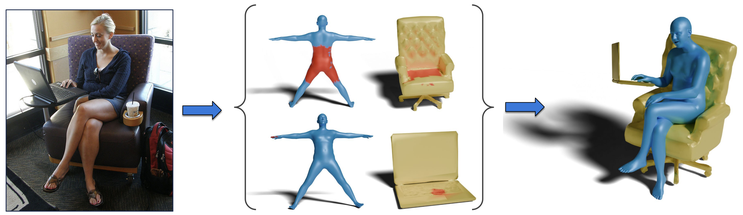
InteractVLM: 3D Interaction Reasoning from 2D Foundational Models

Sai Kumar Dwivedi[1], Dimitrije Antić[2], Shashank Tripathi[1], Omid Taheri[1], Cordelia Schmid[2], Michael J. Black[1], Dimitrios Tzionas[1]

[1] Max Planck Institute for Intelligent Systems, [2] University of Amsterdam, [3] Inria, École normale supérieure

Goal

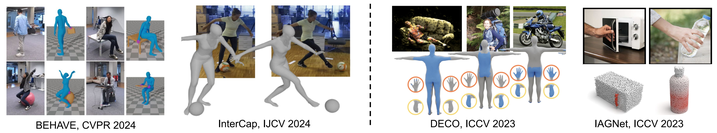


Input RGB Image (In-the-wild) → Contact Estimation on 3D Body & 3D Object → Joint 3D Reconstruction via Contact-Guided Fitting

Need for Contact: Knowing contact points helps resolve occlusion and depth ambiguities in joint reconstruction of human and object from a single RGB image.

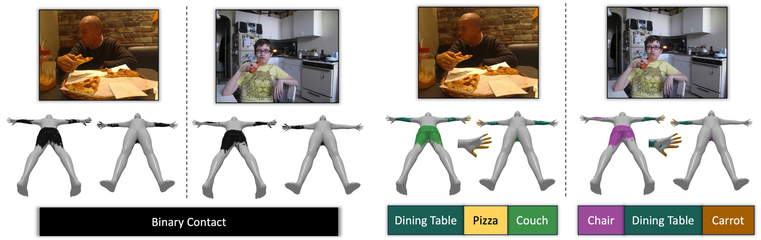
Problem

MoCap or Manual Labelling for Contact Annotations **Does Not scale**



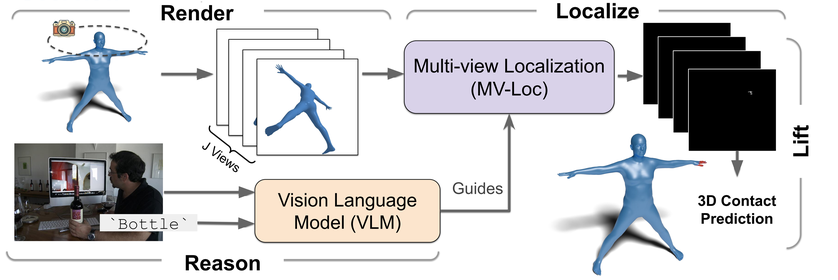
Expensive Motion Capture Systems | Tedious Manual Labelling

Binary Contact **Misses the Rich Semantics** of Human-Object Interaction

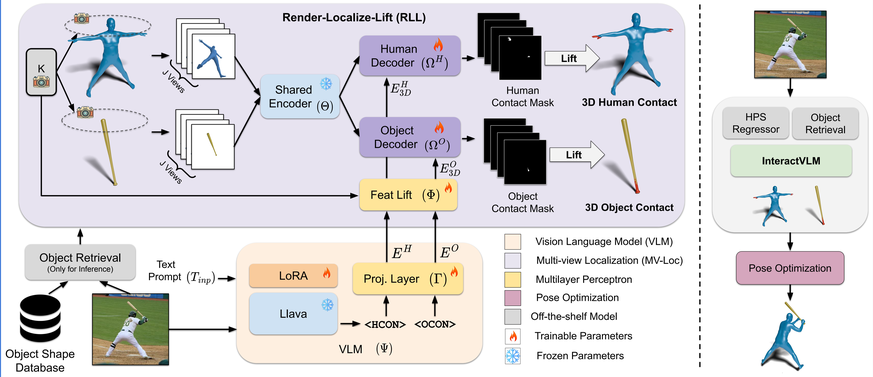


Foundational Vision Language Models (VLMs) can address data scarcity and semantic understanding, but they operate in **2D**, while contact is fundamentally **3D**.

