석사 학위 논문 심사 발표

확산 모델을 활용한 사람 모션 기반 주변 사물 생성 모델
Human Motion-based Interacting Object Generation using Diffusion

2024.12.16

Neuro-Machine Augmented Intelligence Lab (NMAIL)

전윤호 (Younho Jeon)

Advisor: 조성호 교수님

Committee: 김태균 교수님, 최성희 교수님

NMAIL

Introduction

Tesla

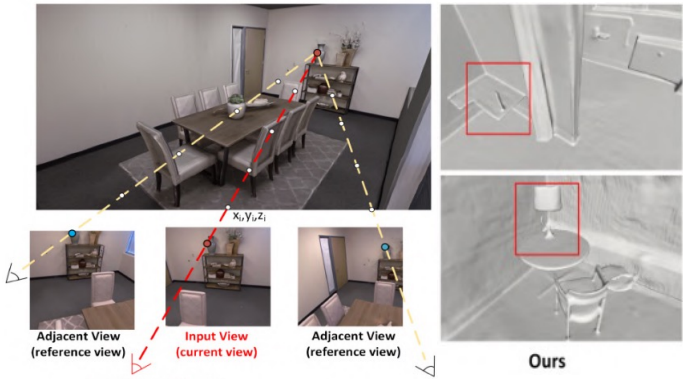
NVIDIA

Humane Ai Pin

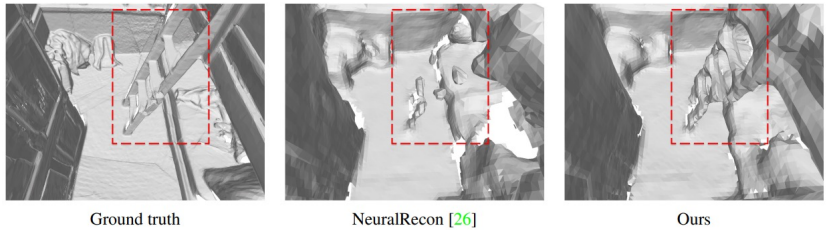
Apple Vision Pro

Meta

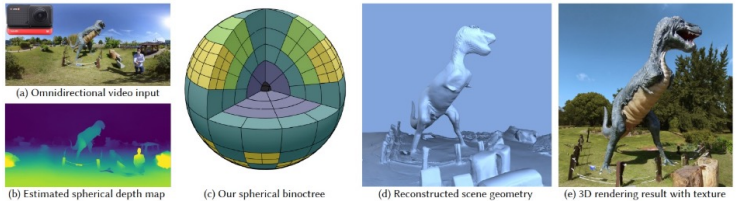
Introduction



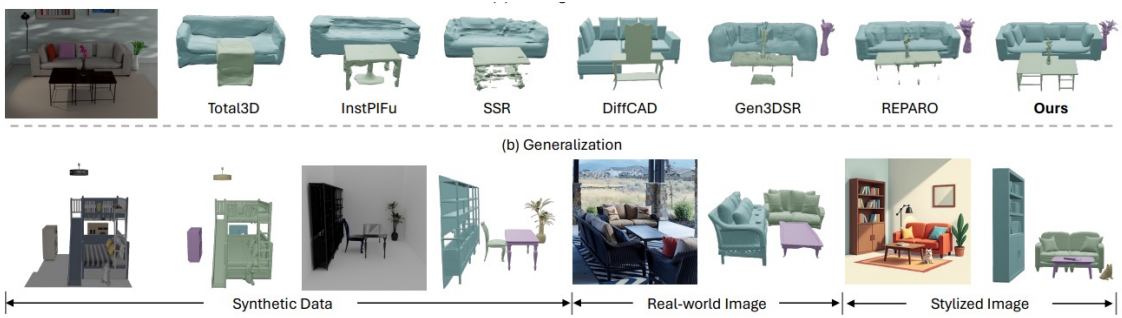
PMVC [Chushan Zhang et al. CVPR. 2024]



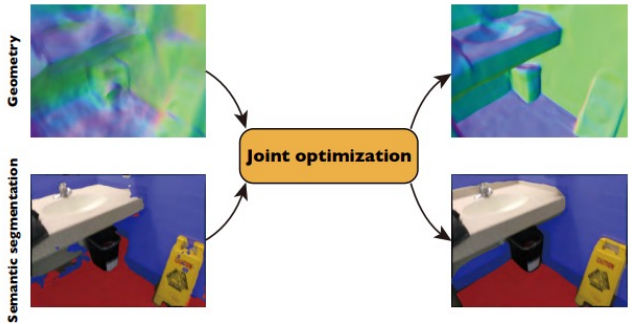
VisFusion [H Gao et al. CVPR. 2023]



EgocentricReconstruction [H Jang et al. SIGGRAPH. 2021]

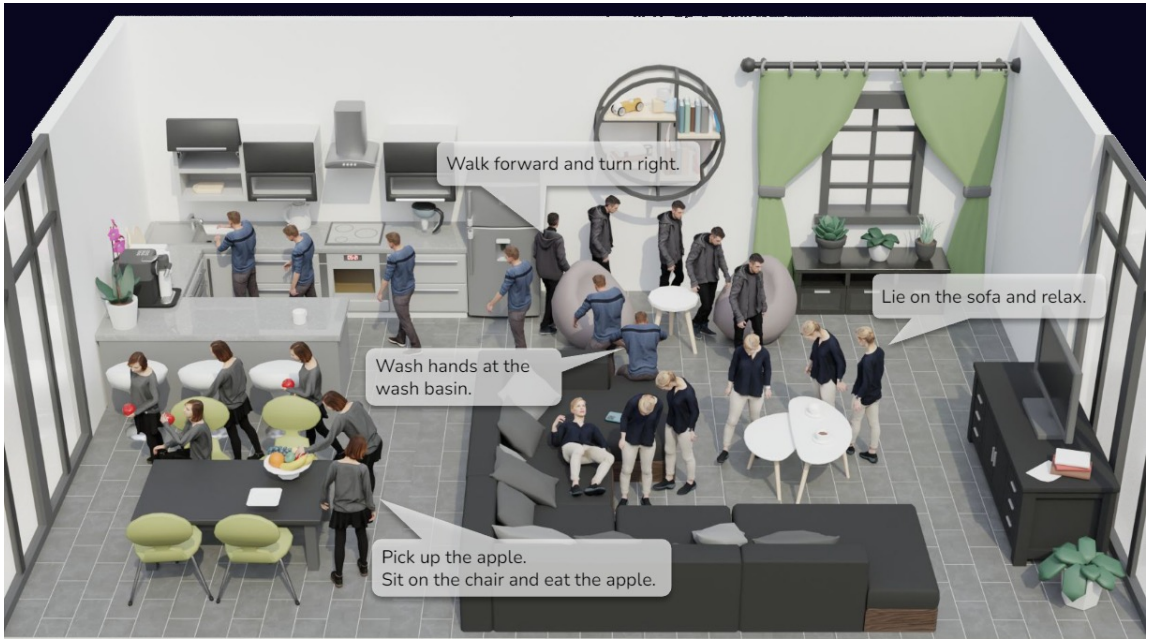


MIDI [Ze huang et al. arxiv. 2024]



Neural 3D Scene Reconstruction [Haoyu Guo et al. CVPR. 2022]

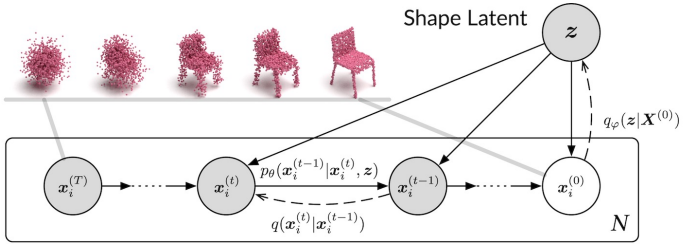
Introduction



If Human Motion is in the Scene, Motion will be interacting constantly!

