CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

CVPR 2024

Presenter: Jiahang Zhang
2024.3.31

Background: SMPL

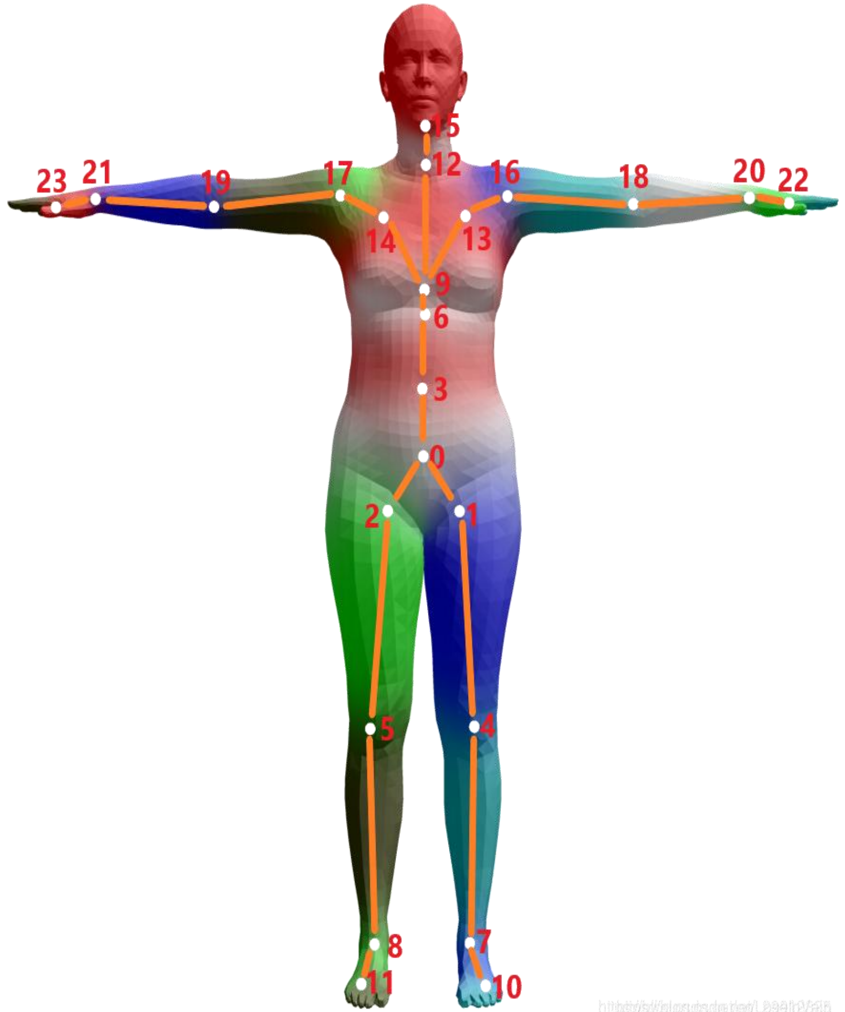
SMPL: A Skinned Multi-Person Linear Model

Matthew Loper *[1,2] Naureen Mahmood†[1] Javier Romero†[1] Gerard Pons-Moll†[1] Michael J. Black†[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Industrial Light and Magic, San Francisco, CA

SMPL Model

- Shape Parameters (10,)
- Pose Parameters (24,3)
- Vertex Number N = 6890



SMPL-H: hand+body

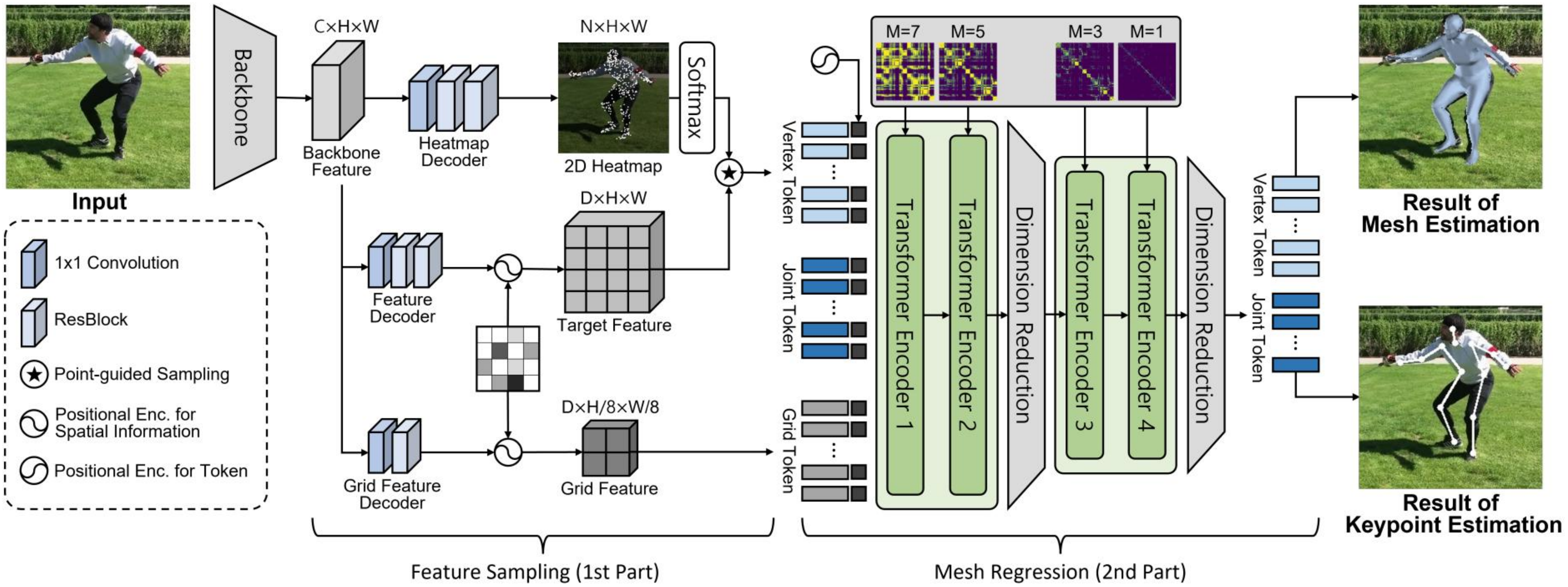
SMPL-X: face+hand+body

Background: 3D Human Reconstruction

Sampling is Matter: Point-guided 3D Human Mesh Reconstruction

Jeonghwan Kim[1*] Mi-Gyeong Gwon[1*] Hyunwoo Park[1]
Hyukmin Kwon[2] Gi-Mun Um[2] Wonjun Kim[1†]
[1]Konkuk University [2]Electronics and Telecommunications Research Institute

Model-Free Methods



PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop

Hongwen Zhang[§‡*], Yating Tian[†*], Xinchi Zhou[♮], Wanli Ouyang[♮], Yebin Liu[‡], Limin Wang[†✉], Zhenan Sun[§✉]
[§]CRIPAC, NLPR, Institute of Automation, Chinese Academy of Sciences, China
[†]State Key Laboratory for Novel Software Technology, Nanjing University, China
[♮]The University of Sydney, Australia [‡]Department of Automation, Tsinghua University, China

Model-Based Methods

- Pose, shape, and camera parameters Θ

$$\mathcal{L}_{reg} = \lambda_{2d}||K - \hat{K}|| + \lambda_{3d}||J - \hat{J}|| + \lambda_{para}||\Theta - \hat{\Theta}||, \quad (2)$$

