

List of Figures

List of Tables

Chapter 1

Introduction

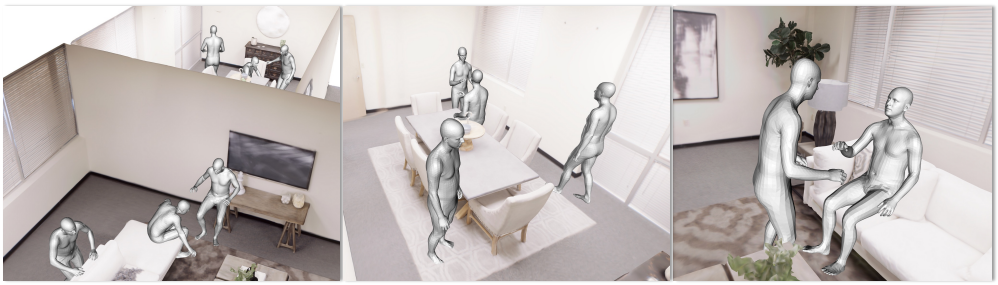


Figure 1.1 **MAMMOS** generates multiple human motions sharing a given indoor space. Generated virtual humans correctly understand the scene context around them, and they are also able to interact with each other. Our system automatically resolve the complex spatio-temporal constraints of the scene geometry and social interactions.

We consider the future of an interactive virtual world, where people in different places can collaborate within a shared space. The space can be a digital twin or a purely virtual asset. We aim to create an experience where the

characters interact with each other and respect the spatial context (Figure 1.1). Few works tackle the problem of generating diverse motions that consider both the scene context and mutual interactions between humans. Most importantly, it is hard to obtain training sequences for motions that interact with nearby objects and humans. A few works suggest placing humans in the scenes in either static poses for a given environment [18, 39, 41] or individual people without interacting with each other [38, 37]. People in the background usually perform a pre-defined sequence of motions created by an artist. This is one of the fundamental bottlenecks to expanding realistic interactions of everyday life to the virtual scene.

Our objectives are three-fold: input should be easy, and output should be diverse yet natural. 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 actions labels (sit, stand, lie, and interact). All the subsequent complexities are automated, given the 3D scene. We also aim for diverse output sequences given the semantic orders. Instead of replaying the same sequence of previously computed motion, we generate a rich set of plausible motion trajectories while they abide by the high-level user input. Most of all, the motion has to be natural. There are complex spatial and temporal constraints and intricate correlations between multiple humans and scene objects. We need to jointly consider all of them and make sure there is no penetration or violation of physical laws and still make eye contact and avoid collisions.

To overcome the complexity of the problem, we design a modular approach and utilize existing datasets for individual steps of the pipeline as shown in Figure 3.1. 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 spatio-temporal domain within the scene to best show the start and end of a given label of action and yet avoid collisions. Then a neural mapper can instantiate a set of plausible trajectories between the anchors from a diverse stochastic distribution. Following the the trajectory, we generate detailed motions for individual characters, where we can only consider local scene or interaction contexts assigned for the intermediate way points of the assigned path. By detaching the holistic analysis from detailed motion generation, we can utilize the prior works on motion generation without exhaustively 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, and diverse, yet 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.

Chapter 2

Related Works

Motion Synthesis Creating the motion of the human body has been investigated in various contexts. Given a frame or a sequence of frames, several works attempt to predict a subsequent sequence of motion [27, 13, 5, 3, 40]. In a practical aspect, many applications require generating natural motion between keyframes of static poses [8, 15]. Recent studies also suggest synthesizing motion without any explicit pose information, but rather from language[1, 2, 11, 28, 29, 4] or music [36, 20, 21, 22]. While these methods generate plausible motions without any skills for animating a character, they lack explicit control of the synthesized sequence. Recent methods observe specific action labels and generate a sequence of motion either with a recurrent unit [11] or a transformer [28]. Petrovich et al. [29] and Athanasiou et al. [4] go beyond a small set of action categories and generate diverse motions from free-form text. 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 [35] or 3D skeleton in RGB, RGB-D, or depth image [12, 23], recent works provide means to place a full 3D mesh of a human in a 3D scene utilizing high-quality parametric models [24, 32, 26]. Recent works utilize the 3D body model with expressive hands and faces [26] and place them on 3D scenes. Zhang et al. [39] observes the local scene context using explicit basis point sets (BPS) [30], while Hassan et al. [18] utilize contact probability between human and motion. A more recent method [41] proposes utilizing the distribution of the scene and poses using conditional variational autoencoder (CVAE) [33] given a training set containing both scenes and human bodies. 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 [38] or synthesizing motions of a given action label within the scene [37]. However, our approach jointly considers multiple humans within the scene, which harmoniously interact with the environment and other humans.

Chapter 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, $A^i = \{(a^i_1, c^i_1), (a^i_2, c^i_2)..., (a^i_{M_i}, c^i_{M_i})\}$. Specifically, the action label $a^i_j \in \{\text{stand}, \text{sit}, \text{lie}\}$ indicates the category of body pose in relation to the scene context. The interaction indicator $c^i_j \in \{0, \ldots, K\}$ represents whether the paired action a^i_j is an independent action of the character ($c^i_j = 0$) 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

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

Figure 3.1 **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.

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 [43], and $\theta_t^i \in \mathbb{R}^{32}$ represents the body pose in the form of VPoser [26].