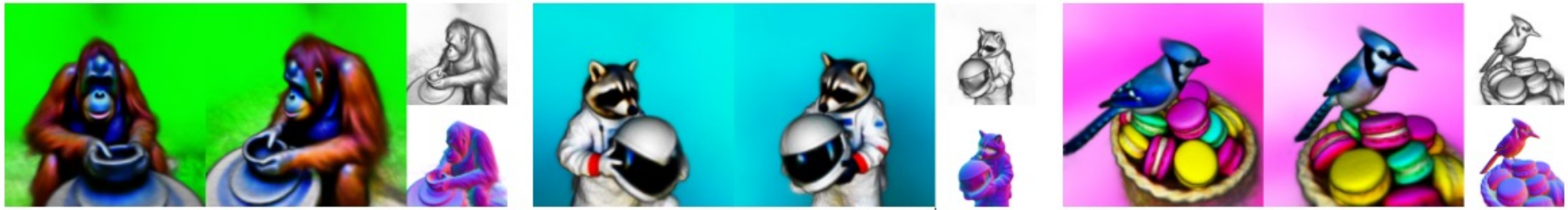
-> Generate Interacting Object and Scene from Human Motion

Generating Using Diffusion



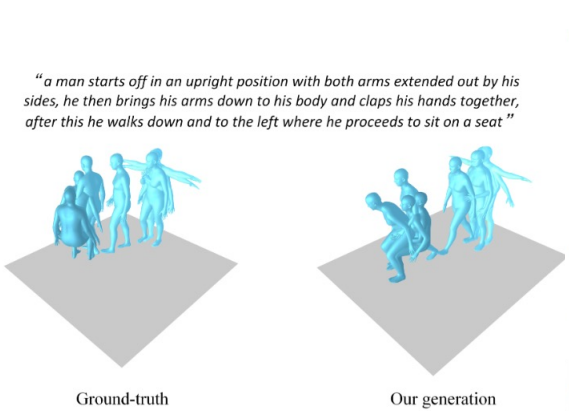
Diffusion-point-cloud, [Shitong Luoet al. CVPR. 2021]

MeshGPT, [Yawar Siddiqui al. CVPR. 2024]

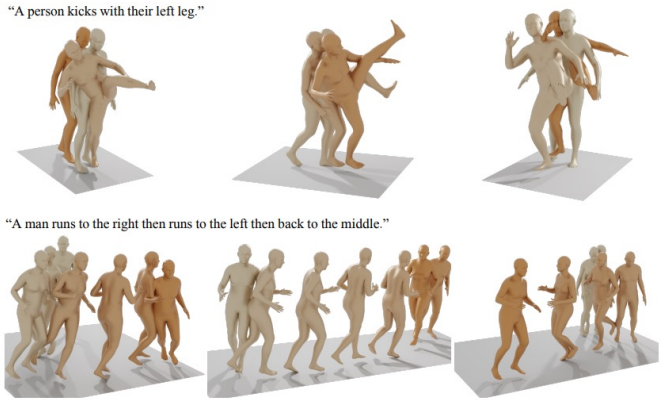


DreamFusion [Poole B et al. arxiv. 2022]

Generating Motion

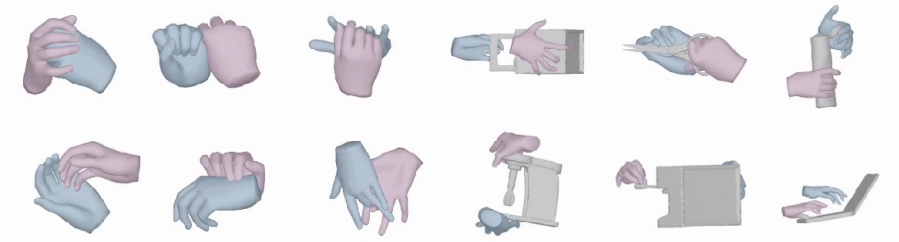


T2M GPT, [J Zhang et al. CVPR. 2023]



MDM [Guy Tevet et al. ICLR. 2023]

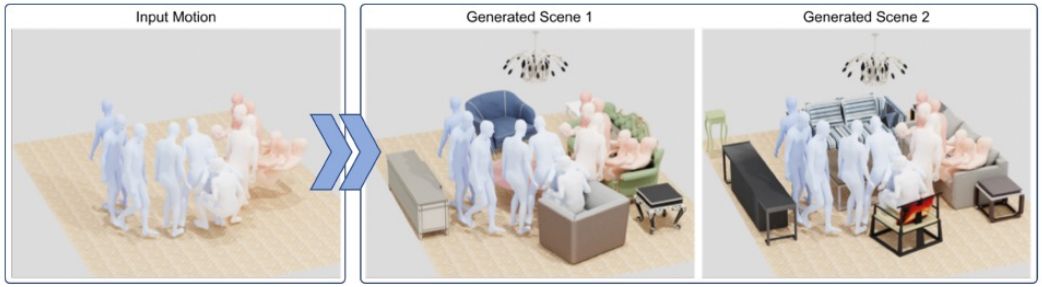
GMD [Korrawe K. et al. ICCV. 2023]



InterhandGen, [Jihyun Lee et al. CVPR. 2024]

SceneDiffuser, [S Huang et al. CVPR. 2023]

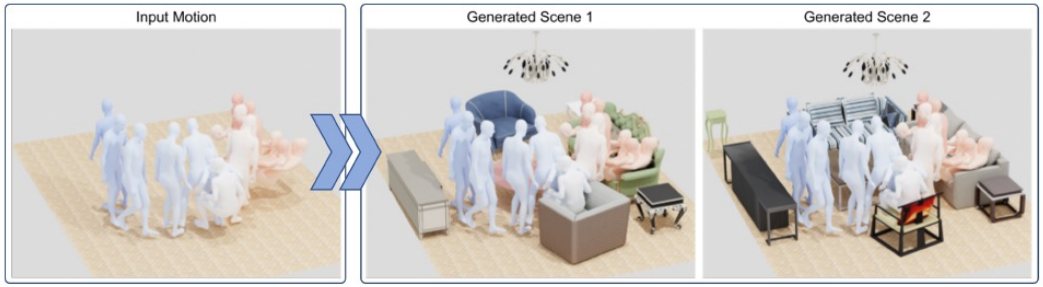
Most Related Works



MIME [Yi et al. CVPR. 2023]

LSDM [An Vuong et al. NeurIPS. 2023]

Most Related Works



MIME [Yi et al. CVPR. 2023]

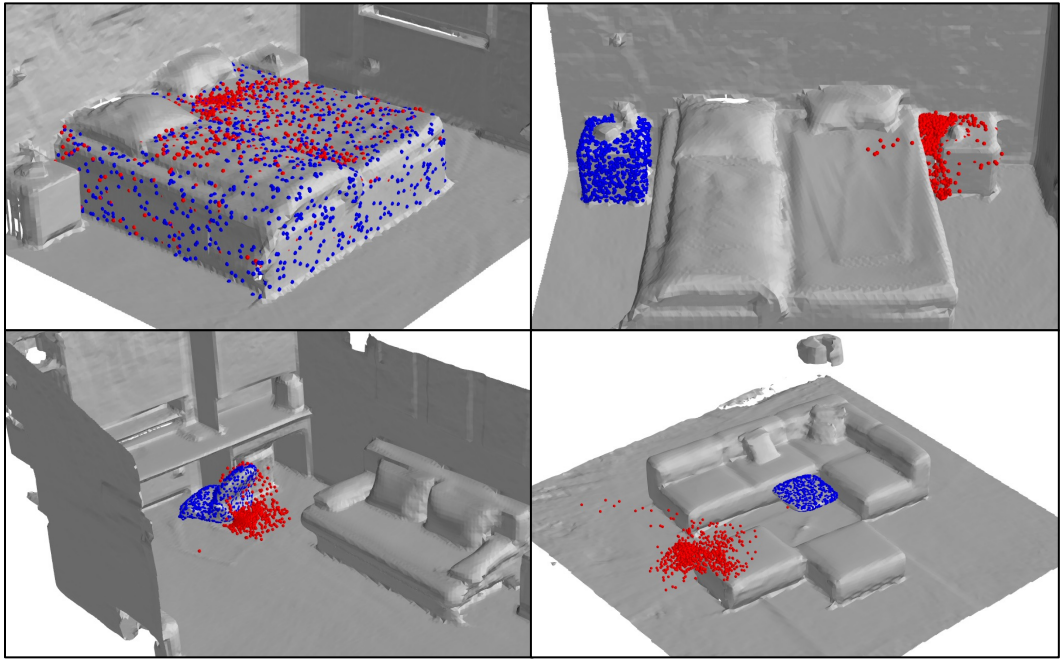
How about using short motions for Live?
How about predicting objects directly?