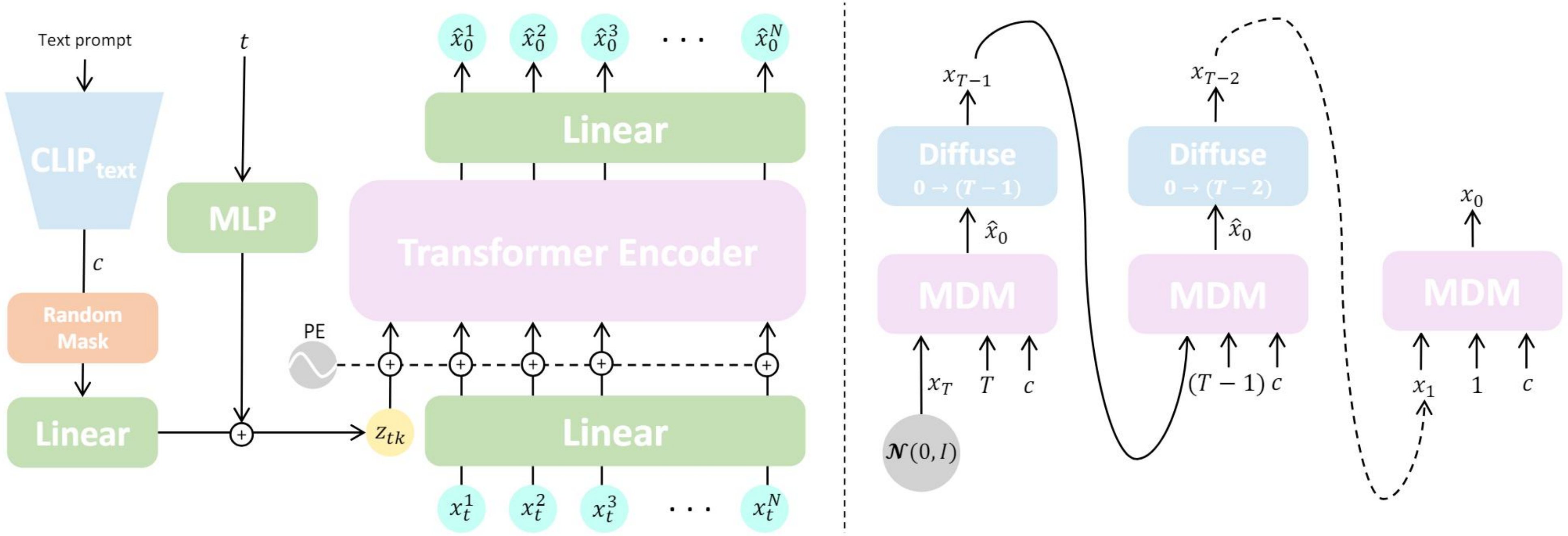
where $||\cdot||$ is the squared L2 norm, $\hat{K}$, $\hat{J}$, and $\hat{\Theta}$ denote the ground truth 2D keypoints, 3D joints, and model parameters, respectively.

Background: 3D Human Motion Generation

HUMAN MOTION DIFFUSION MODEL

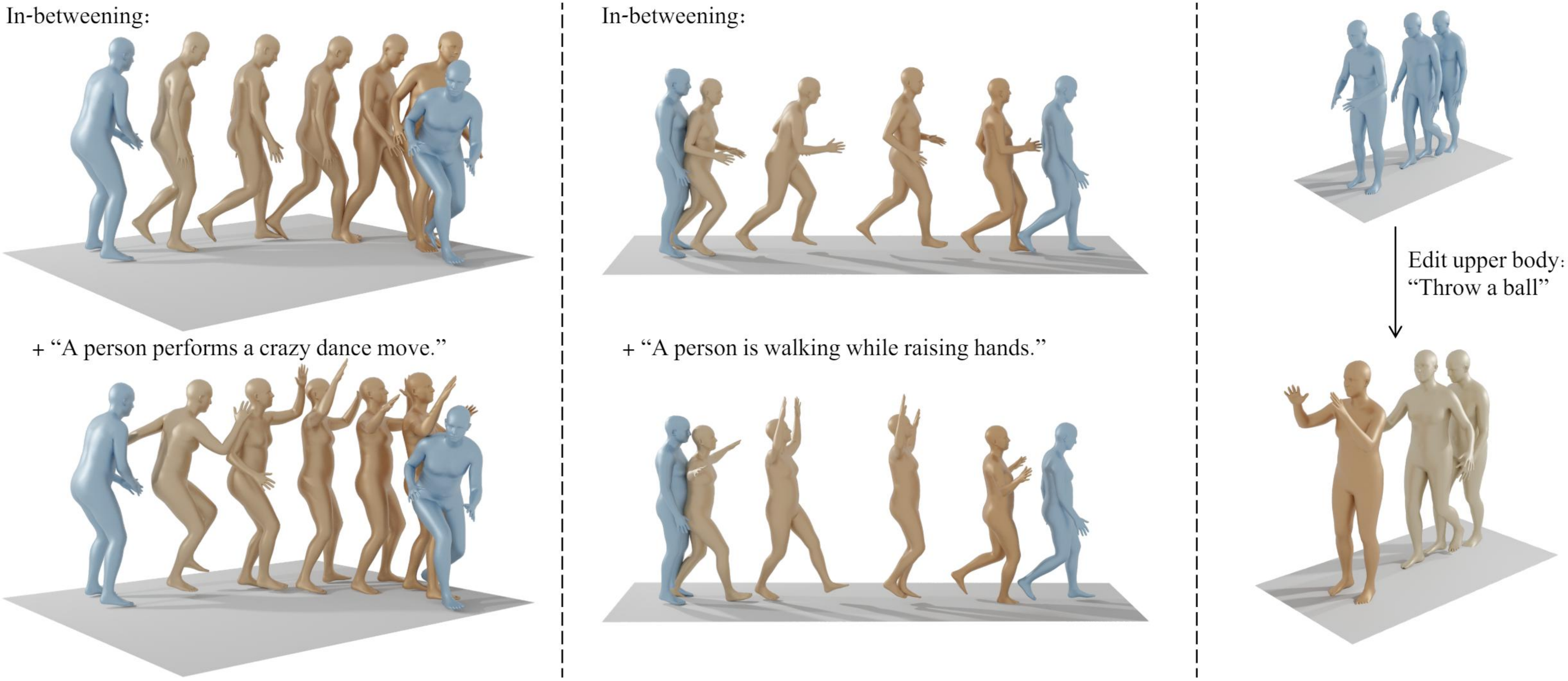
Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or and Amit H. Bermano
Tel Aviv University, Israel
guytevet@mail.tau.ac.il

