

# Supplementary Materials for MAMMOS: Mapping Multiple human MOTion with Scene understanding and natural interactions

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 A) and how the naturalness of motion is evaluated (Section B). Additionally, we also present more qualitative results (Section C) and discuss the limitations of our work (Section D).

## A. Implementation Details

### A.1. Anchor Placement

At the first stage of the pipeline, we generate poses corresponding to the input action labels. Then we place the posed bodies in the given scene with an approach described in POSA [6] without any modification. We do not provide the details of POSA here, but the process of putting humans in scene is described in Section 3.1 of the main paper.

#### A.1.1 Interaction Anchor Placement

Among the anchors with action labels, interaction anchors need a more careful arrangement. For natural interaction between two people, the interaction anchors need to be close to each other, but not to the point where they 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.

### A.2. Path Generation

#### A.2.1 Individual Path Generation

As described in Section 3.2, our path generation module uses a modified  $A^*$  algorithm [4] where the scene-aware

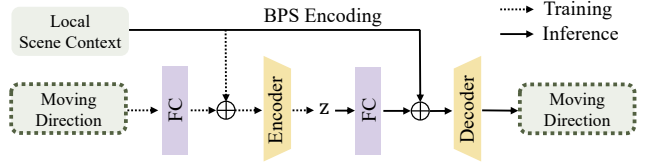


Figure S1: Architecture of Neural Mapper. Changed components from their origin [12] are highlighted with dotted border.

stochasticity and collision avoidance terms are added to the cost of the standard  $A^*$  algorithm (Equation (1)).

For the scene-aware stochasticity, as shown in Figure S1, we identically implement the CVAE model referred as Neural Mapper [12], except that the moving direction in a horizontal plane is encoded and decoded as below [7]

$$\text{Encode: } \theta \rightarrow (\sin \theta, \cos \theta)$$

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

We use the  $\sin/\cos$  encoding to represent the orientation because it is continuous around  $2\pi$  and therefore has better reconstruction property over their original  $[0, 1]$  encoding. As with the original one, our Neural Mapper also expects the input, which is the local scene context, to be encoded by BPS [10]. To train the Neural Mapper, we extract motion sequences of 30 frames with sufficient horizontal movement of the pelvis ( $\geq 0.1\text{m}$ ) from PROX dataset [5], 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  $2\text{m} \times 2\text{m} \times 2\text{m}$  cube centered at the starting point of the motion sequence using BPS with  $10^4$  basis points. Leveraging the trained Neural Mapper, we calculate feasibility score  $m(p, q)$  of each neighbor grid  $q$  from current grid  $p$  as

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

where  $\alpha$  ( $0 \leq \alpha \leq \pi$ ) is the shortest angle in radian between

the 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  $\beta$  is a manually set constant considering the interval between grids. We use 5 for the  $\beta$ . 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.

### A.2.2 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 the regenerated and previous subpath can be handled by either augmenting the idle path or additional iteration. Such a process is repeated until all interaction pairs are synchronized, and no collisions occur along the entire path.

## A.3. Motion Completion

### A.3.1 Moving Motion

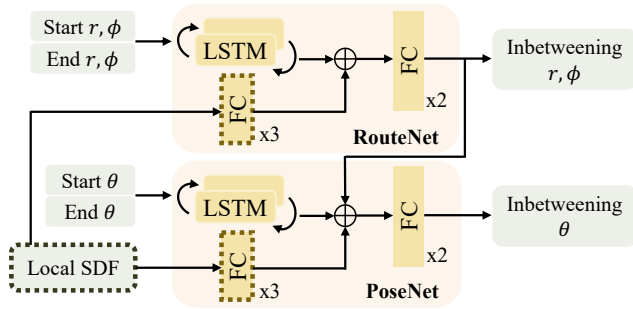


Figure S2: Architecture of moving motion synthesis network. Boxes with dotted lines indicate components modified from their original implementation in [13].

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. For the network architecture, as shown in Figure S2, we borrow RouteNet  $\mathcal{R}$  and PoseNet  $\mathcal{P}$  proposed in [13]. Since we use the local SDF for scene context instead of the global scene point cloud, we replace PointNet [11] 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 calculated based on the original human position of the starting frame. Other omitted training details are the same as those of the original paper [13].

### A.3.2 Interaction Motion

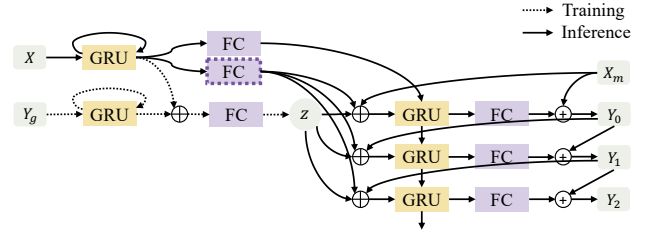


Figure S3: Architecture of interaction motion synthesis network. Changed components from their origin [15] are highlighted with dotted border.