The overall pipeline is depicted in Figure 3.1. 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 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 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 CVAE architecture [37] and place the posed character in appropriate translation r_j^i and rotation ϕ_j^i with an existing method of scene-aware placement 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. How-

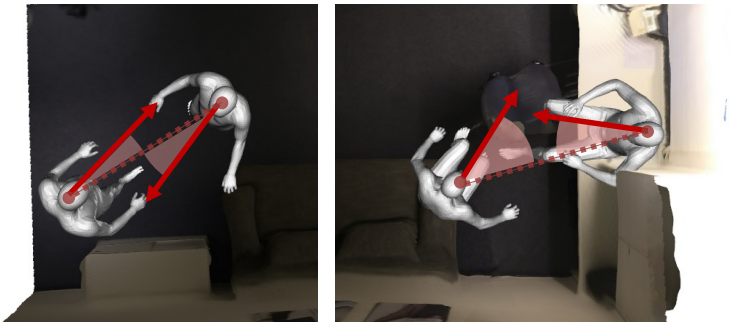


Figure 3.2 **An example of interaction anchor placement.** Considering each human as a point on a 2D overhead map, interaction anchors are placed so that the direction of the body does not deviate more than a threshold angle from the virtual line connecting two humans (dotted line).

ever, 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.

3.1.1 Interaction Anchor Placement

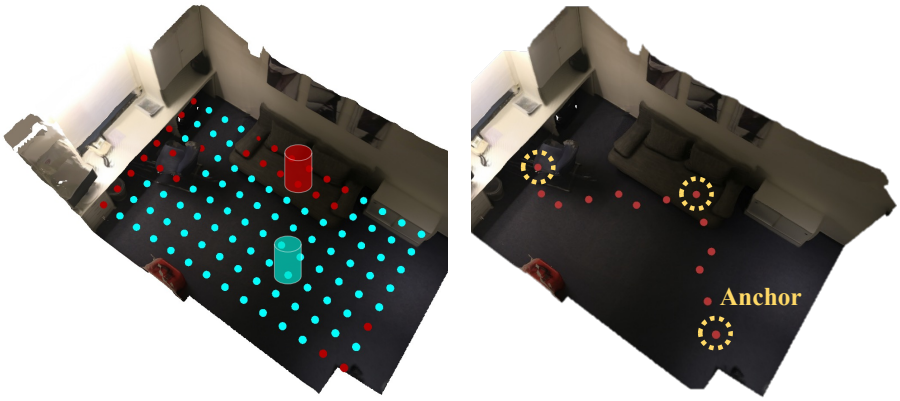
Because the interaction anchors require additional inter-human constraints, we first place interaction anchors ($c_j^i \neq 0$) and place the remaining ones using existing methods. In addition to the spatial context considered for normal anchors, the pair of action anchors with the same interaction indicator number (c_j^i) should be in an appropriate distance and angle to face each other as shown in Figure 3.2. 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) [10].

3.1.2 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 finalizes the coarse spatio-temporal context of the multiple interacting humans within the scene. 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 [38, 37, 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. We first generate spatial paths for individual humans and add an appropriate amount of idle time to satisfy the temporal constraints. For interaction anchors, we also allow a desired duration of stay at the anchor position to allow time to naturally interact with the paired person (Section 3.3).



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

Figure 3.3 **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.

3.2.1 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 3.3. For each grid, 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 $f(q) =$

$$\underbrace{g(q) + h(q)}_{A^*} + \underbrace{(1 - m(p, q))}_{\text{Neural Mapper [37]}} + \underbrace{C \cdot \mathbb{1}_{\text{collision}}(q, t+1)}_{\text{Collision Avoidance (Ours)}}, \tag{3.1}$$

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. [37] 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 trained neural network.

The third group of the terms in Equation (3.1) 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.

3.2.2 Timeline Integration

Timeline integration aligns the temporal windows of interaction anchors with the same indicator number. As individual path planning connects the generated anchors, the lengths of the intermediate paths are different, and the number of action labels is different to start with. As a result, the relative time steps of the interaction anchors do not align. Timeline integration fixes the temporal misalignments by adjusting the time stamps of the coarse grid paths. Specifically, we iterate to add necessary idle times to match the times for interaction and to check possible new collisions as illustrated in Figure 3.4. By adding an