Diffusion-Based



Background: 3D Human Motion Generation

Diffusion-Based



Background: 3D Human Motion Generation

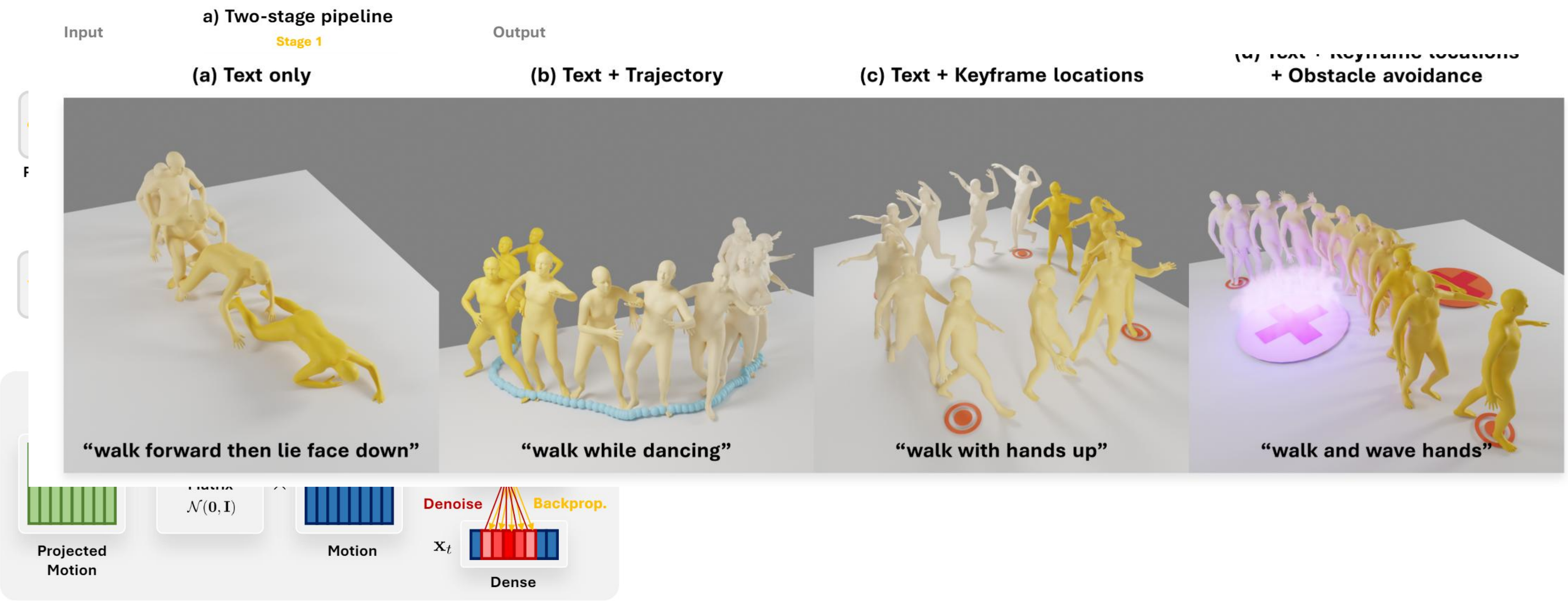
Guided Motion Diffusion for Controllable Human Motion Synthesis

Korrawe Karunratanakul[1] Konpat Preechakul[2] Supasorn Suwajanakorn[2] Siyu Tang[1]

[1]ETH Zürich, Switzerland [2]VISTEC, Thailand

https://korrawe.github.io/gmd-project/

Diffusion-Based

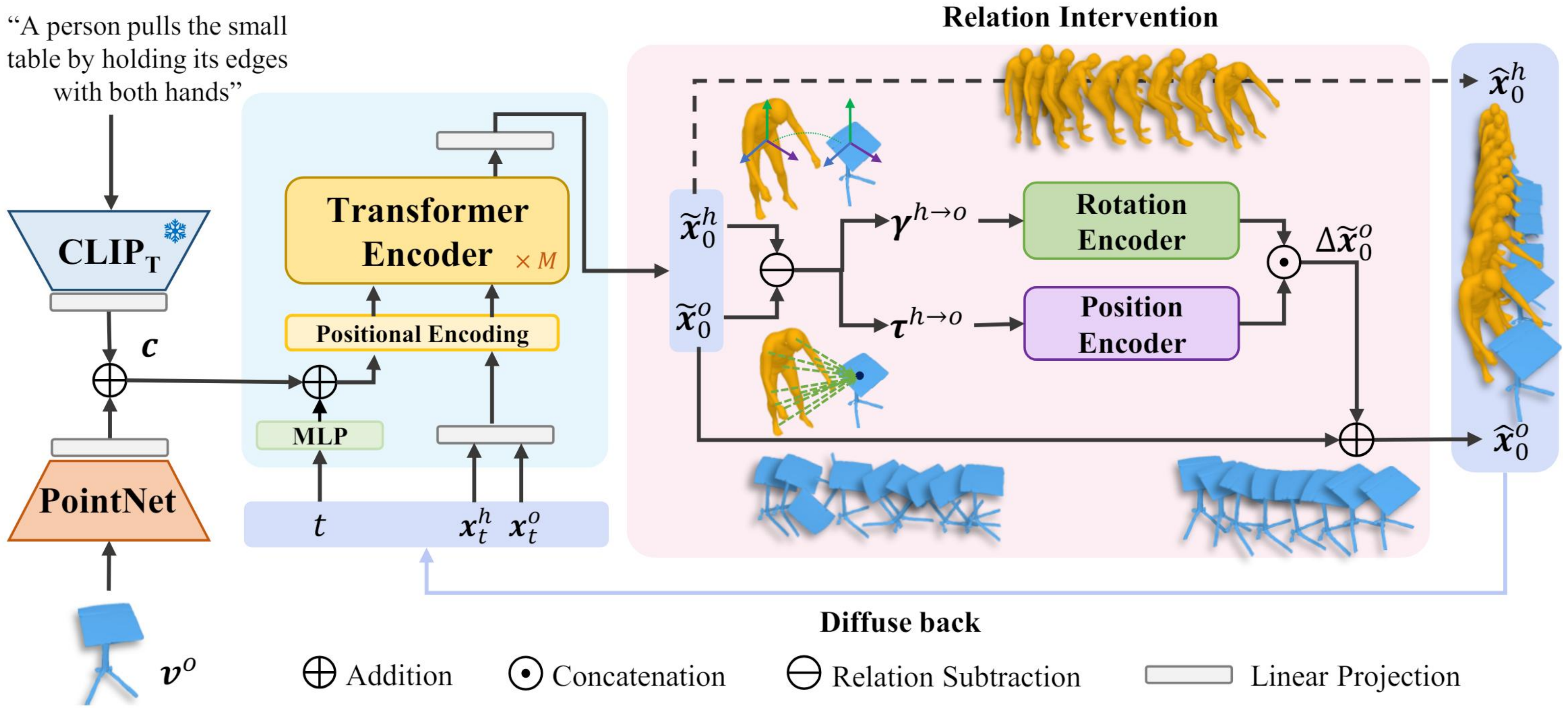


Background: HOI Generation

THOR: Text to Human-Object Interaction Diffusion via Relation Intervention

Qianyang Wu[1], Ye Shi[1], Xiaoshui Huang[2], Jingyi Yu[1], Lan Xu[1], and Jingya Wang[1]