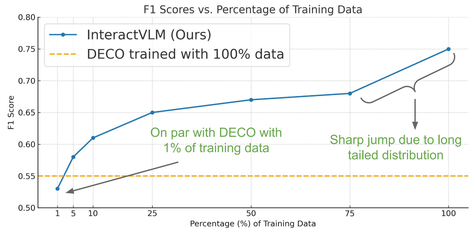
Overview



Method

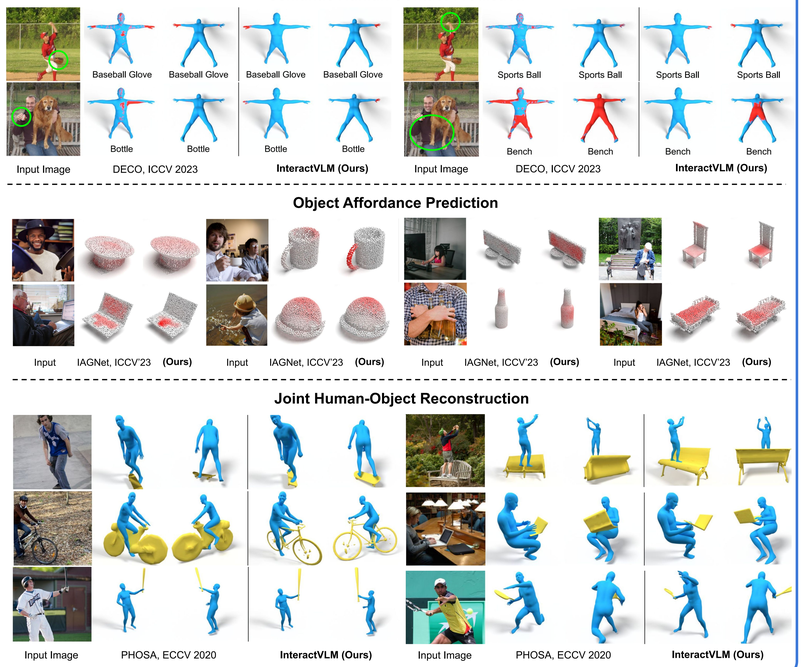


Ablations



	Variants	F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑
MV-Loc	No CamParams	69.4	68.0	71.1
	Concat CamParams	71.9	72.0	71.8
	FeatLift (Φ)	75.6	75.2	76.0
	No MV-Loc	62.3	60.8	63.9
Data	3D Contact Datasets	65.9	64.8	67.0
	+ Contact Parts (text)	74.8	74.5	75.1
	+ HOI-VQA	75.6	75.2	76.0
VLM	No VLM	32.3	30.8	43.0
	VLM-13B-Img	67.2	68.5	66.0
	VLM-13B-NoFT	64.8	65.3	64.2
	VLM-7B	73.3	76.8	73.5
Prompt	Contact parts (fine)	74.8	74.5	75.1
	Contact parts (coarse)	68.4	69.0	67.8
	No object name	71.5	72.1	70.9

Qualitative Results



Results for Contact & Affordance Prediction

Object Categories	Semantic-DECO, ICCV 2023 F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑	Geo (cm) ↓	InteractVLM F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑	Geo (cm) ↓
Accessory	40.1	30.7	**75.6**	21.88	**61.1**	**72.9**	65.6	**6.91**
Daily Obj	26.5	20.1	52.3	60.34	**68.6**	**71.4**	**78.0**	**7.46**
Food	11.7	19.4	12.9	49.61	**66.4**	**66.1**	**77.9**	**7.17**
Furniture	24.5	15.8	**83.7**	29.17	**60.5**	**64.2**	60.1	**6.21**
Kitchen	27.7	24.7	37.2	52.34	**71.8**	**71.3**	**81.1**	**7.61**
Sports	36.4	30.4	80.1	79.21	**77.9**	**76.7**	**83.2**	**7.98**
Transport	52.0	39.1	93.7	31.78	**77.8**	**80.5**	**78.9**	**7.97**

Methods	PIAD, ICCV 2023 SIM (%) ↑	AUC (%) ↑	aIOU (%) ↑	MAE ↓
PFusion, CVPR 2018	43.2	77.50	12.31	1.35
PMF, ICCV 2021	42.5	75.05	10.13	1.41
ILN, SIGGRAPH 2022	42.7	75.84	11.52	1.37
XMF, NeurIPS 2022	44.1	78.24	12.94	1.27
IAGNet, ICCV 2023	54.5	84.85	20.51	0.98
InteractVLM	**62.7**	**86.47**	**21.20**	**0.81**

Resources

Project Page

https://interactvlm.is.tue.mpg.de

Code

https://github.com/saidwivedi/InteractVLM