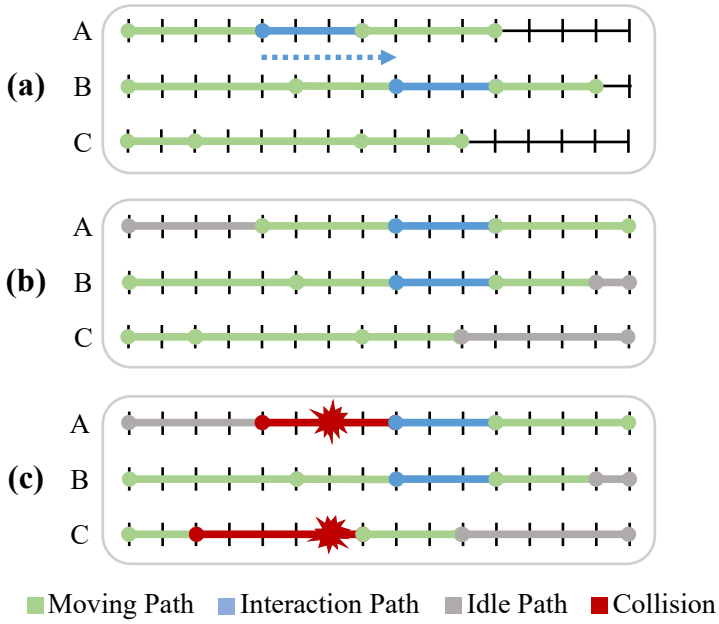


Figure 3.4 **Steps iterated for the timeline integration.** (a) The initial timelines before timeline integration. Note that interactions (blue) are not synched after independent path generation. To synchronize, the timeline of Human A should be shifted by the timestep indicated by the blue dotted-arrow. (b) Timelines after timeline integration. Timelines are adjusted so that all interactions are synced with their pair. By shifting timeline, we need to add idle time (gray) to fill the added empty temporal window. (c) Checking collisions. After each modification for the timeline integration, we check possible collisions for generated moving paths. Only subpaths with collision (red) are recreated, not the entire path.

idle time, the trajectory stays at the same spatial grid to avoid a collision or match the interaction time with another human subject. After every iteration, we regenerate the problematic subpaths subject to collisions.

3.3 Motion Completion

Motion completion creates frames of smooth motion that follows 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 [38, 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.

3.3.1 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 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).

Then we interpolate the keyframes by modifying the motion generation of [38] to encode local scene context. While previous works [38, 37] generate plausible motion and adapt to the scene context, their original works place one person at a time which are trained with 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 process the entire scene as an input. Instead, we train a framework only with

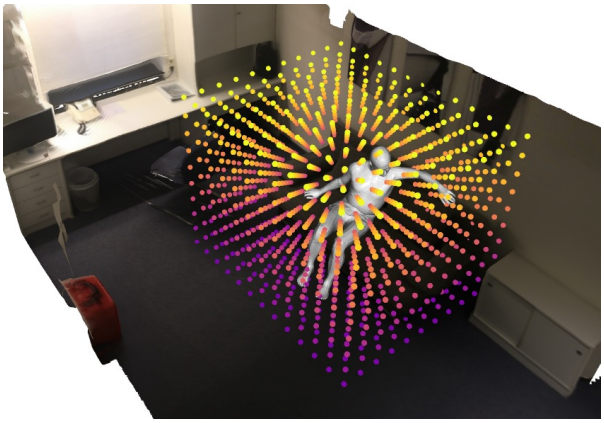


Figure 3.5 **An example of grid points where local SDF is retrieved.** Cube-shaped and human-centered grid points are used to encode local scene context (fewer grid points are depicted than are actually used).

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 3.5, 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.

3.3.2 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 en-

able different duration by employing RNN structure. Specifically, we employ GRU-based CVAE architecture used in the marker predictor of the GAMMA from [42]. It is trained with the TCD Hands dataset [25], which contains the SMPL-X parameters [26] 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.

3.3.3 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 [38] and self-penetration constraints from [6]. The full loss is presented in the supplementary material. Notably, we introduce novel eye contact optimization, which

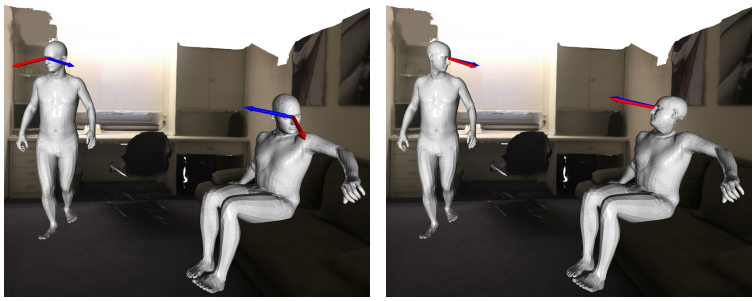


Figure 3.6 **An example of eye contact optimization.** Red arrows are the estimated current eye directions and the blue ones are the target eye directions to be matched to make eye contact.

plays a significant role in realism for natural interaction.

3.4.1 Eye Contact

While two people are interacting with each other, they need to make eye contact. We add the 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 3.6. Let's denote a vertex on the center of the forehead as v_f^k and the one on the back of the head as v_b^k for a human k. Assuming that human i and j are interacting with each other, our eye contact loss can be defined as below,

$$E_{\text{eye}} = \underbrace{\arccos \frac{e^i \cdot w^i}{\|e^i\| \|w^i\|}}_{\text{angle between } e^i \text{ and } w^i} + \underbrace{\arccos \frac{e^j \cdot w^j}{\|e^j\| \|w^j\|}}_{\text{angle between } e^j \text{ and } w^j} \tag{3.2}$$

where $e^i = v_f^i - v_b^i$ is the current eye direction of human i and $w^i = v_f^j - v_b^i$ 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.

Because what we want is to implement eye contact by minimal turning of the head without changing any existing rest poses, our optimization variable is not the VPoser embedded pose $\theta \in \mathbb{R}^{32}$, but the relative rotation between neck and head extracted from full body pose $\Theta \in \mathbb{R}^{63}$. However, unfortunately, direct optimization toward full body pose Θ may cause severe damage to the body shape. We add the pose prior loss [26] to the optimization constraint to avoid undesired deterioration and maintain valid body shapes.