[1] ShanghaiTech University
[2] Shanghai AI Laboratory

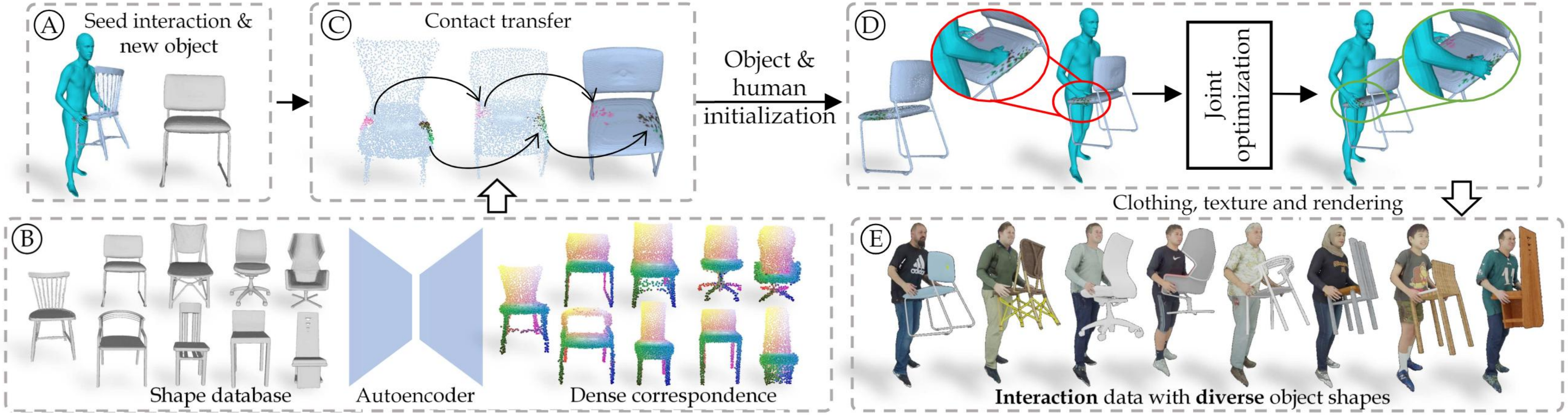


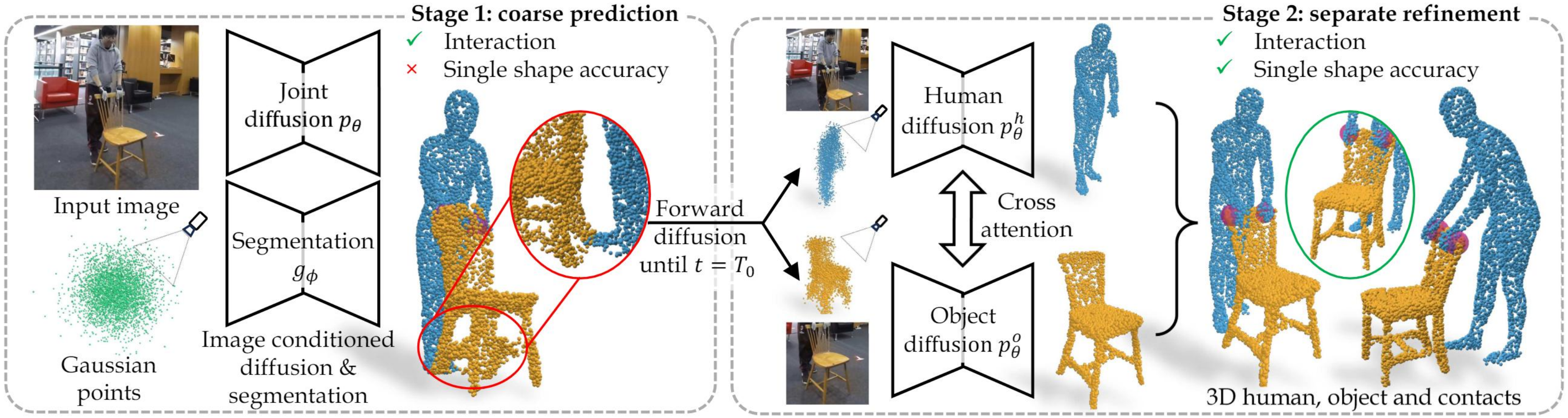
Background: HOI Generation

Template Free Reconstruction of Human-object Interaction with Procedural Interaction Generation

Xianghui Xie[1,2,3] Bharat Lal Bhatnagar[4] Jan Eric Lenssen[3] Gerard Pons-Moll[1,2,3]

[1]University of Tübingen, [2]Tübingen AI Center, Germany
[3]Max Planck Institute for Informatics, Germany
[4]Meta Reality Labs

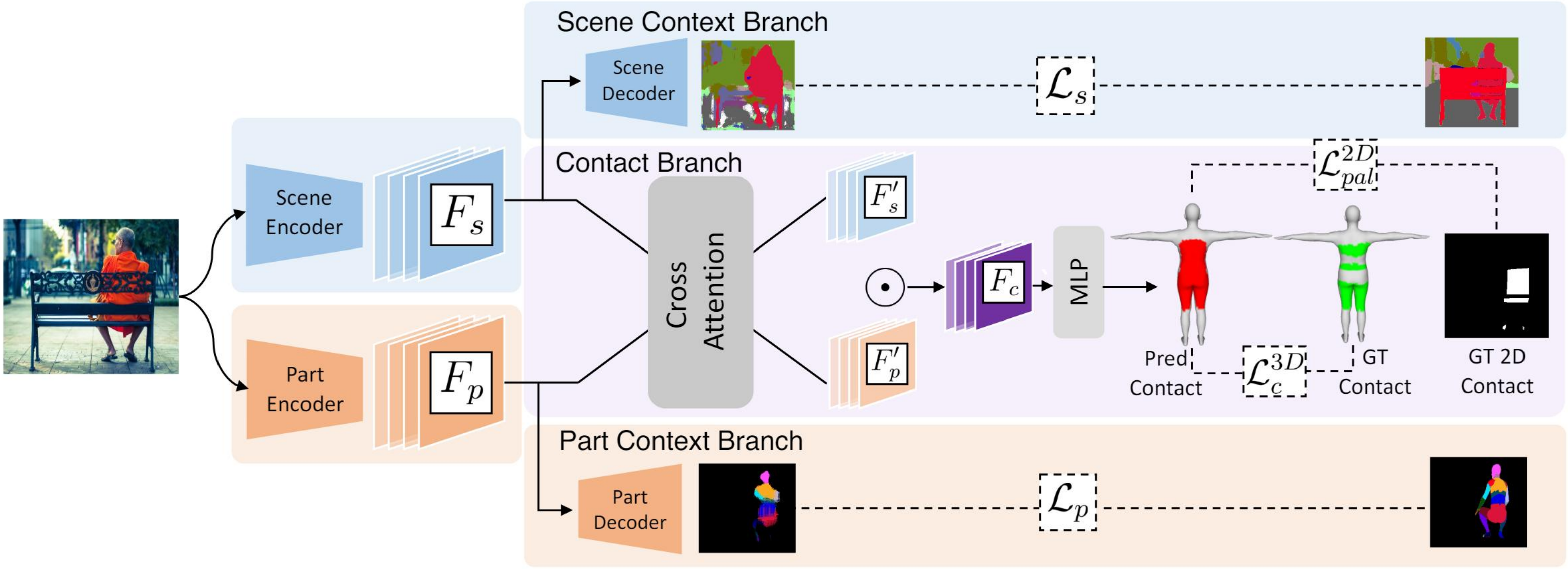




Background: Contact Prediction for HOI

DECO: Dense Estimation of 3D Human-Scene Contact In The Wild

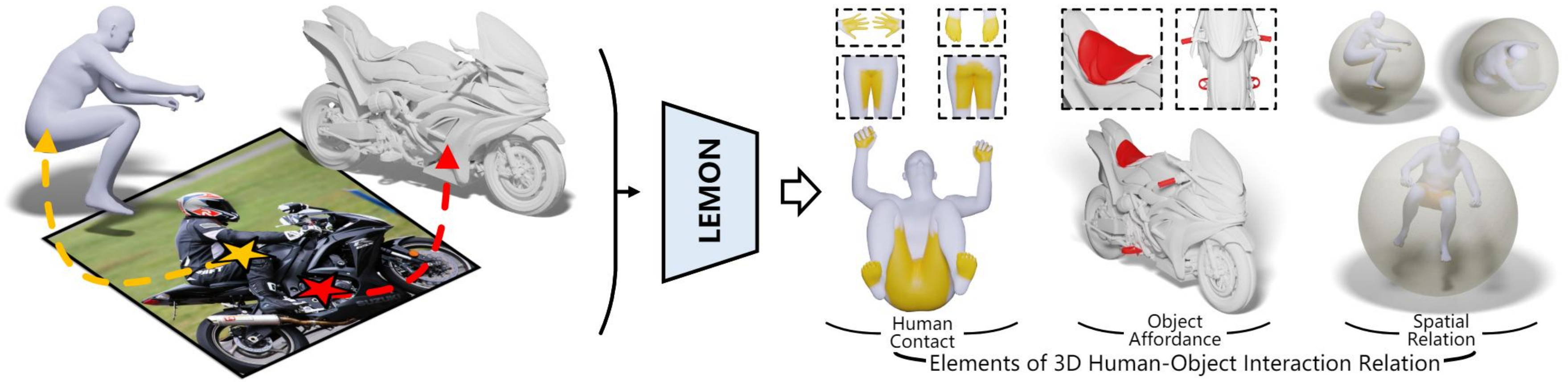
Shashank Tripathi[1*] Agniv Chatterjee[1*] Jean-Claude Passy[1] Hongwei Yi[1]
Dimitrios Tzionas[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Amsterdam, the Netherlands
{stripathi, achatterjee, jpassy, hyi, black}@tue.mpg.de d.tzionas@uva.nl