LSDM [An Vuong et al. NeurIPS. 2023]

How about predicting
without any text-guidance?

ex. *Put Chair Under the me and Next to bed*

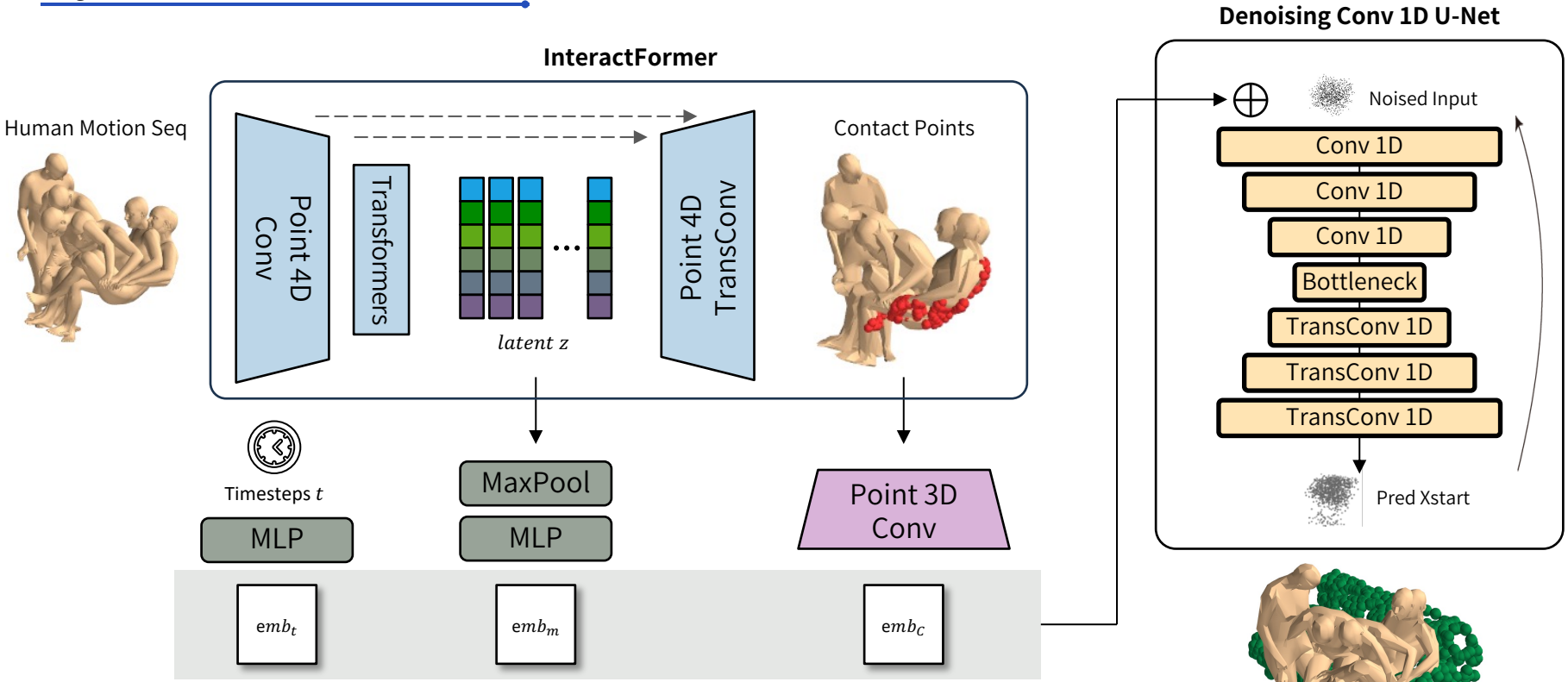
Limitations



Ground Truth (Blue), Predicted (Red)

Baseline	PRO-teXt CD ↓	PRO-teXt EMD ↓	PRO-teXt F1 ↑
Paschalidou et al., 2021)	2.0756	1.4140	0.0663
MMON (Ye et al., 2022)	2.1437	1.3994	0.0673
MIME (Yi et al., 2023)	2.0493	1.3832	0.0990
2023) + text embedding	1.8424	1.2865	0.1032
MCDM	0.6301	0.7269	0.3574
LSDM w.o. text (Ours)	0.9134	1.0156	0.0506
LSDM (Ours)	**0.5365**	**0.5906**	**0.5160**

System Architecture

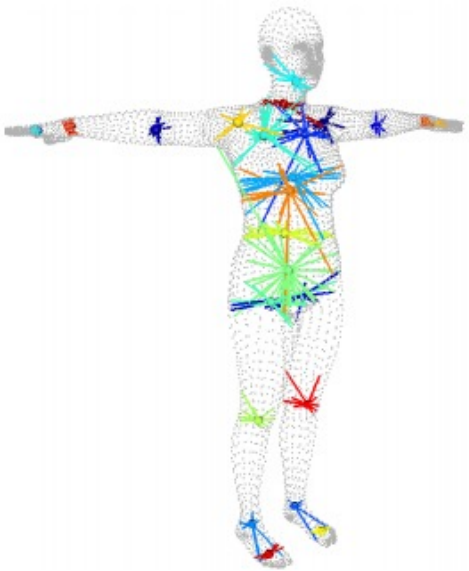


Main Purpose: Generating 3D Object, from Human Motion, Using Contact Points

Final Output (3D Object in Point Cloud)

Efficient Motion Data Format

SMPL Parameters (axis-angle)

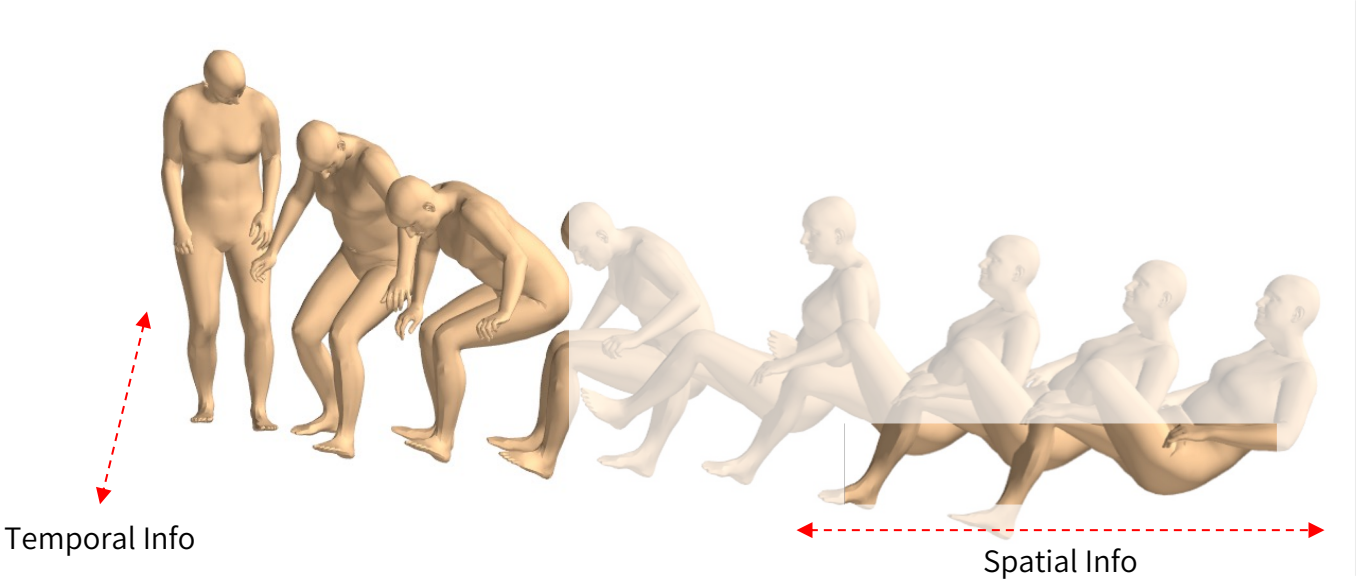


Easy to Understanding Full Body
Low Computation Cost
Less into about specific body (like Hand)