Chapter 4

Experiments

Figure 4.1 **Qualitative results for multi-human motion.** We show sample results of multi-human motions with diverse interactions in different scenes. Please refer to the accompanying videos for the sequence of motions.

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 [31] in the original architecture to fully-connected layers. They are trained and tested with

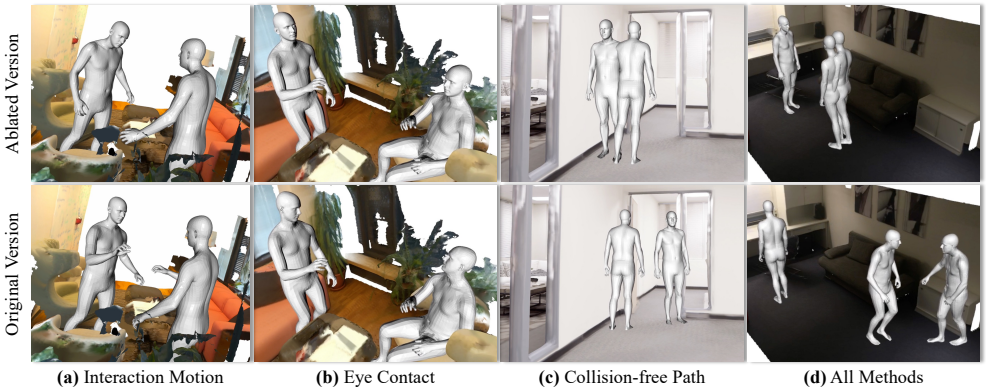


Figure 4.2 **The effect of each technical component.** MAMMOS implements critical components to create natural and physically plausible interactions between people. Please refer to the accompanying videos for the sequence of motions.

the same split of PROX dataset [17] as in [38, 39, 41, 37]. We also evaluate the generalization to larger scale 3D scenes with Replica dataset [34]. 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.

4.1 Multi-Human Motion

MAMMOS generates multi-human motion including natural interactions, which has not been addressed in previous literature. Since there is 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), eye contact optimization (-E). Users are asked to compare two versions of multi-human motions and

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 4.1 **User study results of multi-human motions.** We conducted an ablation study on methods used to generate multi-human motions and collected ratings on the naturalness of multi-human motions. The winning percentage of 'all' indicates the ratio of users who chose that the complete pipeline with all components is more natural. The relative score of 'all' is the mean and standard deviation of the scores that user provided in a scale from 1 to 5 (5 is more natural).

choose a preferred one. They make total 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 is summarized in Table 4.1. 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. Although they are all critical, the physical violation of collision-free paths appears especially noticeable. Sample frames of our multi-human motion are available in Figure 1.1 and 4.1, whereas the comparison against ablated versions is in Figure 4.2. The full videos of motion sequences, including the ones used for the user study, are available in the supplementary material.

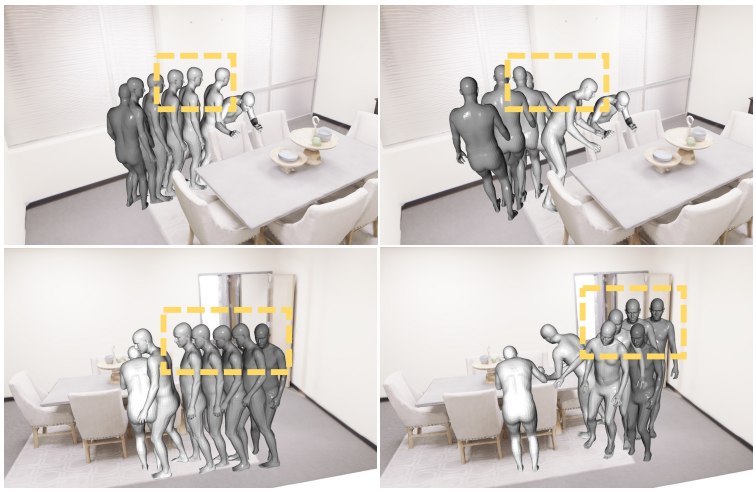
Method	Non-collision↑		Contact↑		Smoothness↑	
	PROX	Replica	PROX	Replica	PROX	Replica
Long-term [38]	94.58	99.39	95.04	90.89	96.41	93.28
Ours	**97.17**	**99.72**	**99.80**	**99.99**	**97.71**	**97.42**

Table 4.2 **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.

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*[38], and *towards* [37]. 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 4.2. The evaluation is presented in three metrics: the non-collision score, the contact score, and the smoothness score. The non-collision score and contact score are defined in [41, 38] to 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 the best non-collision and contact scores, generating physically plausible human motion in various scenes. The smoothness score evaluates the smoothness of the synthesized motion, and



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

Figure 4.3 **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.)

is defined as:

$$\text{score}_{\text{smooth}} = 1 - \frac{1}{T}\sum_{t=1}^{T}\|v_t - v_{t-1}\|_2$$

where v_t is the body vertices at frame t, and the jittering is measured by the mean of l_2 distances between the body vertices of consecutive frames. The smoothness score 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. Figure 4.3 provides the qualitative comparison of motion between ours and *long-term* in large-scale scenes. Note that highlighted region shows that ours framework generate more smooth human motions with less jittering.