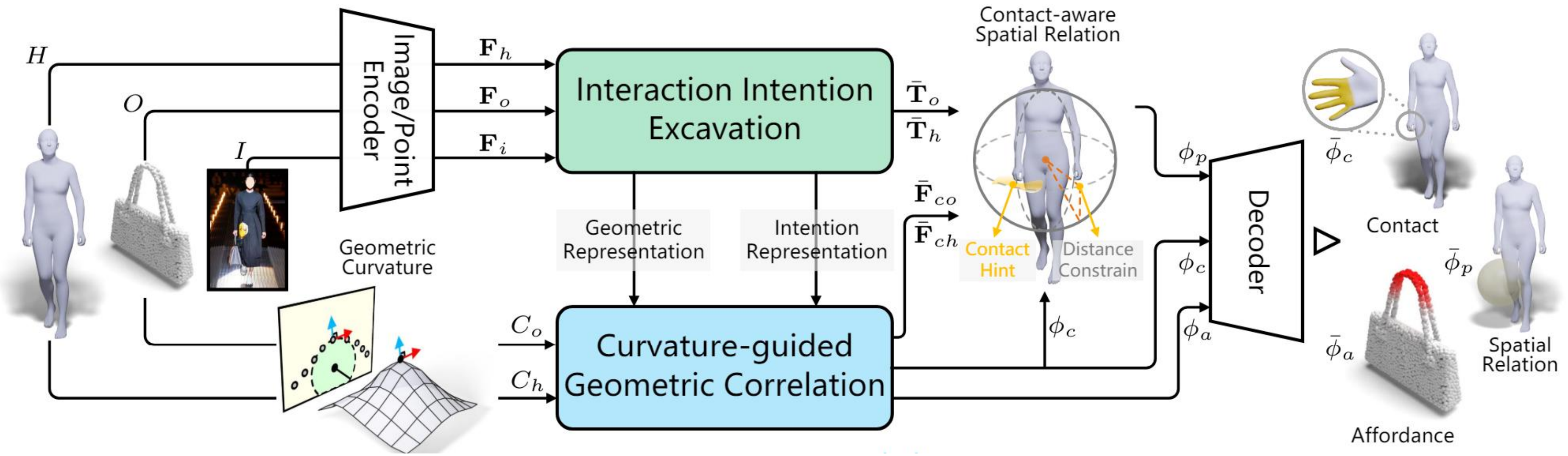


Background: Contact Prediction for HOI

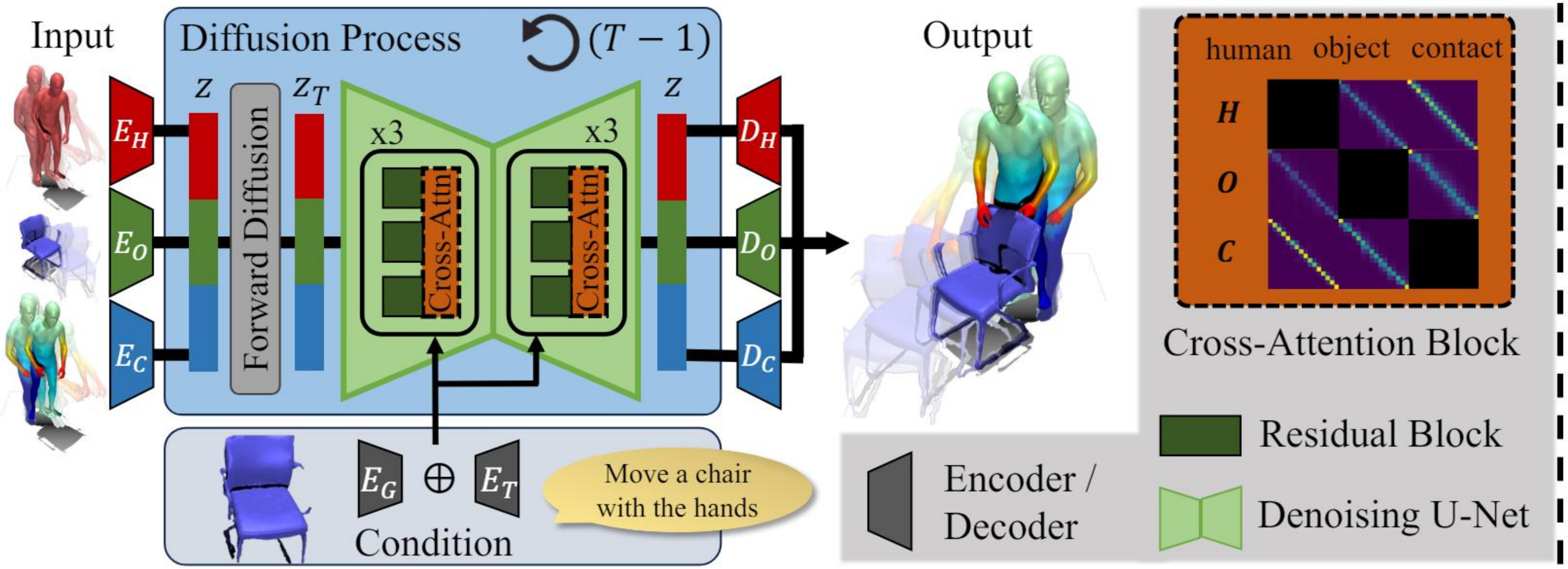
LEMON: Learning 3D Human-Object Interaction Relation from 2D Images

Yuhang Yang[1], Wei Zhai[1,†], Hongchen Luo[1], Yang Cao[1,2], Zheng-Jun Zha[1]
[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center