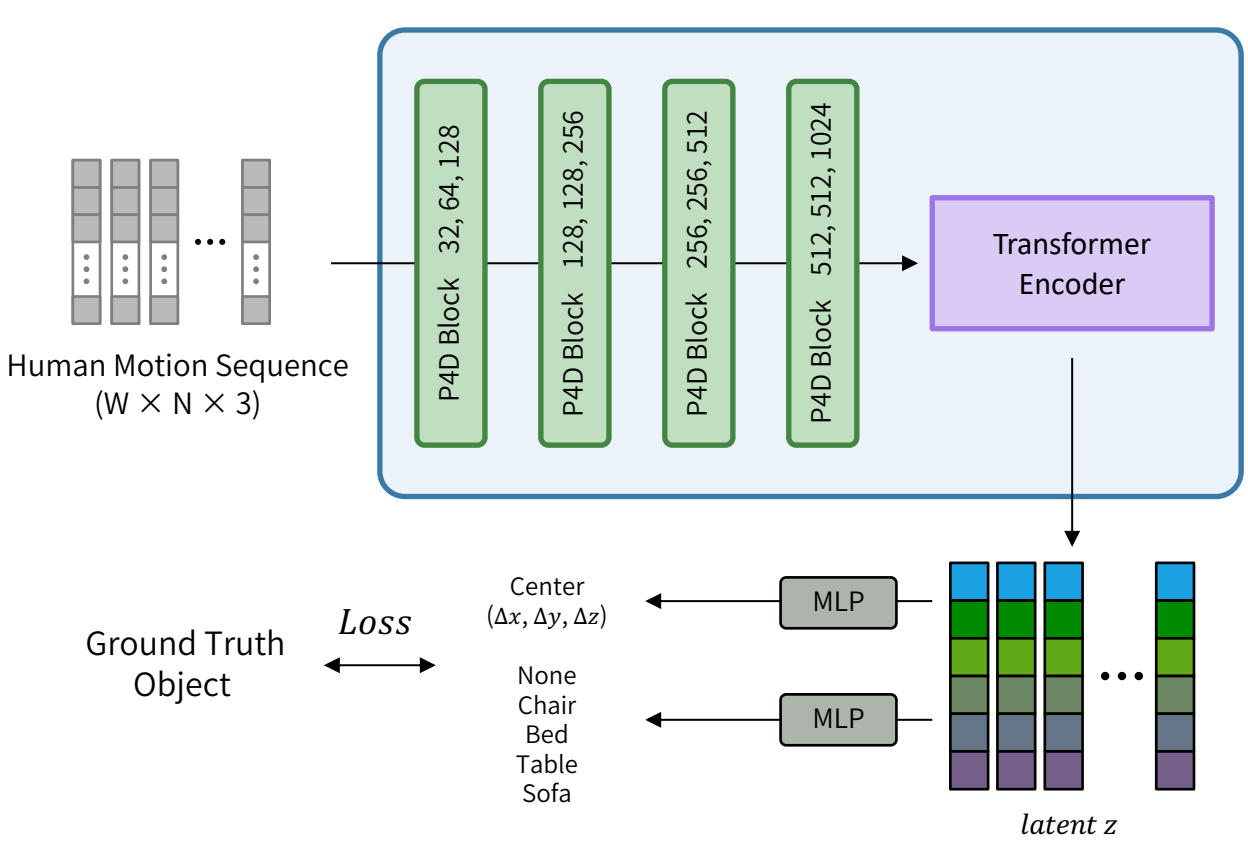
Point Cloud of Human Body Surface

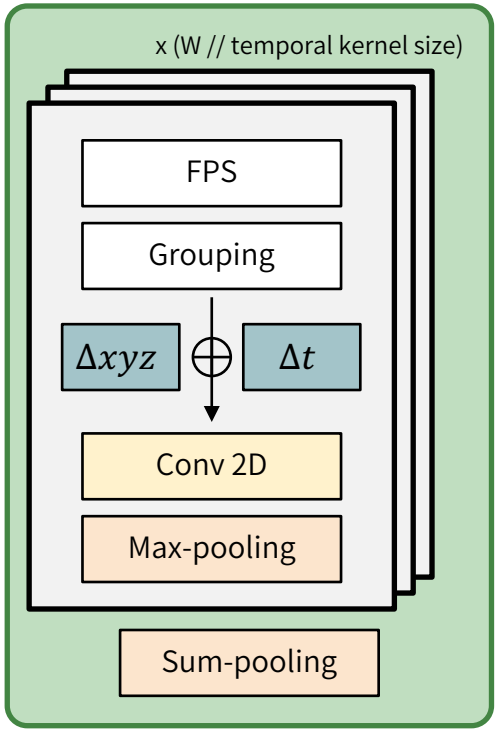
High Computation Cost
Helpful for providing condition
to Generating 3D Object

Dealing with 4D Data (3D Point Cloud + 1D Time)



