InteractVLM: 3D Interaction Reasoning from 2D Foundational Models

Sai Kumar Dwivedi[1] Dimitrije Antić[2] Shashank Tripathi[1] Omid Taheri[1]
Cordelia Schmid[3] Michael J. Black[1] Dimitrios Tzionas[2]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Amsterdam, the Netherlands
[3]Inria, École normale supérieure, France

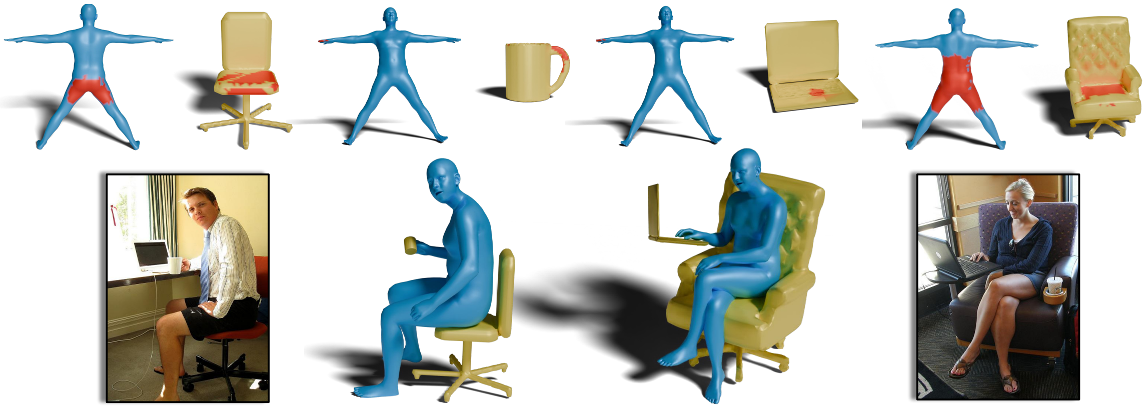

Figure 1. We present InteractVLM, a novel method for estimating contact points on both human bodies and objects from a single in-the-wild image, represented as red patches. Our method goes beyond traditional binary contact estimation methods by estimating contact points on a human in relation to a specified object. We do so by leveraging the broad visual knowledge of a large Visual Language Model.

Abstract

Estimating the 3D pose and shape of interacting humans and objects from single in-the-wild images is important for mixed reality and robotics. This is challenging due to occlusions, depth ambiguities, and widely varying object shapes. Existing work tackles these challenges by exploiting surface contact points on the body and object and using these to guide 3D reconstruction. Unfortunately, obtaining 3D contact annotations requires either expensive 3D ground truth or time-consuming manual labeling. Consequently, obtaining training data at scale is a challenge. We tackle this by developing a novel model called InteractVLM that harnesses the broad visual knowledge of large Visual-Language Models (VLMs). The problem is, however, that these large models do not directly "understand" 3D human-object contact. To address this, we exploit existing small datasets of 3D human-object interaction to fine-tune large models to understand contact. However, this is non-trivial, as such models reason "only" in 2D, while contact is inherently 3D. Thus we introduce a novel "Render-Localize-Lift" module that: (1) embeds 3D body and object surfaces in 2D space via multi-view rendering, (2) trains a novel multi-view localization model (MV-Loc) to infer contacts in 2D, and (3) lifts these to 3D. This lets InteractVLM infer 3D contacts for both bodies and objects from a single in-the-wild image. InteractVLM outperforms existing work on contact estimation and also facilitates 3D reconstruction from an in-the-wild image. To estimate 3D human and object pose, we infer initial body and object meshes, then infer contacts on both of these via InteractVLM, and last exploit these in fitting human and object meshes to image evidence. Results show that our approach performs promisingly in the wild.

1. Introduction

People interact with objects in nuanced ways. Reconstructing these Human-Object Interactions (HOIs) in 3D is central to many applications, from assistive robots to mixed reality. However, achieving this from single images is challenging due to depth ambiguity, occlusions, and the diverse shapes and appearances of objects.

There are methods that estimate 3D human bodies and methods that estimate 3D objects, but few that put these together. Knowing the contacts between these could significantly improve joint reconstruction. Our goal is to infer contact points on both humans and objects from single in-the-wild images, and then use these contacts for jointly reconstructing humans and objects. However, there is a lack of in-the-wild training images paired with ground-truth contact labels for both 3D humans and objects. Acquiring such data is challenging and existing methods do not scale.

The problem gets more challenging due to the complexity of real-world interactions. Humans often contact multiple objects simultaneously; e.g., using a laptop while sitting on a club chair. Yet, current approaches treat contact prediction as simple binary classification; i.e. detecting whether a body part is in contact with "any" object. This simplified assumption fails to capture the rich semantic relationships of multi-object interactions. To address this, we introduce a novel "Semantic Human Contact" estimation task. This involves predicting the contact points on bodies related to particular object given a single in-the-wild image.

