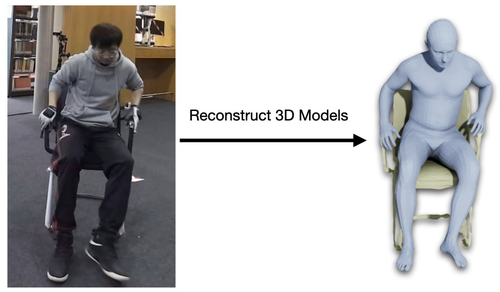
Reconstructing Action-Conditioned Human-Object Interactions Using Commonsense Knowledge Priors

Xi Wang[1]*, Gen Li[1]*, Yen-Ling Kuo[2], Muhammed Kocabas[1,3], Emre Aksan[1], Otmar Hilliges[1]
[1]ETH Zurich, [2]MIT, [3]MPI for Intelligent Systems, Tübingen

3DV 2022
International Conference on 3D Vision

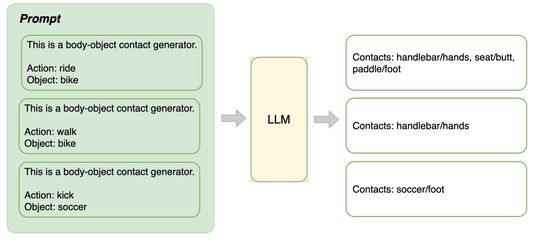
Task

Estimate 3D human-object interaction models from RGB images.



Key Idea

LLMs to extract contact labels



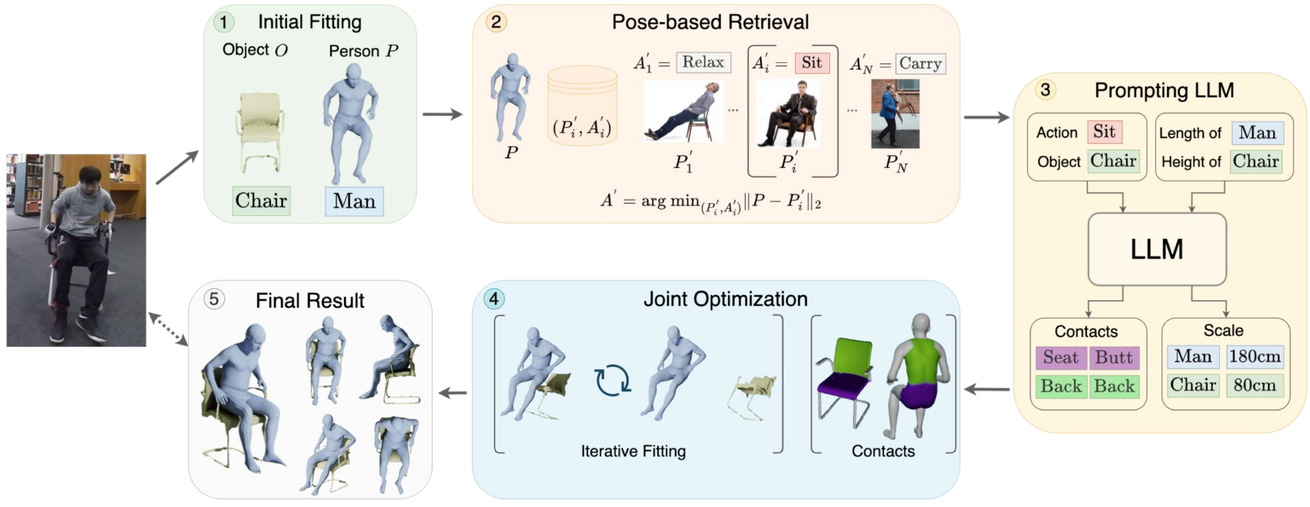
Link part labels

Project page

https://eth-ait.github.io/rhoi/

Method

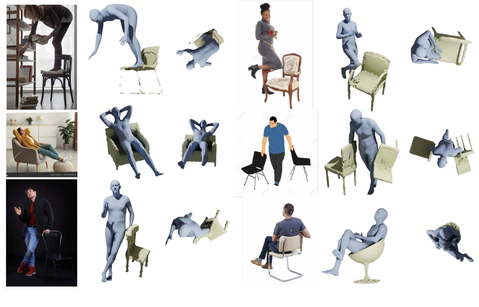
We use large language models to extract action-conditioned contact labels and object size prior, allowing us to reconstruct diverse interactions of different objects.



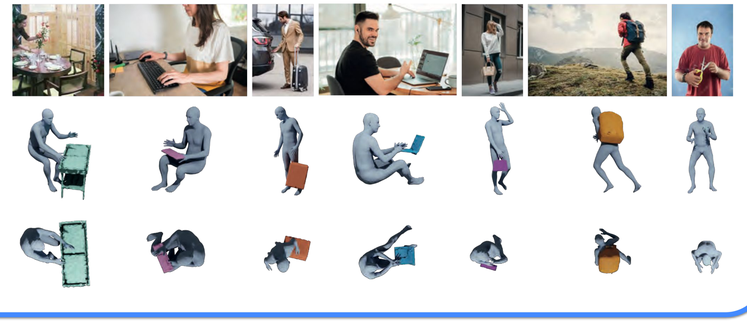
Optimization loss $\mathcal{L} = \mathcal{L}_{contact} + \lambda_1 \mathcal{L}_{normal} + \lambda_2 \mathcal{L}_{penetration} + \lambda_3 \mathcal{L}_{scale} + \lambda_4 \mathcal{L}_{reprojection}$

Generalisability

Diverse interactions

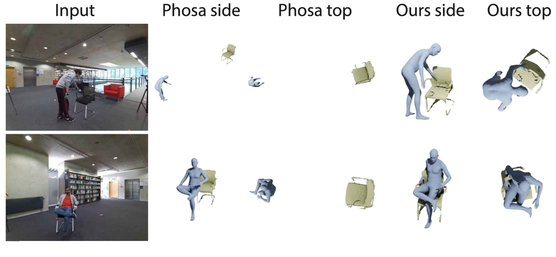


Diverse object categories



Quantitative Evaluation

Quantitative & qualitative evaluation on BEHAVE confirms the benefits of our method over Phosa.



Method	Phosa $\mathcal{H}\downarrow$	Phosa $\mathcal{O}\downarrow$	Ours $\mathcal{H}\downarrow$	Ours $\mathcal{O}\downarrow$
Hand	7.4 (2.5)	83.3 (167.7)	8.1 (2.6)	34.3 (17.6)
Lift	8.0 (2.9)	88.0 (189.0)	8.3 (3.1)	29.1 (16.7)
Liftreal	7.7 (2.2)	80.1 (172.8)	7.8 (2.4)	29.1 (14.3)
Sit	7.1 (2.5)	32.2 (57.6)	7.9 (3.1)	23.4 (12.5)
Sitstand	6.3 (1.5)	26.9 (8.2)	8.1 (2.1)	25.1 (12.7)
Mix	6.5 (2.3)	88.1 (187.7)	7.4 (3.0)	27.5 (16.7)
Avg.	7.2 (2.3)	66.4 (130.5)	7.9 (2.7)	28.1 (15.1)

Ablation

Contact loss significantly improves the accuracy of estimated objects.

Method	$\mathcal{H}$	$\mathcal{O}$
No action	7.9 (2.9)	58.4 (40.3)
No $\mathcal{L}_{contact}$	8.0 (2.7)	103.2 (86.3)
No $\mathcal{L}_{normal}$	7.8 (2.8)	30.7 (14.3)
Ours (full)	7.9 (2.7)	28.1 (15.1)

Failure cases