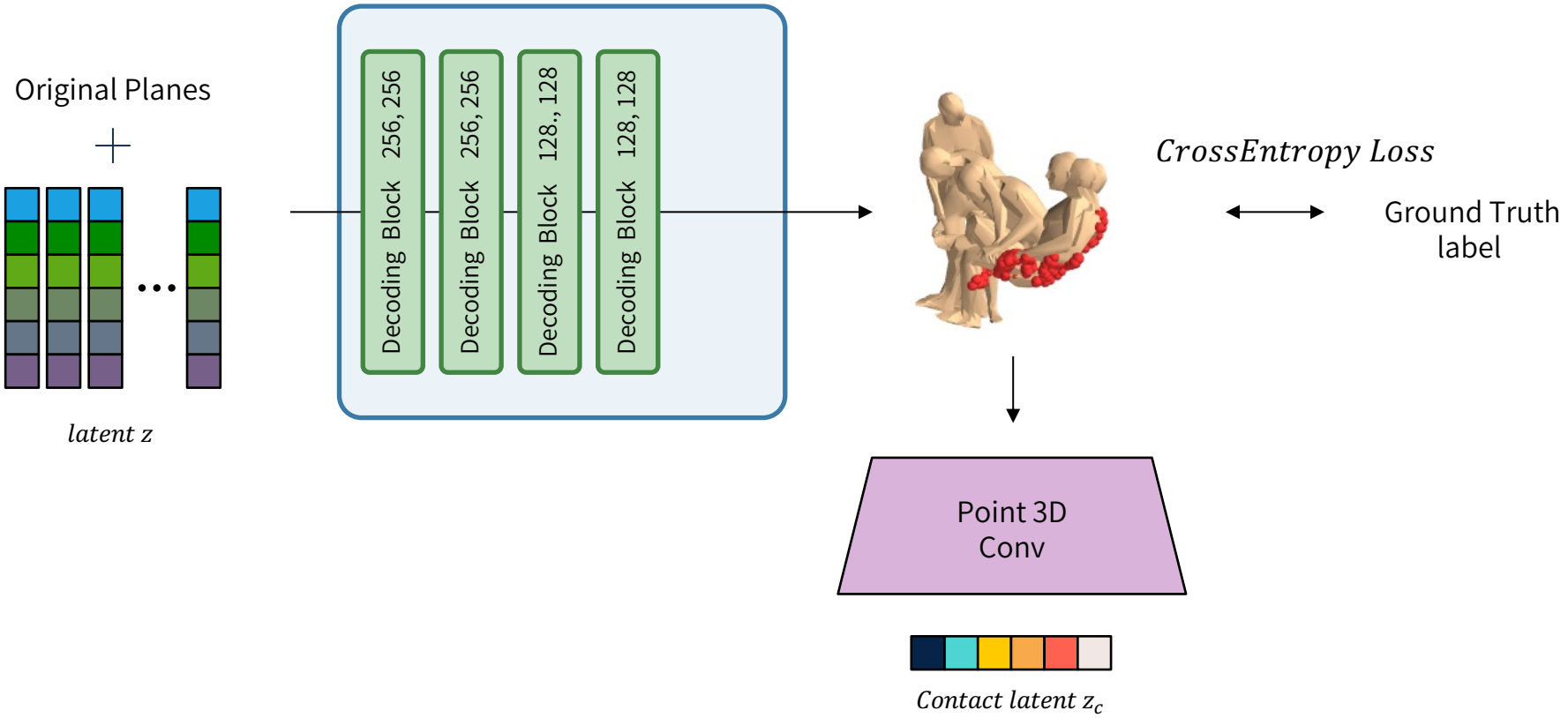
Encoding Human Motion Sequence



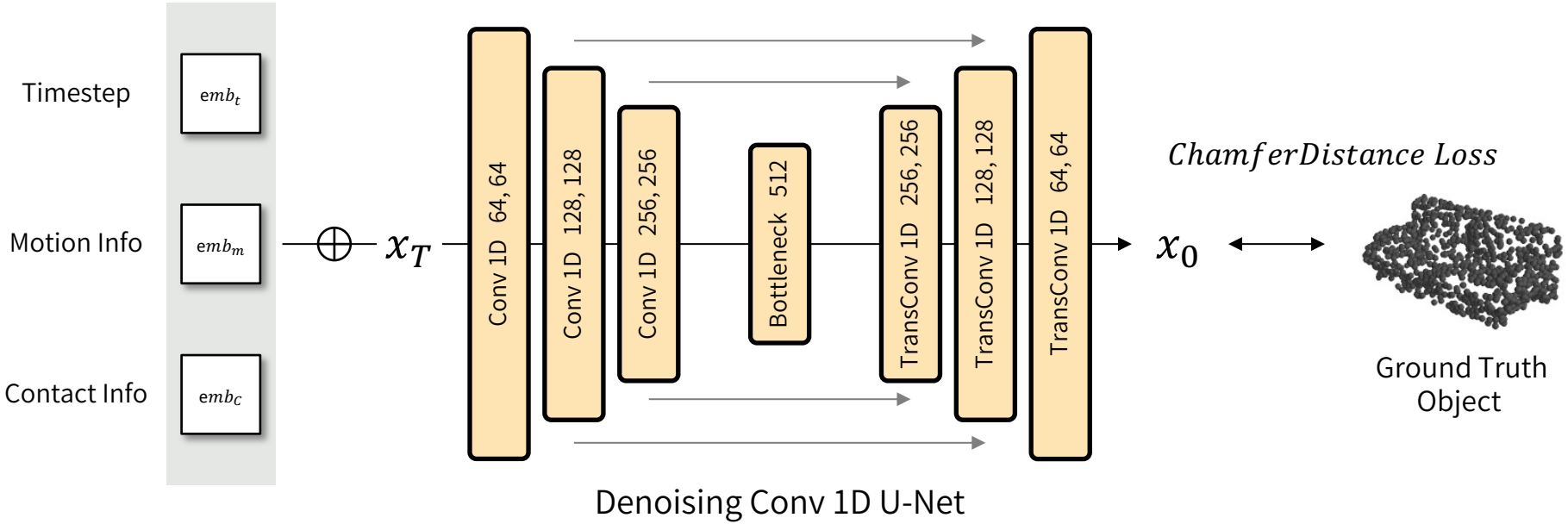


Loss Calculation for Latent Space, with Object Center and Category

Decoding Human Motion Sequence for Contact Points



Conditional Denoising U-Net for Generating Object



Conditional Diffusion – Forward Processing

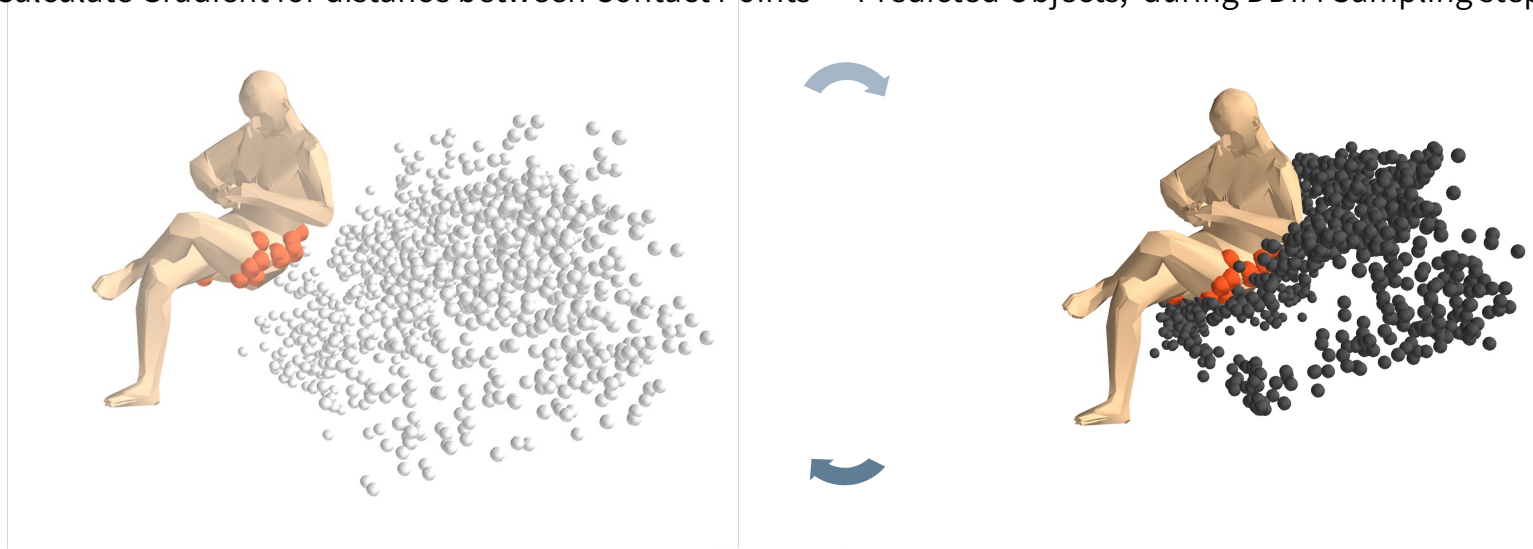
$$q(\mathbf{x}_{t+1}|\mathbf{x}_t) = \mathcal{N}\left(\sqrt{1-\beta_t}\,\mathbf{x}_t,\ \beta_t\mathbf{I}\right), q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=0}^{T-1} q(\mathbf{x}_{t+1}|\mathbf{x}_t)$$

Given Point Cloud X_0, add Gaussian noise each timestep with specific noise standard deviation

Using Cosine Scheduler for adding noise

Conditional Diffusion – CPG Sampling

Based on DDIM sampling, We used Contact Points Guidance (CPG) Sampling.
Calculate Gradient for distance between Contact Points ↔ Predicted Objects, during DDIM Sampling steps



$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\hat{\epsilon}}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\hat{\epsilon} + \sigma_t \epsilon_t$$

$$\mathbf{x}_{t-1} \leftarrow \mathbf{x}_{t-1} - w_{cpg} \cdot \nabla_{\mathbf{x}_{t-1}} CD(p_{cnt}, \mathbf{x}_{t-1})$$

