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]

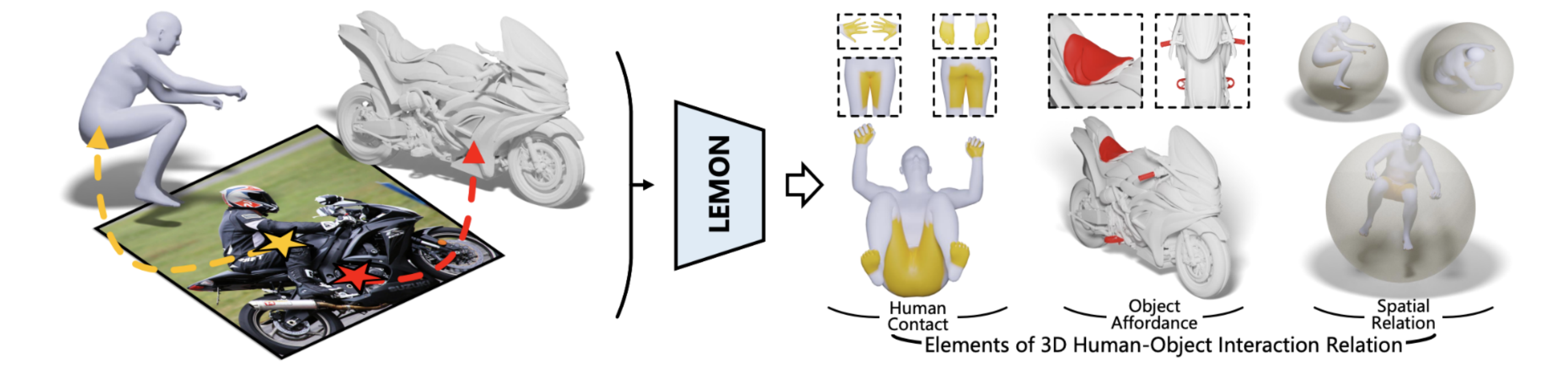
[1]University of Science and Technology of China