For interaction, we generate the upper body motions and combine them with any sitting or standing anchors. As a dataset for upper body interaction motion, we use the SMPL-X [9] fitted TCDHands dataset [8]. 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,  $\Theta_u$  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. The network architecture for synthesizing interaction motion is shown in Figure S3. We use the GRU-based CVAE architecture, which is the same architecture used with the Marker Predictor of the GAMMA [15]. However, we reduce the dimensions of layers to adapt to

the smaller size of the TCDHands dataset and also insert an additional 3-layer MLP to the conditional input path (highlighted one in Figure S3), as same with GRU’s initial hidden state path. The changed dimensions are listed in Table S1 below.

	Ours	Marker Predictor [15]
GRU	128	256
MLP (encoder/decoder)	[256, 256]	[512, 256]
MLP (conditional input)	[256, 256, 128]	[512, 256, 256]
Latent $z$	32	128

Table S1: 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 a condition, and output a 10-frame motion primitive that is 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 [3] to the KL-divergence loss term instead of the robust function [2]  $\Psi(s) = \sqrt{1 + s^2} - 1$  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.

#### A.4. Optimization

We provide the full details on losses incorporated for the optimization step.

**Foot Location Loss [13]** We use the foot location loss to minimize foot sliding when a human walks. The foot location loss is defined as below:

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

where  $S$  is a set of subsequences [13] divided based on the stable foot when a human walks,  $v_t^f$  is the stable foot vertices at frame  $t$  and  $\bar{v}_s^f$  is the mean stable foot vertices of the subsequence  $s$ .

**Penetration Loss [13]** 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 occurred), and  $v_t$  is the body vertices at frame  $t$ .

**Contact Loss [13]** 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  $v_{\text{contact}}$  is the predefined set of body vertices [5] where the contact with the scene is encouraged,  $v_{\text{scene}}$  is the set of scene vertices, and  $\rho$  is the Geman-McClure error function that reduces the weights of  $v_t^c$  that are far from  $v^s$ .

**Smoothness Loss [13]** 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 [1]** We estimate the self-penetration of the upper body during the interaction motion, unlike previous works that consider the entire body [1]. 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 [5]** 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  $t$ .

## B. Naturalness Evaluation

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

### B.1. Modified Non-collision Score and Contact Score

Unlike [14], 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.

## B.2. Smoothness Score

The smoothness score evaluates the smoothness of the synthesized motion, and 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.

## B.3. User Study

For single human motion, we give 3 examples (ours, *long-term* [13], *towards* [12]) 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 S2 shows the comparison result of how much more natural each rank is. For multi-human motion, using our method and 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.

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 S2: 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.

## C. Qualitative Results

### C.1. Collision-free Path Generation

Two examples of collision-free path generation are presented in Figure S4 and S5. Our modified  $A^*$  algorithm can generate plausible, yet collision-free paths considering both spatial and temporal contexts.

### C.2. Interaction Motion

Figure S6 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.

### C.3. Results in Diverse Scenes

More sample frames of our final results from various scenes are presented in Figure S7. As presented, our framework is capable of generating multi-human motion with diverse scenarios in various scenes.

## D. Limitations

While MAMMOS can generate diverse natural motions to imitate human-human interaction, the interaction motion and eye contact still have room for improvement. For example, humans in the real world usually interact by alternating talking and listening and may not constantly stare at each other during an interaction. Incorporating social context and nuances in response to other people can significantly enhance the realism of conversation or social interaction within the scene.

The sequential path generation pipeline limits the diversity of paths for those generated later in the order. Instead of concurrently generating the trajectories, we handle the complex spatio-temporal constraints of multiple people by generating one path at a time. When the scene size is small, collision avoidance against previous paths can impose a severe restriction on succeeding paths. The resulting paths can be unnatural, or it can even be impossible to create collision-free paths without modifying the anchors.

## References

- [1] 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. 3
- [2] 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. 3
- [3] Xiaodong Liu Jianfeng Gao Asli Celikyilmaz Lawrence Carin Hao Fu, Chunyuan Li. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *NAACL*, 2019. 3
- [4] 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. 1
- [5] 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. 1, 3
- [6] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of*



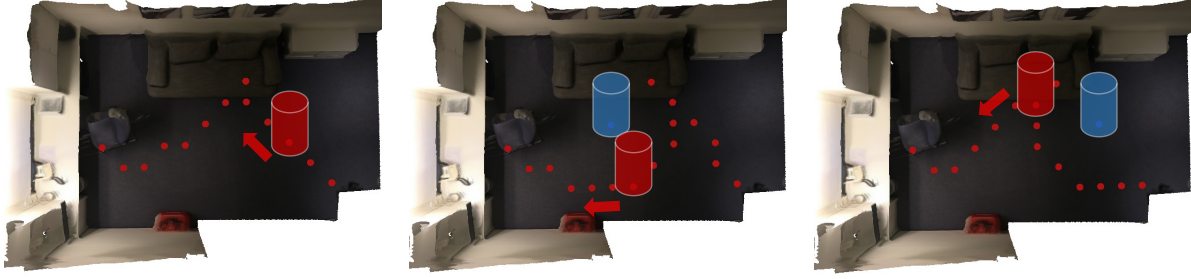


Figure S4: **Collision-free path planning example.** Our collision-avoidance term (Equation (1)) enables  $A^*$  algorithm to generate collision-free paths. The leftmost figure shows the original path. Middle and right show the modified paths (red) for the same anchors to avoid the standing human (blue).