Datasets

Real Data

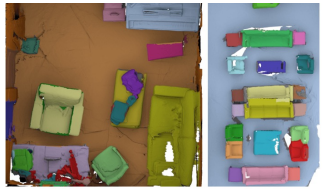
PROX

PROX, 43 sequences (about 1 min)
HUMANISE, 126 sequences (less 1 min)

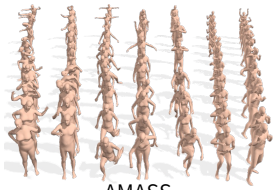
↓

Split it into 1-second short sequences, with the object it is most frequently in contact

Synthetic Data (Only Testing)



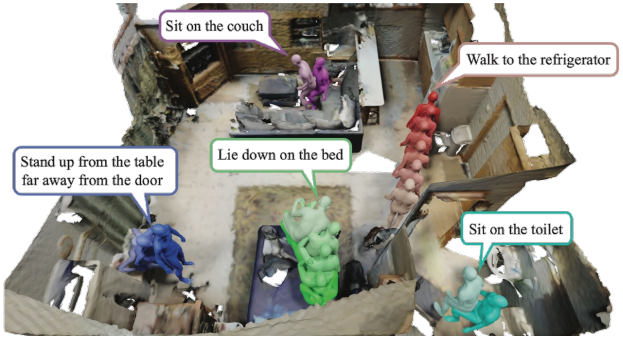
ScanNet v2

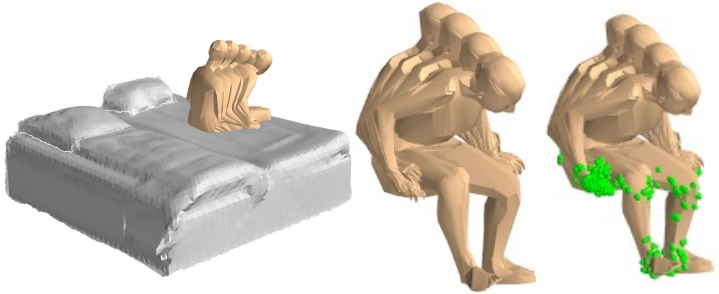


AMASS

→

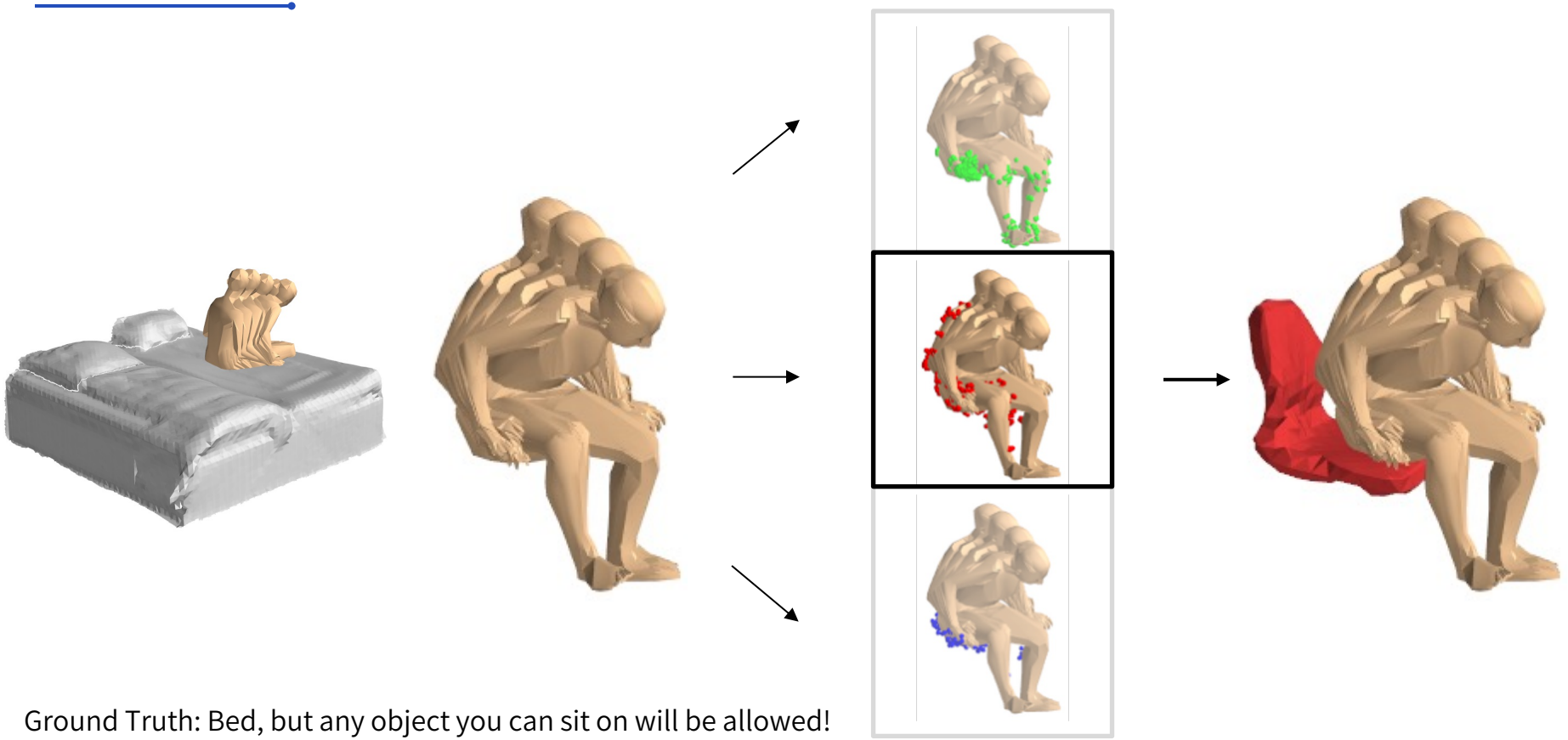
HUMANISE





Human Motion Sequence: (W × N × 3)
Target Object : (M × 3)
Contact Label: (W × N × num_class)

Datasets



Ground Truth: Bed, but any object you can sit on will be allowed!

Quantitative Results – InteractFormer

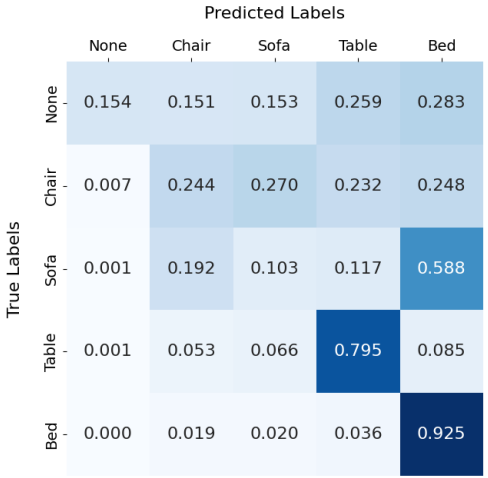
	PROX			HUMANISE		
	F1 ↑	Acc. ↑	Contact. ↑	F1 ↑	Acc. ↑	Contact. ↑
POSA	0.1269	0.1964	0.2257	0.0607	0.1100	0.1550
ContactFormer	0.2548	0.7646	0.7793	0.2151	0.7195	0.7416
Ours	**0.4577**	**0.8828**	**0.9029**	**0.2279**	**0.8048**	**0.8423**

Contact*: Evaluates whether points are in contact with any object,

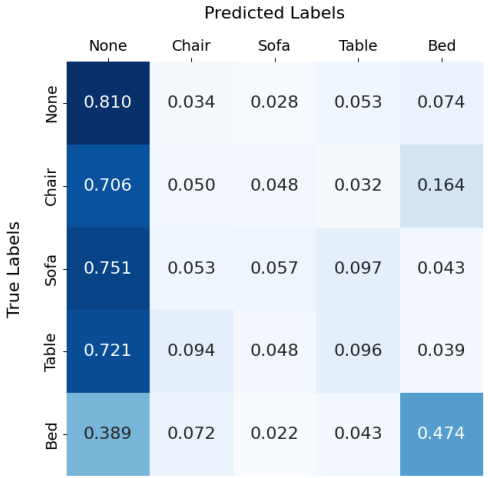
Our model achieves the best results in all three metrics for predicting Contact Points.