[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

CVPR JUNE 17-21, 2024 SEATTLE, WA

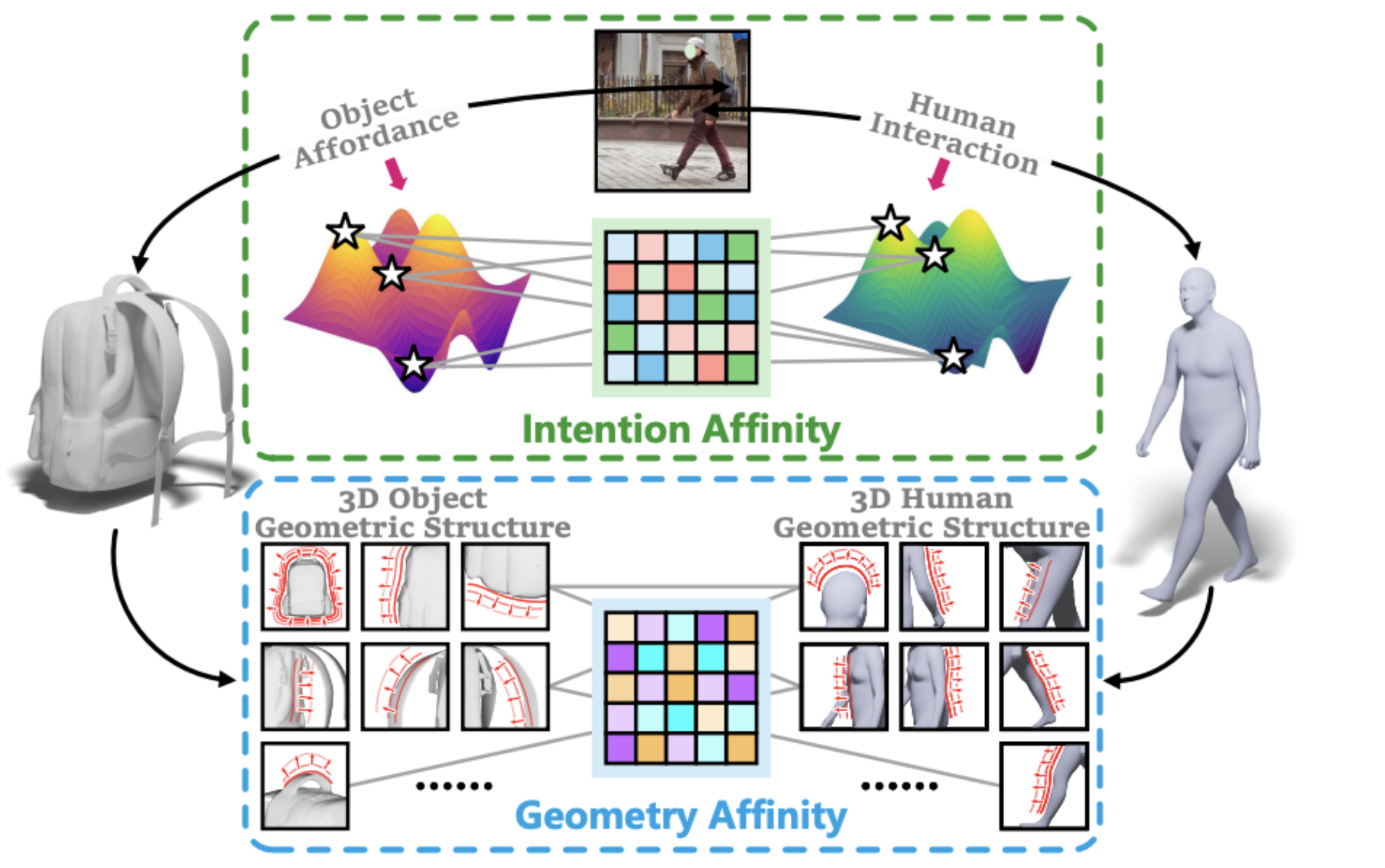
Introduction

Overview of LEMON:

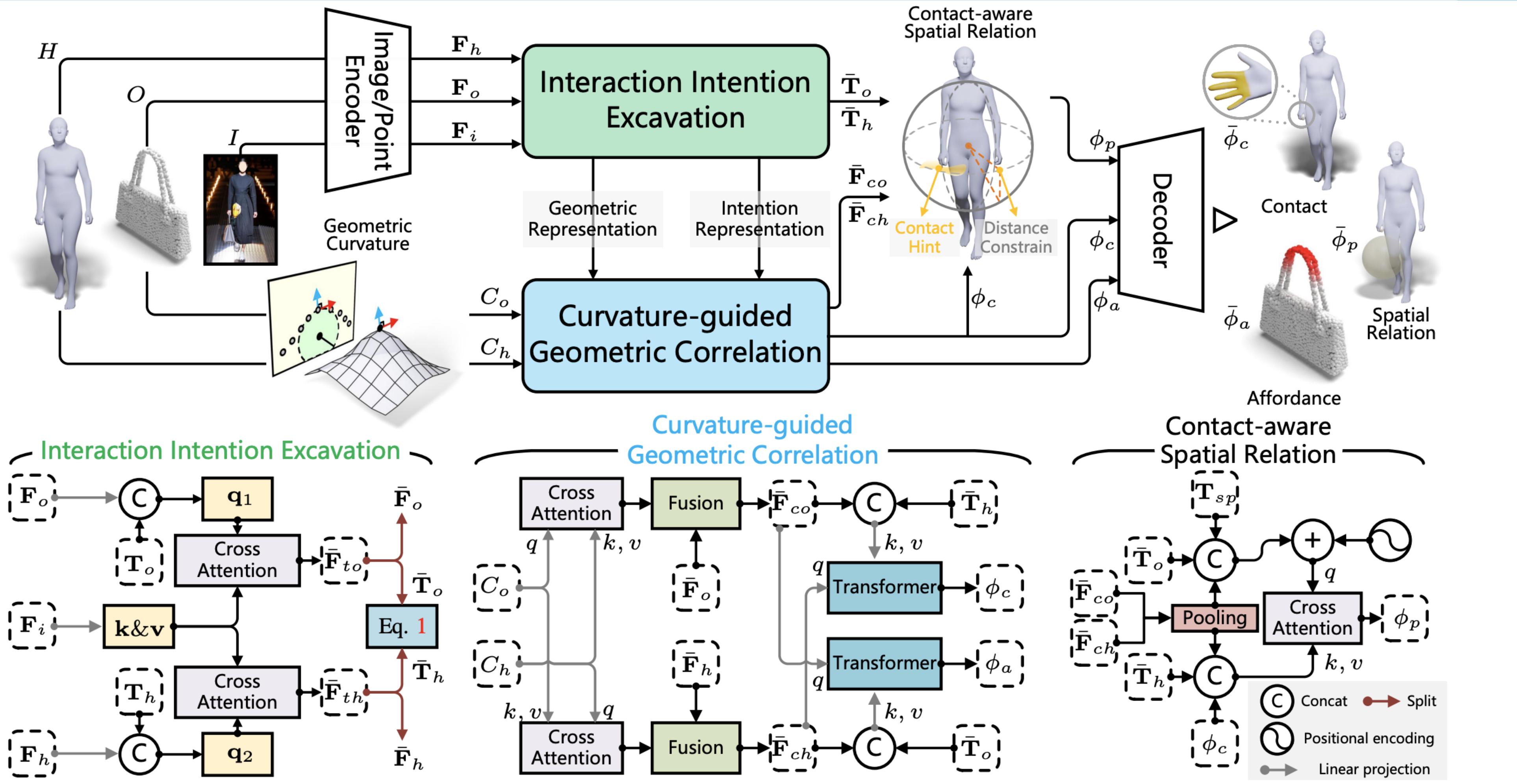


Goal: For an interaction image with paired geometries of the human and object, LEMON learns 3D human-object interaction relation by jointly anticipating the interaction elements, including human contact, object affordance, and human-object spatial relation. Vertices in yellow denote those in contact with the object, regions in red are object affordance regions, and the translucent sphere is the object proxy.

Motivation: The design of daily objects typically adheres to certain human needs, suggesting that object affordances inherently hint at human actions, revealing the intention affinity. Meanwhile, the interacting human and object exhibit matching geometries (either posture or configuration), which presents "how" to interact, arising the geometry affinity.