CG-HOI: Method

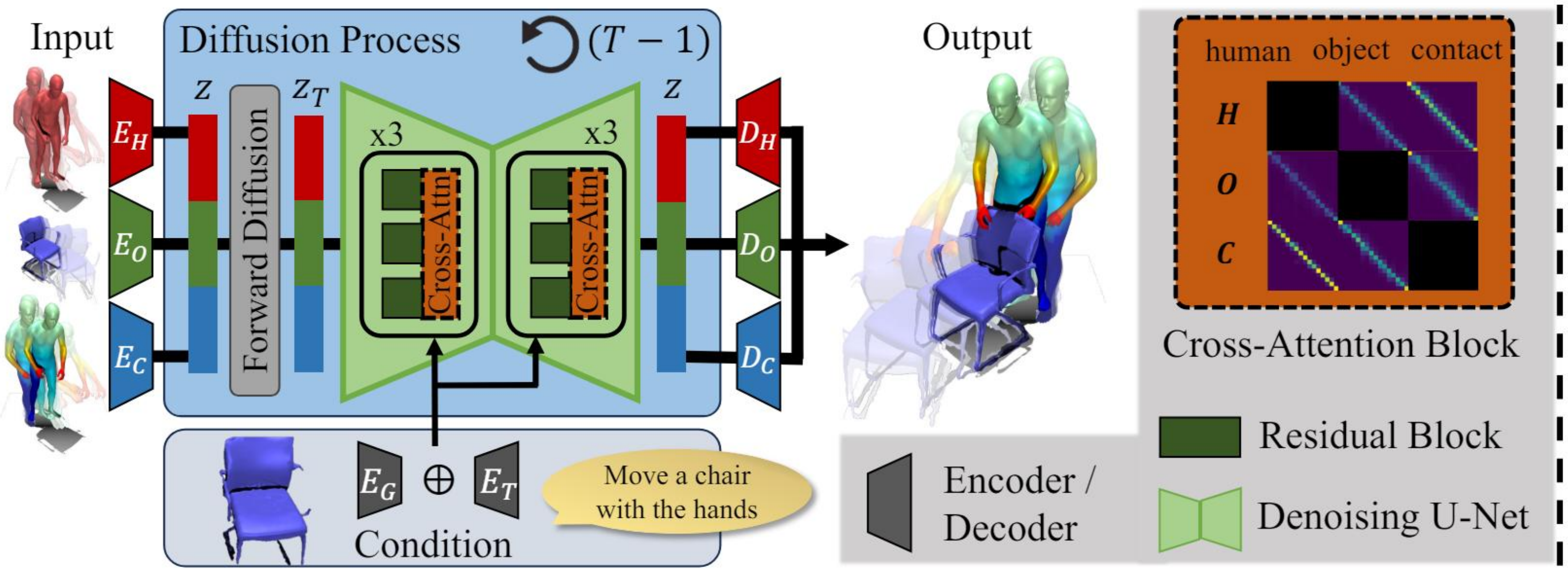


- Diffusion-based
- Conditioned on text description T, static geometry G
- Contact-guided inference

CG-HOI: Method

Diffusion Process

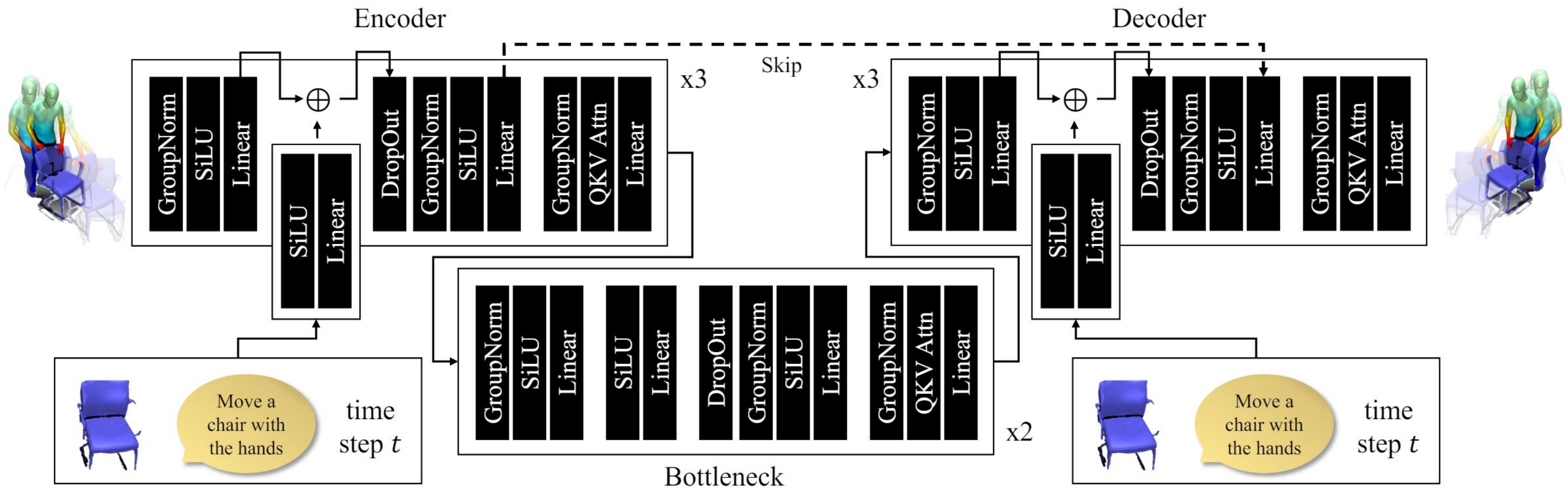
- Human surface: SMPL
- Object Transformation: global translation, rotation matrix
- Contact: distance between markers to the closest object geometry point



CG-HOI: Method

Diffusion Process

- Text encoder: CLIP
- Object Geometry encoder: PointNet
- Cross-Attention Mechanism for hi, oi, ci, to model inter-dependency



Object Motion Modeling

- *Object motion is naturally most influenced by parts of the human body in very close contactto the object.*
- Predict one hypotheses for each marker, instead one for the whole seq.

Formally, we predict object transformation hypotheses o_i^j for each contact point on the human body, and weigh them with the inverse of their predicted contact distance c_i^j:

$$o_i = \frac{1}{\sum_j c_i} \sum_{j=0}^{N} (\max(|c_i|) - |c_i^j|) o_i^j, \qquad (5)$$

CG-HOI: Method

Loss Formulation

- Between Prediction and Ground-Truth
- Classifier-Free Guidance

$$\mathbf{L} = \lambda_h ||h_i - \hat{h_i}||_1 + \lambda_o ||o_i - \hat{o}_i||_1 + \lambda_c ||c_i - \hat{c}_i||_2,$$

CG-HOI: Method

Interaction Generation

- Explicit constraints during inference on human-object contact
- Apply a a cost function $G((x)_t)$ at each time step

$$\hat{\mu}_t = \mu_t + s \sum_t \nabla_{x_t} G(x_t),$$

- The input can also be replaced by the object trajectory without any new training.

CG-HOI: Experiment

Dataset

- CHAIRS: 46 subjects as their SMPL-X, interaction with chairs and sofas
- BEHAVE: 8 participants as their SMPL-H alongside 20 different objects

Evaluation Metrics

- **R-Precision**: the closeness of the text condition and generated HOI in latent feature space.
- **FID**: the similarity between generated and ground-truth distribution in encoded feature space.
- **Diversity and MultiModality**: motion variance across all text descriptions.
- **Perceptual User Study**

CG-HOI: Experiment

Quantitative Result

		BEHAVE		
Task	Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →
	Real (human)	0.73	0.09	4.23
Text-Cond. Human Only	MDM [71]	0.52	4.54	5.44
	InterDiff [84]	0.49	5.36	3.98
	Ours	**0.60**	**4.26**	**4.92**
	Real	0.81	0.17	6.80
Motion-Cond. HOI	InterDiff [84]	0.68	3.86	5.62
	Ours	**0.71**	**3.52**	**6.89**
Text-Cond. HOI	MDM [71]	0.49	9.21	6.51
	InterDiff [84]	0.53	8.70	3.85
	Ours	**0.62**	**6.31**	**6.63**