Quantitative Results – InteractFormer

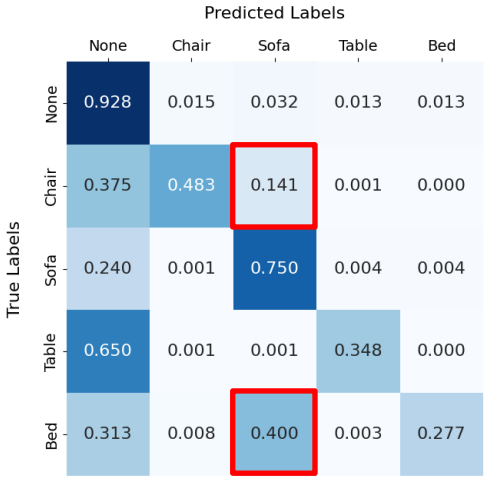
POSA



ContactFormer



InteractFormer (Ours)



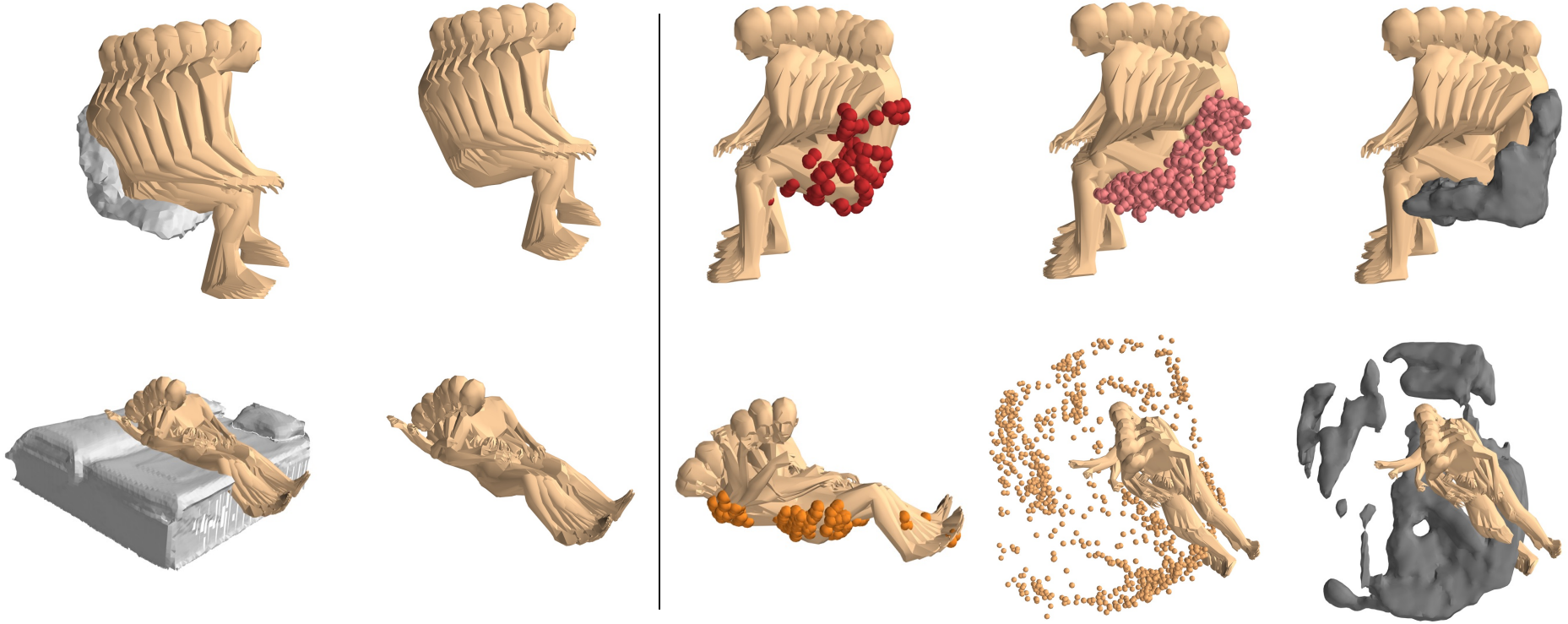
Confusion matrices for each of the three models

Quantitative Results

	PROX			HUMANISE		
	CD ↓	EMD ↓	F1 Score ↑	CD ↓	EMD ↓	F1 Score ↑
LSDM with no text	0.6252	0.9319	0.1961	0.8239	0.8665	0.1521
LSDM with GT Category text	0.5854	0.8874	0.2416	0.6098	0.8428	0.0826
Ours	**0.2982**	**0.6134**	**0.3662**	**0.4259**	**0.7704**	**0.1940**

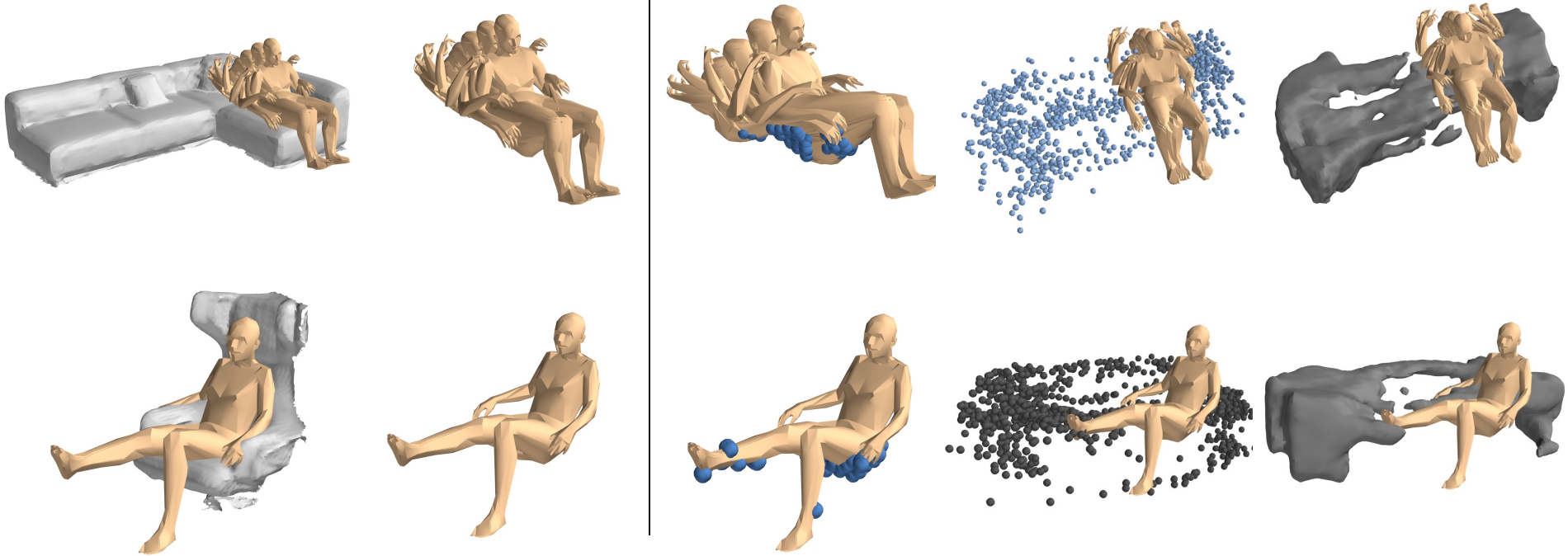
Our model achieves the best results in all three metrics