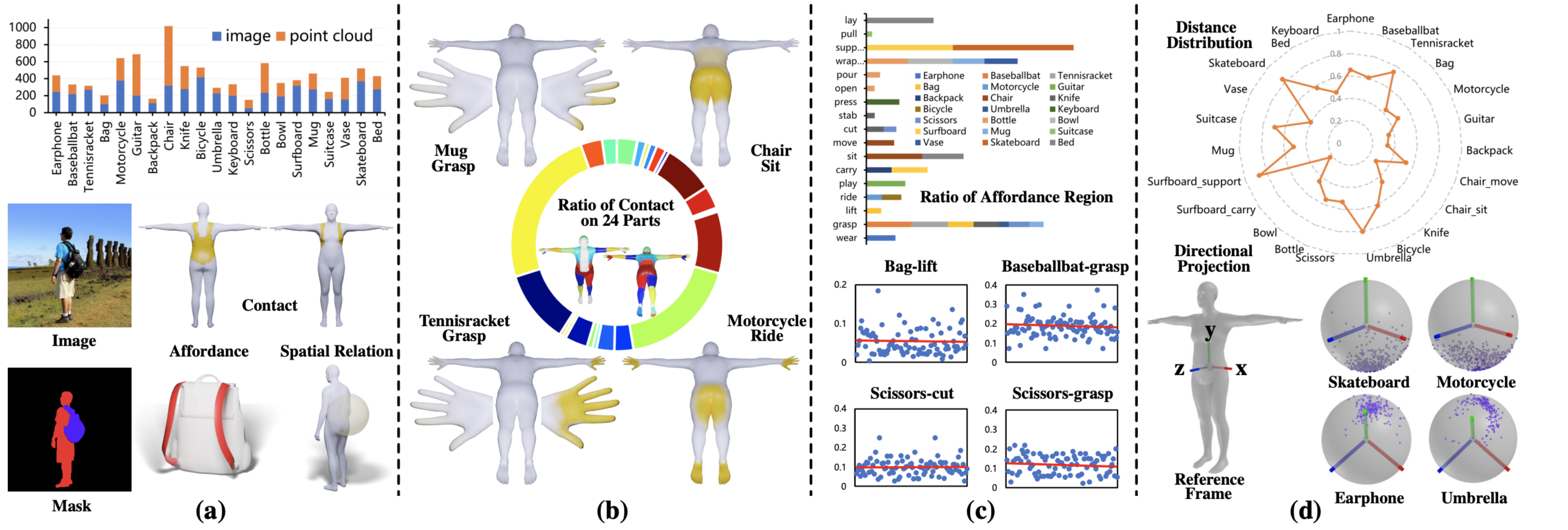


Method



The pipeline of LEMON. Initially, it takes modality-specific backbones to extract respective features $\mathbf{F}_h, \mathbf{F}_o, \mathbf{F}_i$, which are then utilized to excavate intention features ($\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$) of the interaction. With $\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$ as conditions, LEMON integrates curvatures (C_o, C_h) to model geometric correlations and reveal the contact ϕ_c, affordance ϕ_a features. Following, the ϕ_c is injected into the calculation of the object spatial feature ϕ_p. Eventually, the decoder projects ϕ_c, ϕ_a, ϕ_p to the final outputs $\bar{\phi}_c, \bar{\phi}_a, \bar{\phi}_p$. Please refer to the paper for more details.