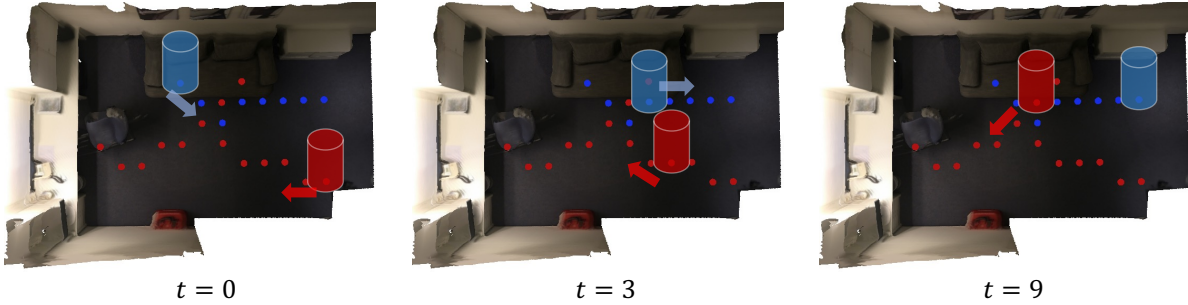


Figure S5: **Another collision-free path planning example.** When creating a collision-free path, we consider both spatial and temporal contexts. Please note that while the red and blue paths collide, if we only consider the spatial context, but there is no collision when we jointly consider the temporal context.

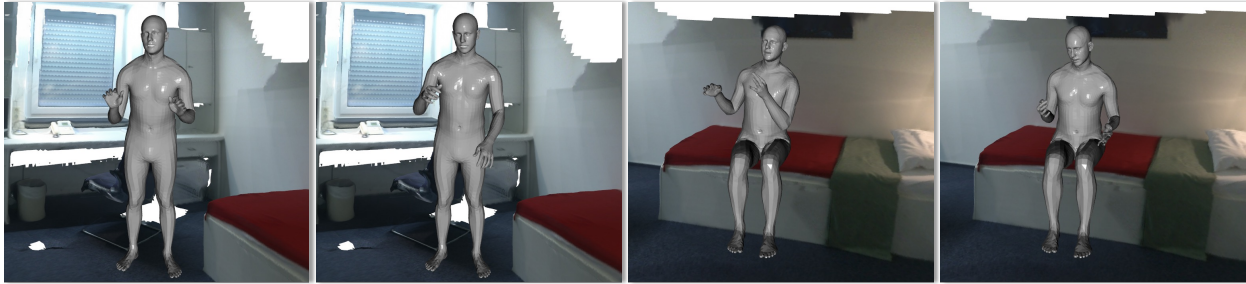


Figure S6: **Interaction motion examples.** Our framework generates an upper body interaction motion applicable to both sit and stand anchors. Various interaction motions can be generated from the same anchor pose.

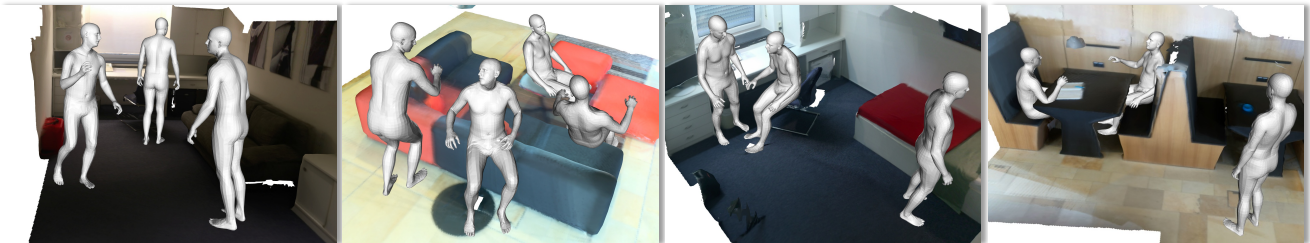


Figure S7: **Qualitative result of multi-human motions in diverse scenes.** Our framework generates diverse scenarios where multiple humans interact with each other in various scenes.

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 1

- [7] Ari Herman (<https://stats.stackexchange.com/users/119623/ariherman>). Encoding angle data for neural network. Cross Validated. URL:<https://stats.stackexchange.com/q/218407> (version: 2018-09-14). 1
- [8] 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. 2
- [9] 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. 2
- [10] 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. 1
- [11] 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. 2
- [12] 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. 1, 4
- [13] 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. 2, 3, 4
- [14] 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. 3
- [15] 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. 2, 3