Qualitative Results



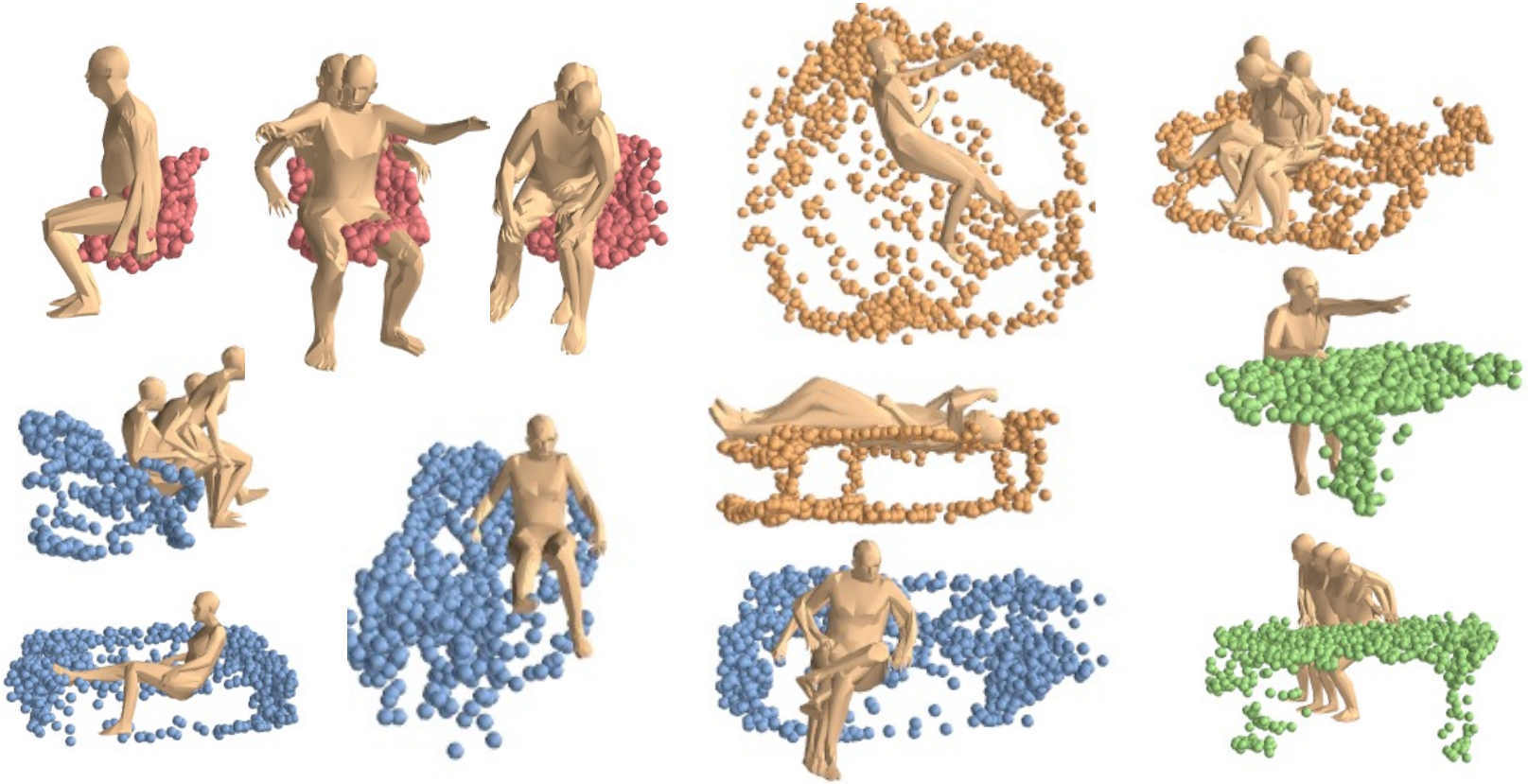
Ground Truth | Input | Pred Cont. Points | Pred Object | Applying SAP

Qualitative Results



Ground Truth | Input | Pred Cont. Points | Pred Object | Applying SAP

Qualitative Results



Conclusion

1. **Without any guidance or other information**, Generating interactive objects using only human motion.

2. **Using contact points predicting models,** Explainable results are derived.

3. **Live applications** are possible using short motion sequences (about 1 second).

Future Works

1. Using Synthetic data with exist Motion Diffusion Model

2. Apply for Scene Reconstruction, using Generated Objects

3. Combine with Sensor-to-Motion, for Sensor-to-Object

References

[1] Kingma, Diederik P. ”Auto-encoding variational bayes.” arXiv preprint arXiv:1312.6114 (2013).

[2] Goodfellow, Ian, et al. ”Generative adversarial networks.” Communications of the ACM 63.11(2020): 139-144.

[3] Lee, Jihyun, et al. ”InterHandGen: Two-Hand Interaction Generation via Cascaded Reverse Dif-fusion.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2024.

[4] Yi, Hongwei, et al. ”MIME: Human-aware 3D scene generation.” Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition. 2023.

[5] Ye, Sifan, et al. ”Scene synthesis from human motion.” SIGGRAPH Asia 2022 Conference Papers.2022.

[6] Nie, Yinyu, et al. ”Pose2room: understanding 3d scenes from human activities.” European Confer-ence on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[7] Hong, Xiaolin, et al. ”Human-Aware 3D Scene Generation with Spatially-constrained DiffusionModels.” arXiv preprint arXiv:2406.18159 (2024).

[8] Paschalidou, Despoina, et al. ”Atiss: Autoregressive transformers for indoor scene synthesis.” Ad-vances in Neural Information Processing Systems 34 (2021): 12013-12026.

[9] Siddiqui, Yawar, et al. ”Meshgpt: Generating triangle meshes with decoder-only transformers.”Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[10] Qi, Siyuan, et al. ”Human-centric indoor scene synthesis using stochastic grammar.” Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[11] Yi, Hongwei, et al. ”Human-aware object placement for visual environment reconstruction.” Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

References

[12] Luo, Shitong, and Wei Hu. ”Diffusion probabilistic models for 3d point cloud generation.” Proceed-ings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[13] Fan, Hehe, Yi Yang, and Mohan Kankanhalli. ”Point 4d transformer networks for spatio-temporalmodeling in point cloud videos.” Proceedings of the IEEE/CVF conference on computer vision andpattern recognition. 2021.

[14] Loper, Matthew, et al. ”SMPL: A skinned multi-person linear model.” Seminal Graphics Papers:Pushing the Boundaries, Volume 2. 2023. 851-866.

[15] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. ”Denoising diffusion probabilistic models.” Advancesin neural information processing systems 33 (2020): 6840-6851.

[16] Song, Jiaming, Chenlin Meng, and Stefano Ermon. ”Denoising diffusion implicit models.” arXivpreprint arXiv:2010.02502 (2020).

[17] Vuong, An Dinh, et al. ”Language-driven scene synthesis using multi-conditional diffusion model.”Advances in Neural Information Processing Systems 36 (2024).

[18] Taheri, Omid, et al. ”GOAL: Generating 4D whole-body motion for hand-object grasping.” Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[19] Wang, Zan, et al. ”Humanise: Language-conditioned human motion generation in 3d scenes.” Ad-vances in Neural Information Processing Systems 35 (2022): 14959-14971.

[20] Brock, Andrew, et al. ”Generative and discriminative voxel modeling with convolutional neuralnetworks.” arXiv preprint arXiv:1608.04236 (2016).

[21] Deng, Jiajun, et al. ”Voxel r-cnn: Towards high performance voxel-based 3d object detection.”Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 2. 2021.

[22] Tahir, Rohan, Allah Bux Sargano, and Zulfiqar Habib. ”Voxel-based 3D object reconstruction fromsingle 2D image using variational autoencoders.” Mathematics 9.18 (2021): 2288.

감사합니다

Q&A

2024.12.16

Neuro-Machine Augmented Intelligence Lab (NMAIL)

전윤호 (Younho Jeon)

Advisor: 조성호 교수님

Committee: 김태균 교수님, 최성희 교수님

NMAIL