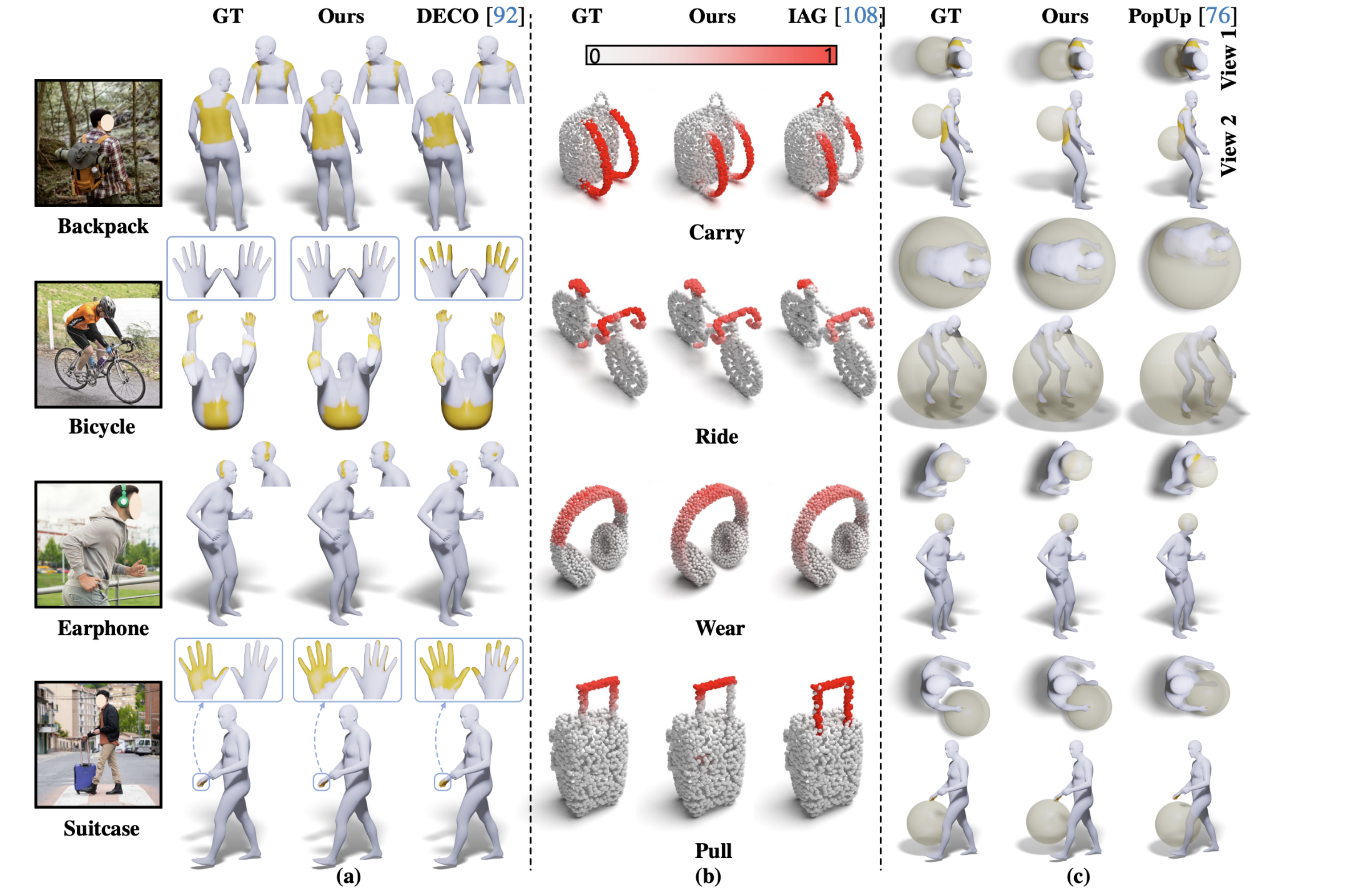
Dataset



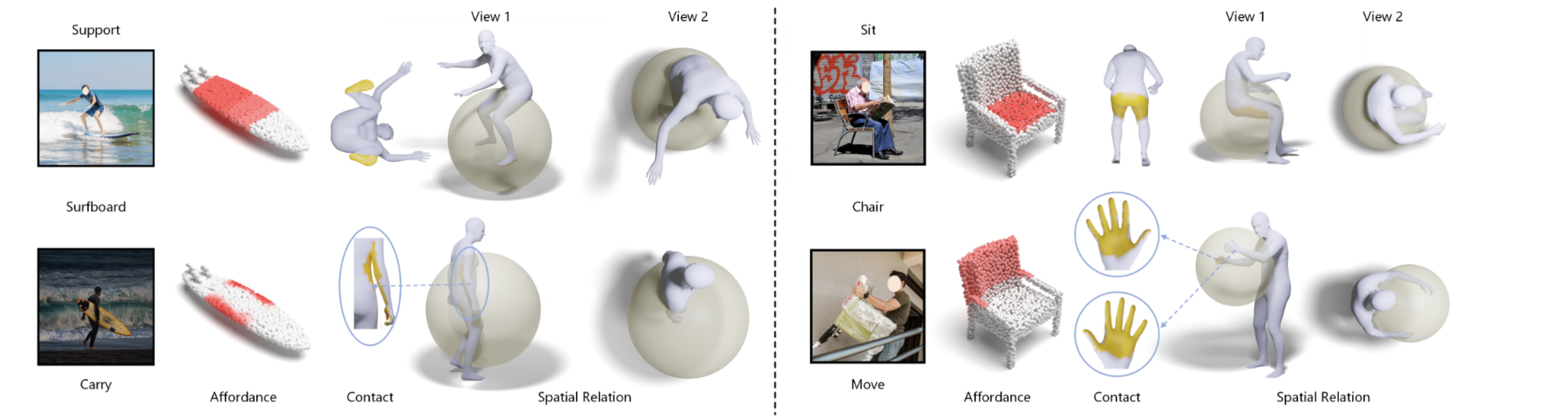
3DIR Dataset. **(a)** The quantity of images and point clouds for each object, and a data sample containing the image, mask, dense human contact annotation, 3D object with affordance annotation, and the fitted human mesh with the object proxy sphere. **(b)** The proportion of our contact annotations within 24 parts on SMPL, and distributions of contact vertices for certain HOIs. **(c)** The ratio of annotated affordance regions to the whole object geometries, and the distribution of this ratio for some categories. **(d)** Mean distances (unit: m) between annotated object centers and human pelvis joints, and directional projections of annotated centers for several objects.