We also provide the results of user study on the visual comparisons in Fig-

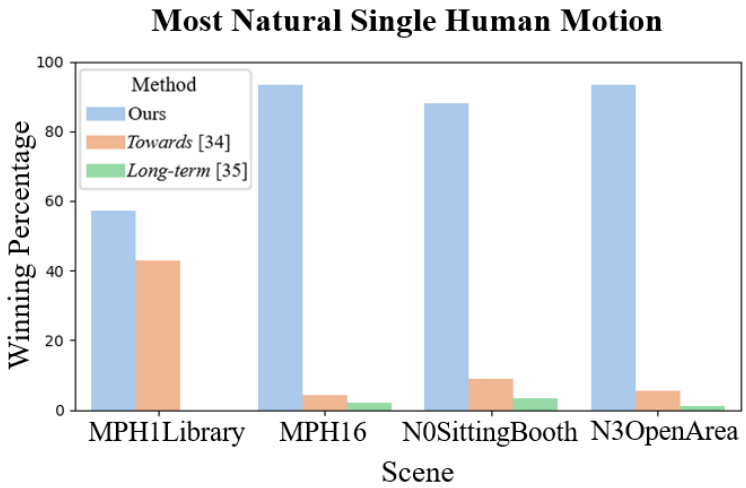


Figure 4.4 **User study on naturalness of single-human motion.** Our method generates the most natural human motions.

ure 4.4. 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. The videos of motion sequences are available in the supplementary material.

Chapter 5

Conclusion

In this paper, we present MAMMOS, the multi-human motion generation framework that can populate natural and diverse virtual humans into a given scene with a proper scene understanding. We simultaneously consider human-human interaction and human-scene interaction, which have not been addressed in other previous researches. With our framework, a user can easily create multi-human motions in large scale scenes with a simple text guidance. Our framework solves the complex spatio-temporal constraints in multi-human scenario by gradually resolving them in each modularized stage. By introducing essential elements in multi-human situations such as collision-free path, interaction motion, and eye contact, our framework can generate natural scenes where multiple people are interacting with each other in a given space. We improve the motion quality of individual humans, and nicely generalize the performance in a larger-scale scene by encoding the local scene context using SDF.

Chapter 6

Supplementary

MAMMOS creates motions of multiple humans within a 3D scene by a modular approach. In the supplementary material, we provide further details on how each module is implemented (Section 6.1) and how the naturalness of motion is evaluated (Section 6.2). Additionally, we also present more qualitative results (Section 6.3).

6.1 Implementation Details

6.1.1 Interaction Anchor Placement

For natural interaction between two people, the interaction anchors need to be close to each other, but at the same time should not be too close to collide. We set the distance between two interaction anchors to a value between 0.75m and 1.29m. In order for the interaction anchors to face each other, at least one of the following two conditions must be satisfied. The first condition is that the angles of the facing directions of the interacting humans with respect to the

line connecting the two interacting agents should be less than 30 degrees. The second condition is that the two rays of eye directions of the two agents meet at one point and the angles should be less than 60 degrees. The conditions are visualized in Figure 3 of the main paper.

6.1.2 Individual Path generation

As described in Section 3.2, our path generation module uses a modified A^* algorithm [14] where the scene-aware stochasticity and collision avoidance terms are added to the cost of standard A^* algorithm (Equation (1)).

For the scene-aware stochasticity, we identically implement CVAE model refered as Neural Mapper [37], except that the moving direction in a horizontal plane is encoded and decoded as below [19]

$$\text{Encode: } \theta \rightarrow (\sin\theta, \cos\theta)$$
$$\text{Decode: } (o_1, o_2) \rightarrow \arctan 2(o_1, o_2).$$

We use the sin/cos encoding to represent the orientation because it is continuous around 2π and therefore has better reconstruction property over their original $[0, 1]$ encoding. As with the original one, our Neural Mapper also expects the input of the moving direction and the local scene context to be encoded by BPS [30]. To train the Neural Mapper, we extract motion sequences of 30 frames with sufficient horizontal movement of the pelvis ($\geq$ 0.1m) from PROX dataset [17], and obtain the moving direction and the local scene context for each motion sequence. We set the moving direction as the direction from the start position to the end position of the motion sequence, and acquire local scene context by encoding scene vertices inside the $2m \times 2m \times 2m$ cube centered at the starting point of motion sequence using BPS with 10^4 basis points. Leveraging the trained Neural Mapper, we calculate feasiblility score $m(p, q)$ of

each neighbor grid q from current grid p as

$$m(p, q) = 1 - \frac{\alpha}{\pi},$$

where α $(0 \leq \alpha \leq \pi)$ is the shortest angle in radian between estimated moving direction and the direction from p to q.

And for the collision-avoidance, we additionally added tiny congestion-avoidance cost $\beta \sum_k e^{-d_k}$ to Equation (1), where d_k is the horizontal distance between the grid point q and human k's position at timestep $t + 1$ and β is a manually set constant considering the interval between grids. We use 5 for the β. The congestion-avoidance cost does not work as the main factor of our path finding algorithm, but it makes our algorithm slightly prefer the direction that is distant from other existing humans.