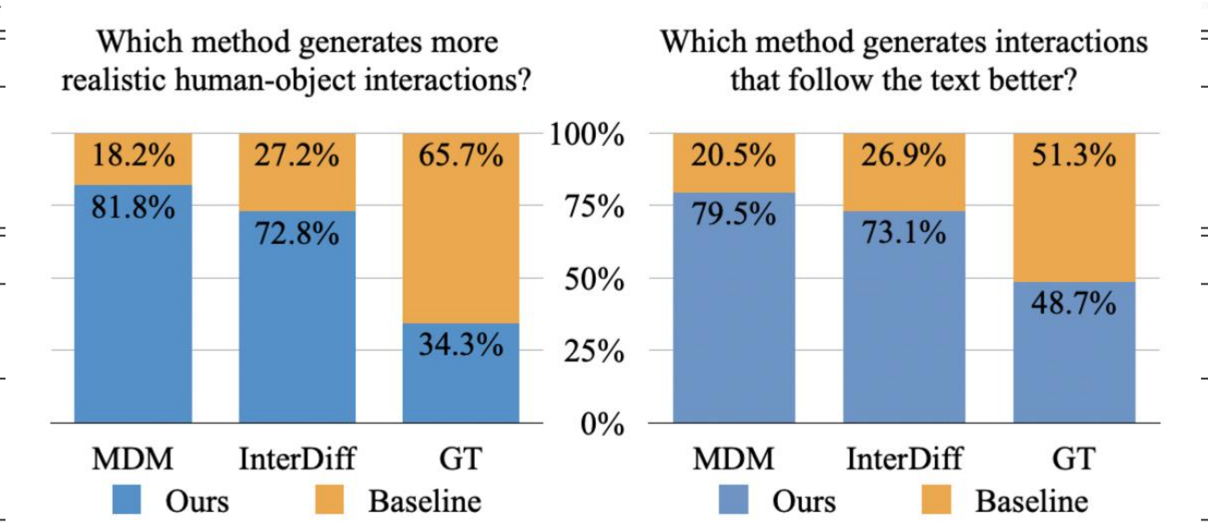
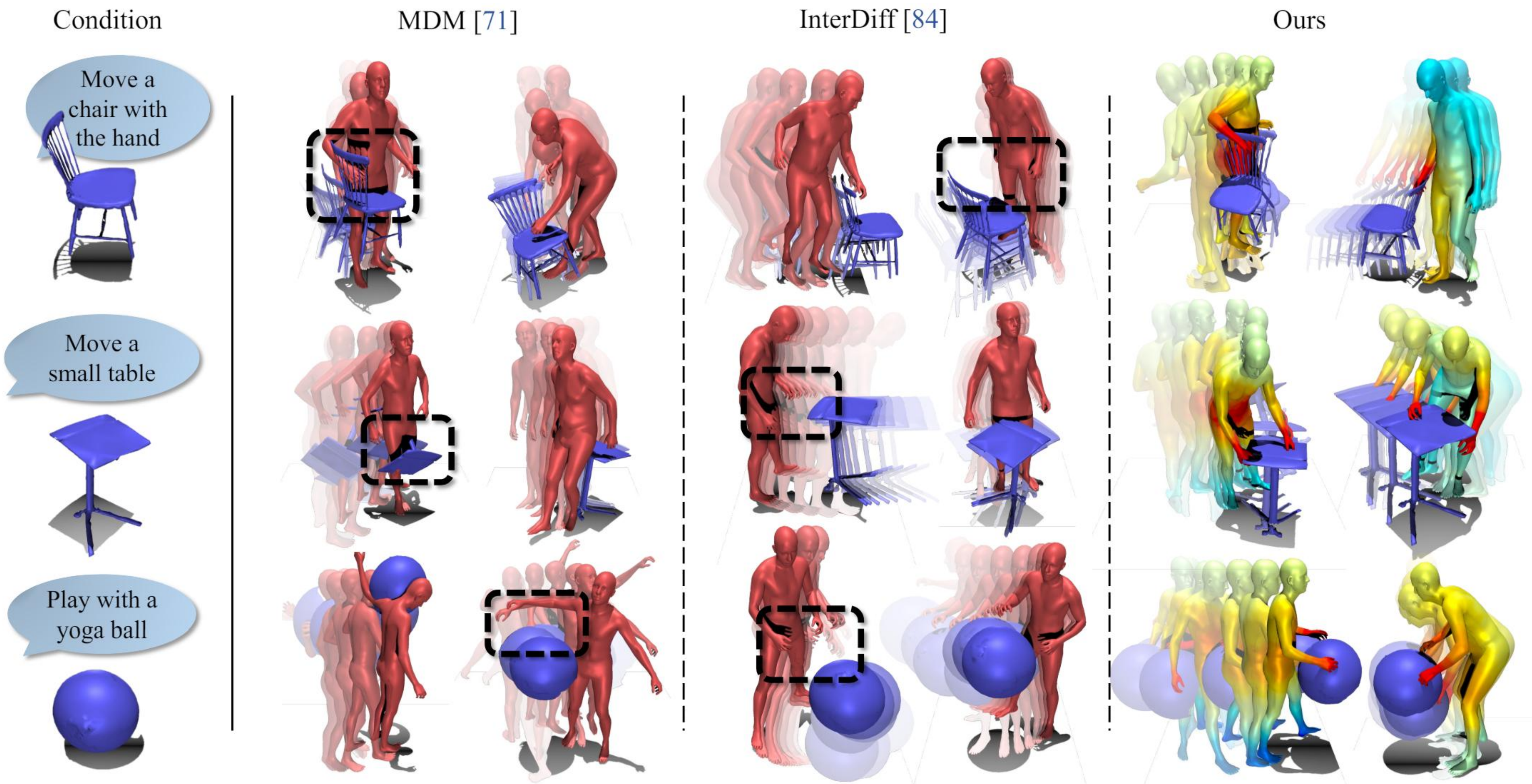


Figure 5. Perceptual User Study. Participants significantly favor our method over baselines, for overall realism and text coherence.

Table 1. Quantitative comparison with state-of-the-art approaches, ... y on the human pose sequence, and motion-cond. denotes predictions ... ject behavior. For metrics with →, results closer to the real distribution are better. Our approach outperforms these baselines in all three settings, indicating a strong learned correlation between human and object motion.

