MIME: Human-Aware 3D Scene Generation
-Supplemental Document-

Hongwei Yi[1] Chun-Hao P. Huang[2*] Shashank Tripathi[1] Lea Hering[1] Justus Thies[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Adobe Inc.
{firstname.lastname}@{tuebingen.mpg.de} chunhaoh@adobe.com

1. Training Details

We demonstrate more technical details in this section. For data representation, we project the foot vertices onto a 2D mask of fixed size (e.g. 64x64 resolution for a bedroom of size 6.3m x 6.3m). The dimension of free space feature F is 64. The category embedding λ is a matrix of size $K \times 64$, that stores a per-object category vector, for all K object categories in the dataset. The position encoding $p(\cdot)$ is:

$$p(x) = \left(\sin\left(2^0\pi x\right), \cos\left(2^0\pi x\right), \ldots, \sin\left(2^{L-1}\pi x\right), \cos\left(2^{L-1}\pi x\right)\right),$$

where x can be any attribute t, r, s and $L = 32$.

During training, we apply the Adam optimizer [1] with learning rate $1e^{-4}$ and no weight decay. In Adam optimizer, we use the default PyTorch implemented parameters, i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$. We train MIME with the batch size 128 for $100k$ iterations. We perform random global rotation augmentation between $[0, 360]$ degrees on the holistic populated scene, including the floor plane, all objects, the free space and all contact humans.

During inference, MIME can handle not just motion data, but also multiple static people. We use NMS to eliminate redundant contact information when a person stays in one position for a while (e.g. sitting on a chair). Potential future work could include synthesizing human motion with temporal contact in 3D scenes [3] and incorporating this information into our method.

2. Ablation Study on Various Size of Floor Plans

In Fig. S.1 a), MIME generates a large sofa that accommodates multiple sitting poses. In Fig. S.1 b), we vary the size of the floor plan. Larger floor plans result in more objects generated given the same input motion. The average number of objects generated per room category is 7.1 for bedrooms, 17.5 for living rooms, 6.2 for libraries, and 15.1 for dining rooms. If the input motion sequence is repeated, the encoded 2D free space and human contact information will be the same as the original input, so MIME will not add new objects in the same place. The generated objects are represented as 3D oriented bounding boxes, so there is no symmetry information for each object. Since MIME conditions the generated object on the input humans *and existing objects*, it does consider the group relationship between them.

3. More Discussion

MIME does not support object-on-object generation because the 3D FRONT dataset does not have small objects on big furniture like cups or lamps. This is an interesting future direction, going beyond methods like ATISS. And adding physics to MIME would be interesting future work.

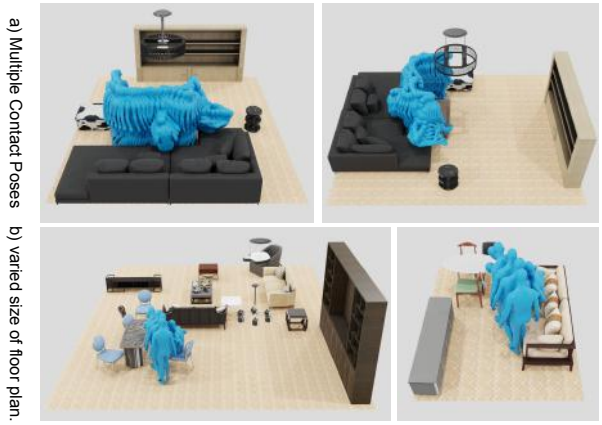

Figure S.1. Ablation study on various size of floor plan . MIME can generate more objects generated with larger floor plans and the same input motion.

4. More Qualitative Examples

We present more qualitative examples for different kinds of rooms, in Fig. S.2, Fig. S.3, and Fig. S.4. Compared with our baseline methods [2], our method can generate more plausible 3D scenes that input motions can interact with.

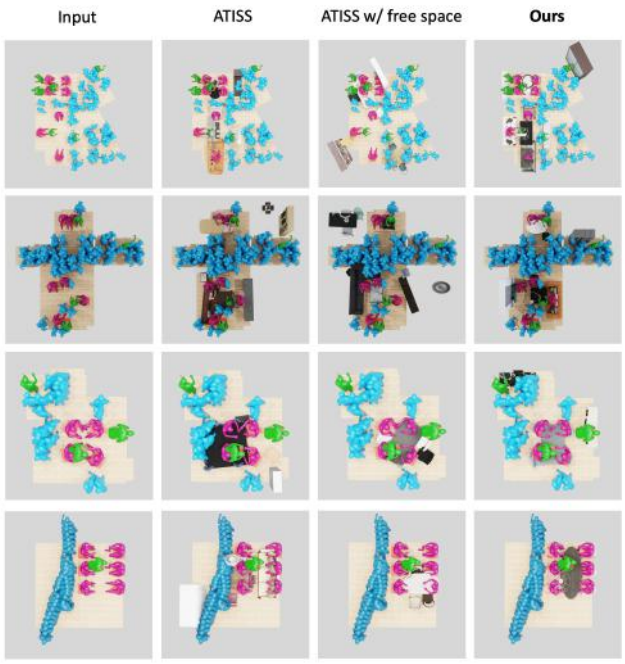

Figure S.4. Qualitative comparison on living rooms (the first two rows) and dining rooms (the last two rows) in the test split of *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects. We also show the original ATISS w/ or w/o the free space mask as input. All results are w/o refinement. Each row represent an example input.

Figure S.2. Qualitative comparison on bedrooms in the test split of *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects. We also show the original ATISS w/ or w/o the free space mask as input. All results are w/o refinement. Each row represents an example input.

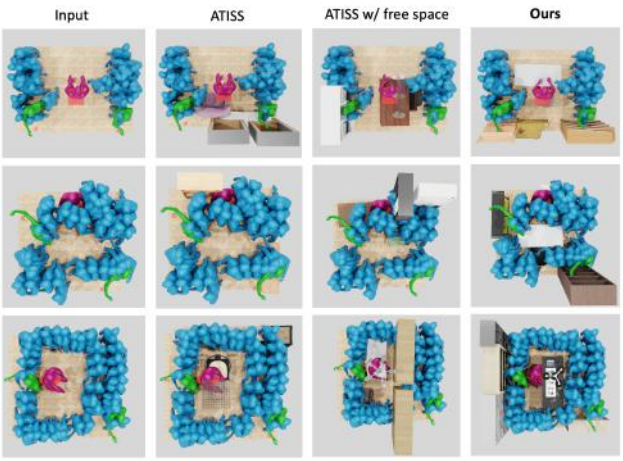

Figure S.3. Qualitative comparison on libraries in the test split of *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects. We also show the original ATISS w/ or w/o the free space mask as input. All results are w/o refinement. Each row represent an example input.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 1

[2] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1

[3] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia Conference Papers*, pages 1–9, 2022. 1