6.1.3 Timeline Integration

To reduce the complexity of timeline integration (Section 3.2), we gradually align the temporal windows of interactions, focusing on one interaction pair at each iteration. In every iteration, we find unmatched interaction pair with the lowest indicator value, and shift the timeline that is earlier than the other by adding idle paths in such a way as to avoid possible collisions. But sometimes, there may be no possible cases to synchronize interaction without any collisions no matter how the timeline is shifted. In that case, we just sync interaction first, then regenerate the problematic subpaths in the shifted timeline as shown in Figure 5-(c). Discordance of length between regenerated and previous subpath can be handled by either augmenting the idle path or additional iteration. Such process is repeated until all interaction pairs are synchronized and no collisions occur along the entire paths.

6.1.4 Moving Motion

When we create generalizable scene-aware moving motion in Section 3.3, we leverage the local SDF as the human-centric local scene context. To acquire the local SDF, we create a $30 \times 30 \times 30$ cube-shaped grid centered at human's pelvis location, as shown in Figure 6, then gather the corresponding SDF value at each grid point and concatenate them. As a network architecture, we use modify the RouteNet $\mathcal{R}$ and PoseNet $\mathcal{P}$ proposed in [38]. Since we use the local SDF for scene context instead of the global scene point cloud, we replace PointNet [31] in $\mathcal{R}$ and $\mathcal{P}$ with fully-connected layers so that our local SDF is encoded through the total 3 fully-connected layers with 512, 256, and 256 hidden dimensions. Our modified $\mathcal{R}'$ and $\mathcal{P}'$ are trained to interpolate moving motion in the local space where the origin is fixed to the root position of the starting keyframe. We use 30-frame motion sequences from the PROX dataset transformed to the local space and the local SDF calcaulated based on the original human position of the starting frame. Other omitted training details are the same as those of the original paper [38].

6.1.5 Interaction Motion

For interaction, we generate the upper body motions and combined with any sitting or standing anchors. As a dataset for upper body interaction motion, we use the SMPL-X fitted TCDHands dataset [25]. We firstly filter out the action sequences that are not interacting gestures; The final action sequences used are: "bottle", "counting", "direction", "finger", "grasp", "object", "ok", "pointing", "sign", "talking", "tposefinger", and "v". Then, we extract the upper body motions from the filtered dataset according to the following equation obtained

using the joint map of SMPL-X body model.

$$\Theta_u = \Theta_{6:9} \parallel \Theta_{15:18} \parallel \Theta_{24:27} \parallel \Theta_{33:63}$$

In here, Θ_u is the upper body pose parameter, $\Theta \in \mathbb{R}^{63}$ is the full body pose parameter, $\parallel$ is the concatenation operator, and the interval is expressed in right-open form with 0-based indexing. As for a network, we use the GRU-based CVAE architecture, the same architecture used with the Marker Predictor of the GAMMA [42], but with the dimensions of layers reduced to adapt to the smaller size of the TCDHands dataset and one additional 3-layer MLP inserted to condition input path as same with GRU's initial hidden state path. The exact dimensions are listed in the Table 6.1 below.

	Ours	Marker Predictor [42]
GRU	128	256
MLP (encoder/decoder)	[256, 256]	[512, 256]
MLP (condition)	[256, 256, 128]	[512, 256, 256]
Latent z	32	128

Table 6.1 **Comparison of the layer dimensions between ours and the original Marker Predictor**

We configure our interaction motion CVAE to take 2-frame motion history as condition, and output a 10-frame motion primitive that are smoothly continuing from the conditioned motion history. We train the model with the identical training loss and settings proposed in their original paper, but we apply cyclical KL annealing [9] to the KL-divergence loss term instead of the robust function [7] $\Psi(s) = \sqrt{1+s^2} - 1$ since it was experimentally found that the cyclical KL annealing produces more diverse motion outputs. For cyclical

KL annealing, we use cyclical cosine scheduling of total 2 cycles and half ratio per cycle to increase the weight of KL-divergence loss.

6.1.6 Optimization

We provide the full implementation details on our understated optimization losses here.

Foot Location Loss [38] We use the foot location loss to minimize foot sliding when a human walks. The foot location loss is defined as below:

$$E_{\text{foot}} = \sum_{s \in S} \mathbb{E}_{t \in s}(\|v_t^f - \overline{v}_s^f\|_2)$$

where S is a set of subsequences [38] divided based on the stable foot when a human walks, v_t^f is the stable foot vertices at frame t and $\overline{v}_s^f$ is the mean stable foot vertices of the subsequence s.

Penetration Loss [38] The penetration loss is defined as below:

$$E_{\text{pene}} = \sum_{t=0}^{T} \mathbb{E}(|\Psi_{\text{sdf}}^{-}(v_t)|)$$

where $|\Psi_{\text{sdf}}^{-}|$ returns the absolute SDF value of points with the negative SDF values (points where penetration occured) and v_t is the body vertices at frame t.

Contact Loss [38] The contact loss is defined as below:

$$E_{\text{contact}} = \sum_{t=0}^{T} \sum_{v_t^c \in v_{\text{contact}}} \min_{v^s \in v_{\text{scene}}} \rho(\|v_t^c - v^s\|_2),$$

where v_{contact} is the predefined set of body vertices [17] where the contact with the scene is encouraged, v_{scene} is the set of scene vertices, and ρ is the

Geman-McClure error function that reduces the weight of v_t^c that are far from v^s.

Smoothness Loss [38] The smoothness loss is defined as below:

$$E_{\text{smooth}} = \sum_{t=1}^{T} \|v_t - v_{t-1}\|_2,$$

where v_t is the body vertices at frame t.