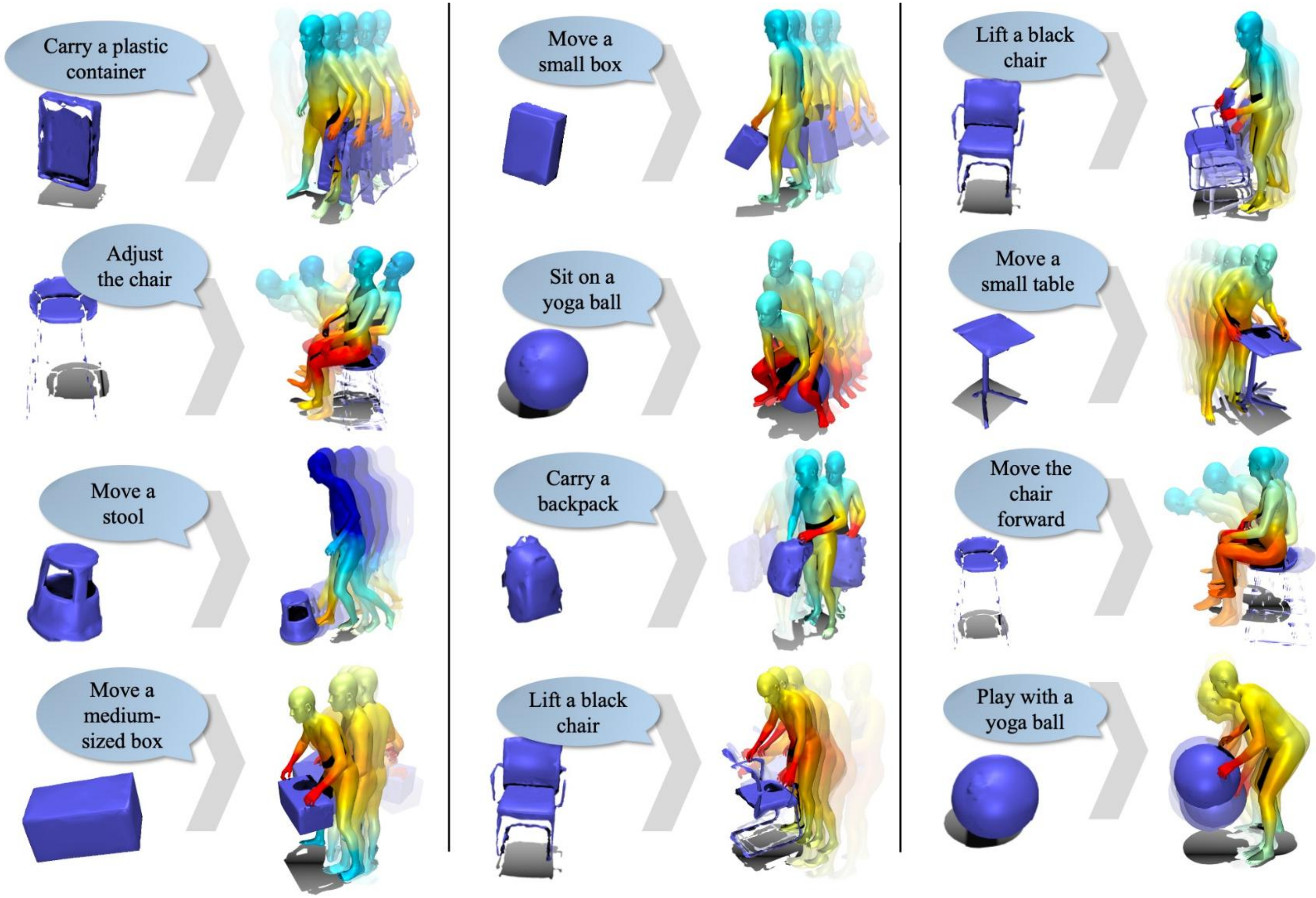
CG-HOI: Experiment

Qualitative Result



CG-HOI: Experiment

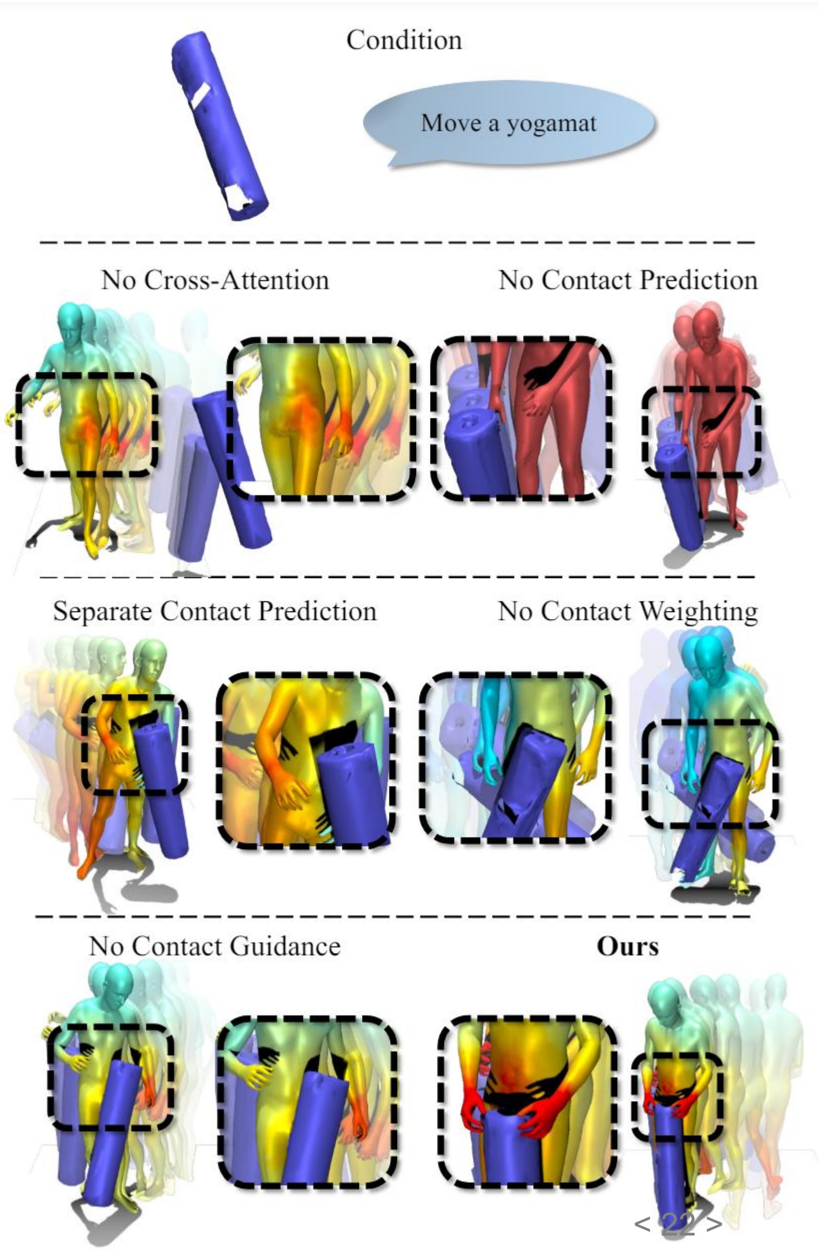
Qualitative Result



CG-HOI: Experiment

Ablation Study

	BEHAVE				CHAIRS			
Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
No cross-attention	0.35	10.44	8.23	7.40	0.49	10.84	12.22	10.64
No contact prediction	0.41	9.64	10.10	6.89	0.41	8.53	11.56	9.15
Separate contact pred.	0.47	8.01	5.12	5.12	0.52	9.34	7.65	4.62
No contact weighting	0.55	8.54	6.52	5.29	0.64	7.55	8.56	5.45
No contact guidance	0.59	7.22	7.84	5.30	0.70	7.41	8.05	5.76
Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.74**	**8.91**	**5.94**



CG-HOI: Experiment

Some Applications

- Human motion generation given object trajectory.

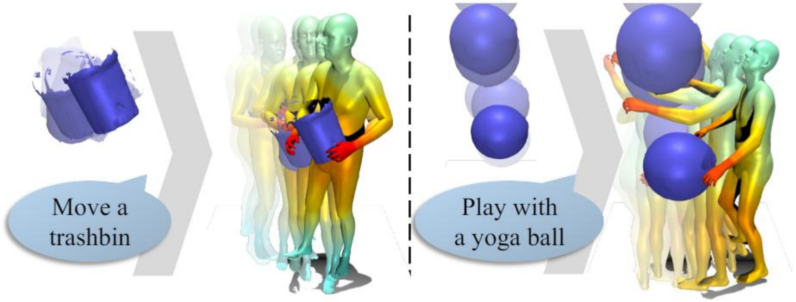
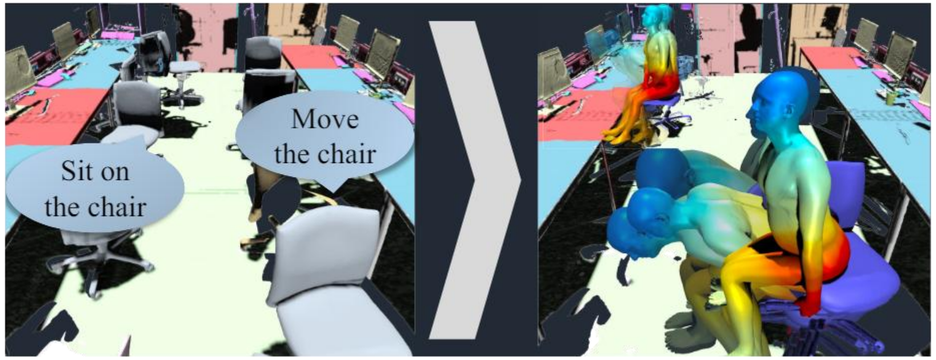


Figure 7. Given an object trajectory at inference time, our method can generate corresponding human motion without re-training.

- Populating 3D scans. Generate realistic human motion sequences given a static scene.



Conclusion

- Contact guided method for diffusion-based HOI generation.
- The first approach to address the task of generating realistic 3D HOI from text.
- Can generate motion that lasts up to 3 seconds.

- Seem to lack the temporal consistency contraint?
- 3D HOI training data is costly.

Thanks!