Results

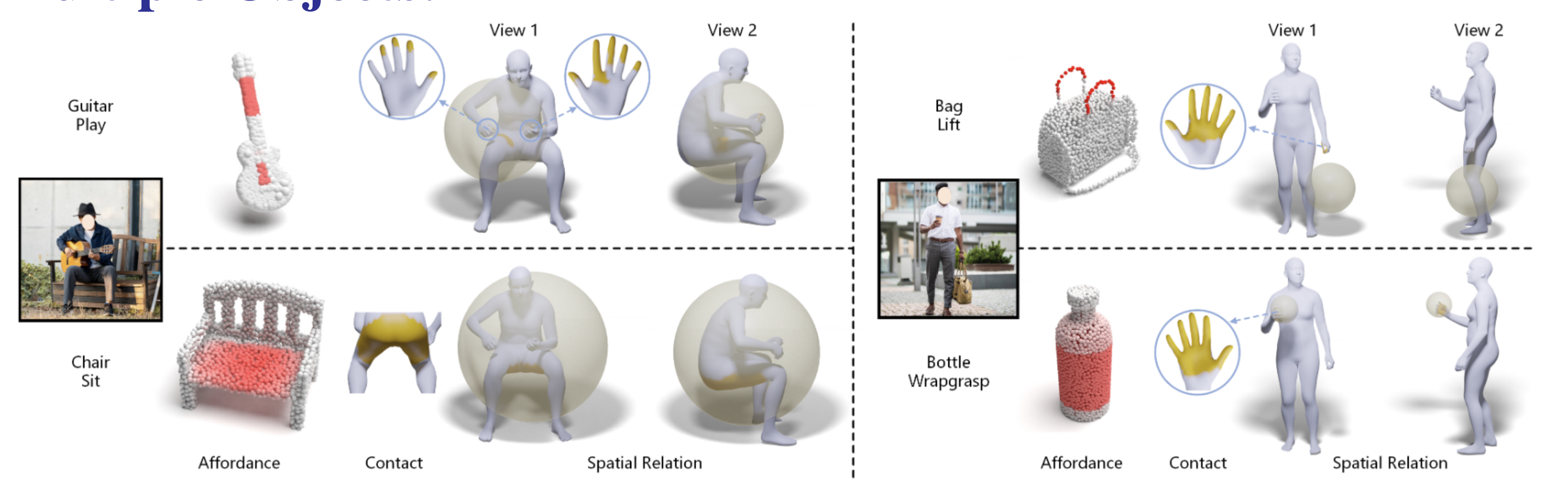
Visual results of comparison:



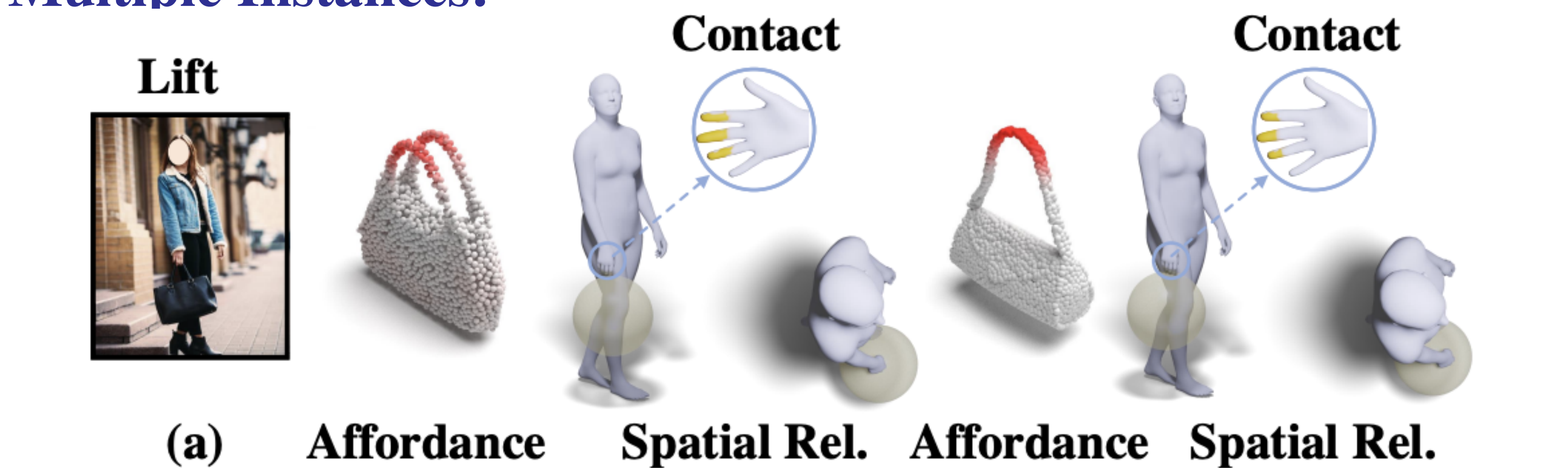
Multiple Interactions:



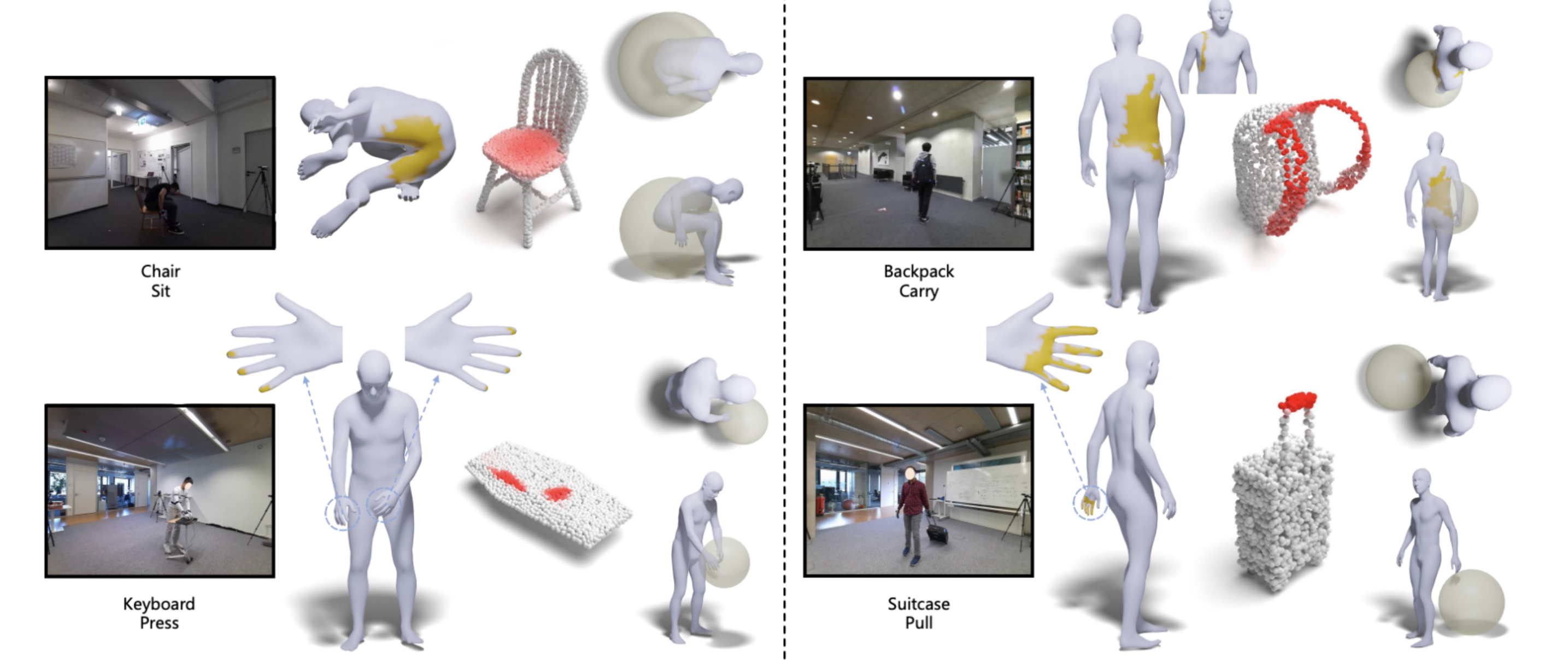
Multiple Objects:



Multiple Instances:



Generalization on unseen BEHAVE dataset:


Chair Sit
Backpack Carry
Keyboard Press
Suitcase Pull