Self-Penetration Loss [6] We estimate the self-penetration of the upper body during the interaction motion unlike previous works that consider the entire body [6]. We approximate the occupied volumes of the two forearms and thighs with individual cylinders that bound the volumes, which are in turn approximated with a set of spheres. Specifically, the self-penetration loss is defined as below:

$$E_{\text{self-pene}} = -\sum_{t=0}^{T} \sum_{i \in S} \sum_{j \in I(i)} \exp\left(\frac{\|c_t^i - c_t^j\|_2}{r_i^2 + r_j^2} \right)$$

where S is a set of spheres approximating cylinders, $I(k)$ is the set of spheres overlapped with sphere k while belonging to another cylinder, c_t^k is the center of sphere k at frame t, r_k is the radius of sphere k.

Pose Prior Loss [17] In optimizing the eye contact, we additionally use the pose prior loss to penalize impossible neck rotations. The pose prior loss is defined as below:

$$E_{\text{pose-prior}} = \sum_{t=0}^{T} \|\theta_t\|_2$$

where $\theta_t \in \mathbb{R}^{32}$ is a VPoser embedded pose parameter at frame t.

6.2 Naturalness Evaluation

We provide more details about the modified non-collision score, contact score, and user study in this section.

6.2.1 Modified Non-collision Score and Contact Score

Unlike [41], we give a margin of 0.01 for a signed distance value of 0 for both contact and non-collision. In our case, we take it as contact when the signed distance value is less than 0.01 for the contact score and non-collision when the signed distance value is greater than -0.01 for the non-collision score.

6.2.2 User Study

For single human motion, we give 3 examples (ours, *long-term* [38], *towards* [37]) with the same inputs and ask the users to choose the most natural and most unnatural that interpolates motion between the start and end anchors. We also ask users to rate on a scale of 1-5 on how much the most natural is more natural than the second, and the same questions are asked about the unnatural. Table 6.2 shows the comparison result of how much more natural each rank is. For multi-human motion, using our method and using ablated versions, we ask which one is more natural and ask to rate how much more natural it is on a scale of 1 to 5.

6.3 Qualitative Results

6.3.1 Collision-free Path Generation

Two examples of collision-free path generation are presented in Figure 6.1 and 6.2. Our modified A^* algorithm can generate plausible, yet collision-free paths considering both spatial and temporal context.

Rank(A>B>C)	Natural(A-B)	Unnatural(C-B)	Number of samples
Ours>*Towards*>*Long-term*	3.41	3.59	220
Ours>*Long-term*>*Towards*	3.54	3.38	59
Towards>Ours>*Long-term*	2.57	3.53	49

Table 6.2 The table shows the scores(1-5) for how natural 1st place compared to 2nd place is(A-B), and how unnatural 3rd place is compared to 2nd place(C-B) for users who select the corresponding rank. Only the rank selected by 10% or more among all users are shown.

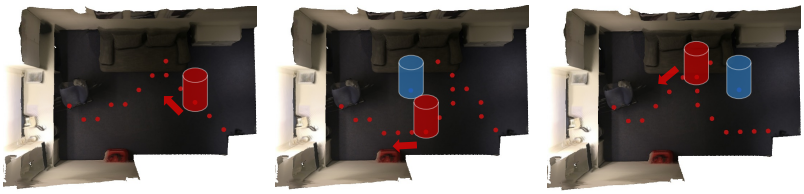


Figure 6.1 **Collision-free path planning example.** Our collision-avoidance term (Equation (1)) enables A^* algorithm to generate collision-free paths. The leftmost figure shows the original path. Middle and right show the modified paths (red) for the same anchors to avoid the standing human (blue).

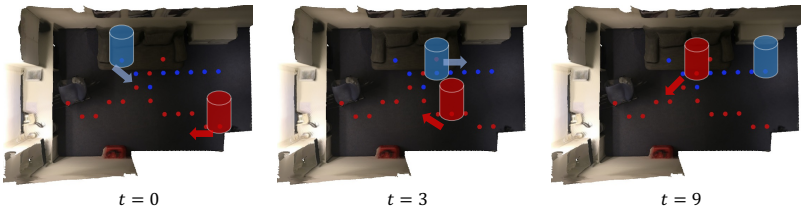


Figure 6.2 **Another collision-free path planning example.** When creating a collision-free path, we consider both spatial and temporal contexts. Please note that while the red and blue paths collide, if we only consider the spatial context, but there is no collision when we jointly consider the temporal context.

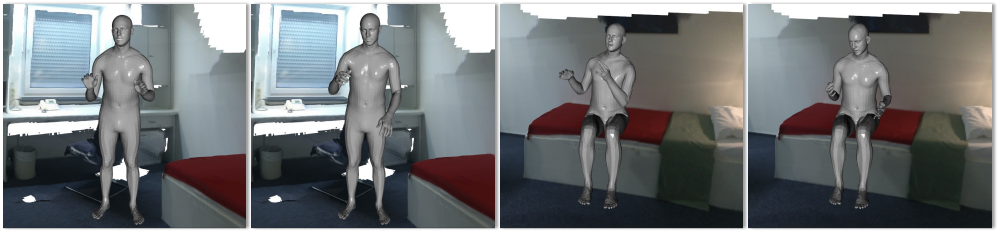


Figure 6.3 **Interaction motion examples.** Our framework generates an upper body interaction motion applicable to both sit and stand anchors. Various interaction motions can be generated from the same anchor pose.

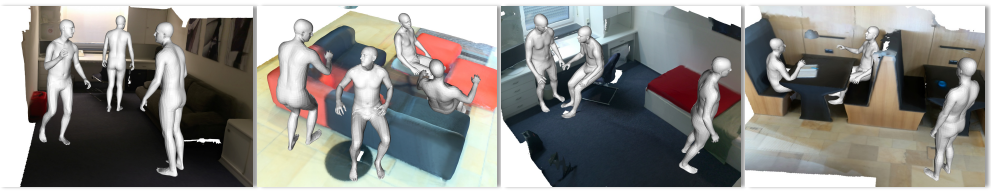


Figure 6.4 **Qualitative result of multi-human motions in diverse scenes.** Our framework generates diverse scenarios where multiple human interact with each other in various scenes.

6.3.2 Interaction Motion

Figure 6.3 shows sample frames of interaction motion derived from the stand and sit anchor. Our framework is capable of generating various interaction motions from the same anchor pose and expresses plausible hand gestures that would actually be seen when people are interacting with each other.

6.3.3 Results in Diverse Scenes

More sample frames of our final results from various scenes are presented in Figure 6.4. As presented, our framework is capable of generating multi-human motion with diverse scenarios in various scenes.

Bibliography

[1] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. 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] C. Ahuja and L.-P. Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.

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

[4] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol. Teach: Temporal action composition for 3d humans. *arXiv preprint arXiv:2209.04066*, 2022.

[5] E. Barsoum, J. Kender, and Z. 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] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. 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] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172 vol.2, 1994.

[8] Y. Duan, T. Shi, Z. Zou, Y. Lin, Z. Qian, B. Zhang, and Y. Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

[9] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

[10] T. P. Grosvenor. *Primary care optometry.* Elsevier Health Sciences, 2007.

[11] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[12] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, pages 1961–1968. IEEE, 2011.

[13] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017.

[14] P. E. Hart, N. J. Nilsson, and B. 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] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[16] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, Oct. 2021.

[17] M. Hassan, V. Choutas, D. Tzionas, and M. 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] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. 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] A. H. (https://stats.stackexchange.com/users/119623/ari herman). Encoding angle data for neural network. Cross Validated. URL:https://stats.stackexchange.com/q/218407 (version: 2018-09-14).

[20] B. Li, Y. Zhao, S. Zhelun, and L. 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.

[21] J. Li, Y. Yin, H. Chu, Y. Zhou, T. Wang, S. Fidler, and H. Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020.

[22] R. Li, S. Yang, D. A. Ross, and A. 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.

[23] X. Li, S. Liu, K. Kim, X. Wang, M.-H. Yang, and J. 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.

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[25] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[26] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. 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.

[27] D. Pavllo, D. Grangier, and M. Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018.

[28] M. Petrovich, M. J. Black, and G. 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.

[29] M. Petrovich, M. J. Black, and G. Varol. Temos: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109*, 2022.

[30] S. Prokudin, C. Lassner, and J. 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.

[31] C. R. Qi, H. Su, K. Mo, and L. 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.

[32] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[33] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

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

[35] F. Tan, C. Bernier, B. Cohen, V. Ordonez, and C. 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.

[36] G. Valle-Pérez, G. E. Henter, J. Beskow, A. Holzapfel, P.-Y. Oudeyer, and S. Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.

[37] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. 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.

[38] J. Wang, H. Xu, J. Xu, S. Liu, and X. 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.

[39] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. 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] Y. Zhang, M. J. Black, and S. 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] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. 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] Y. Zhang and S. 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] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. 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.

초록

우리는 주어진 3차원 장면 구조에서 자연스럽게 상호작용하는 여러 사람의 움직임을 매핑하는 프레임워크인 MAMMOS를 제시한다. 궁극적인 메타버스를 위해서는 환경이나 다른 사람들에 반응하여 상호 작용하는 여러 캐릭터를 만들 수 있어야 한다. 그러나 아티스트가 다양한 3차원 장면에서 여러 캐릭터를 생성하거나 얽힌 시공간적인 문맥을 이해하는 자동화 시스템을 훈련하기 위한 훈련 데이터를 수집하는 일에는 어려움이 따른다. MAMMOS는 뉘앙스와 의도가 있는 사실적인 모션에 대한 복잡한 제약 조건을 성공적으로 처리하는 모듈식 접근 방식이다. MAMMOS는 간단한 텍스트 입력으로 받아 개별 사람에 대하여 충돌을 피하면서도 필요한 상호 작용을 가능하게 하는 시간과 위치에 앵커를 먼저 배치한다. 그런 다음 장면 내에서 여러 사람의 시공간 경로를 생성하고 연결하여 다양하고 자연스러운 동작을 수행하게 한다. 우리가 아는 한, MAMMOS는 풍부한 상호작용을 갖는 여러 사람의 동작 시퀀스를 생성해 내는 최초의 프레임워크이다. MAMMOS를 통해 3D 장면에 사실적인 여러 사람의 모션을 자동으로 채울수 있게 되었다.

주요어: 모션 합성, 사람-장면 상호작용, 메타버스

학번: 2021-24285