To tackle this, as well as to overcome data scarcity, we propose a new paradigm for scaling and reasoning in the wild. Specifically, we observe that large Vision-Language Models (VLMs) can "reason" about in-the-wild images because they are trained on internet-scale data and possess a broad visual knowledge about humans and their interactions with the world. We also observe that this knowledge can be re-purposed for novel tasks by fine-tuning these large models on small datasets. Thus, we exploit VLMs for developing a novel framework, called InteractVLM.

At the heart of InteractVLM lies a reasoning module based on a VLM; see Fig. 2. We enrich this with skills for 3D human-object "understanding," by extending the VLM with a LoRA [25] adaptation layer. As a result, given a color image, this module can be "asked" to produce "reasoning tokens" that facilitate 3D contact localization.

However, exploiting these tokens to localize contact is non-trivial. A natural choice would be to employ a foundational "localization" model [31] that takes these tokens as "guidance," and highlights the 3D contacts. But there exists a key practical problem. Existing foundational models inherently operate only in 2D space, while we need them to do so in 3D. To tackle this, we need to re-cast our problem so it is appropriate for 2D foundational models.

To this end, we develop a novel "Render-Localize-Lift" (RLL) framework that has three main steps; see Fig. 2: (1) We render the 3D shape of a canonical SMPL+H [54] body and an object as 2D images from multiple viewpoints. Furthermore, an object mesh is efficiently retrieved from a large-scale 3D database [10] through OpenShape [43]. (2) We pass the above images into a foundation model to predict 2D contact masks for both the body and the object. (3) We lift the predicted 2D contact points to 3D points via back-projection, i.e., performing the inverse of the first step.

However, even after recasting the problem to 2D by rendering multi-view images, existing foundational models are still not 3D aware, i.e. they treat each view independently, ignoring multi-view consistency. This means contact detections in one view do not necessarily agree with ones in adjacent views. To tackle this, just appending camera parameters to multi-view renderings is insufficient. Instead, we build a novel "Multi-view Localization" model called MV-Loc. This has two steps: (1) We transform the "reasoning token" provided by the VLM with the camera parameters used to render the multi-view images. (2) We ensure multi-view consistency, by lifting the inferred 2D contacts in each view to 3D and computing a 3D loss.

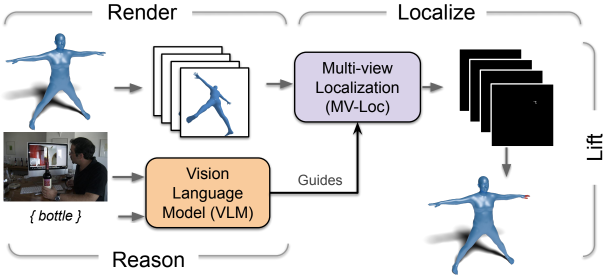

Figure 2. **Overview of InteractVLM.** Given a color image, our VLM performs the core reasoning, and guides a novel MV-Loc model to localize contacts on both bodies and objects in 3D. For visualization purposes, here we show only the body; for details, including object contact, please see Fig. 3.

Our method, InteractVLM, utilizes a VLM in tandem with our novel multi-view localization model, MV-Loc, to perform 3D contact prediction for humans and objects. See Fig. 1 for some examples. We quantitatively evaluate the efficacy of our method for in-the-wild 3D contact prediction on bodies and on objects, using the DAMON [59] and PIAD [66] datasets, respectively. For bodies, we evaluate both for the traditional "binary contact" estimation, and for our new task of "semantic contact" estimation. For all tasks, we find that InteractVLM outperforms the prior work.

Finally, we demonstrate how contact estimated by InteractVLM is used to improve the accuracy of 3D HOI recovery from an in-the-wild image. To this end, we develop an optimization-based method to fit a SMPL-X body mesh and an OpenShape-retrieved object mesh to an image. This is challenging due to occlusions and depth ambiguities. To tackle these, we use InteractVLM's inferred contacts for both humans and objects to guide fitting. To the best of our knowledge, this is the first such approach for estimating 3D HOI for in-the-wild images using inferred contacts.

In summary, we make four key contributions:

1. We build InteractVLM, a novel method that facilitates HOI reconstruction from an in-the-wild image by detecting 3D contacts on both bodies and objects.
2. We demonstrate a way to minimize reliance on 3D contact annotations via exploiting the broad visual knowledge of Vision-Language Models.
3. We build a novel "Multi-view Localization" model that helps in estimating contacts in 3D by transforming the

reasoning of foundational models from 2D to 3D.

4. We introduce the novel "Semantic Human Contact" task for inferring body contacts conditioned on object labels.

Our code and trained models will be released for research.

2. Related Work

2.1. Large Vision-Language Models

Recent advancements in large language models (LLMs) have led to the development of multimodal models that integrate vision and language reasoning. Models like Flamingo [2] and BLIP-2 [36] use cross-attention mechanisms and visual encoders to align image features with text, supporting a variety of vision-language tasks. More recent works, such as VisionLLM [61] and Kosmos-2 [51], use grounded image-text data to enhance spatial understanding, while GPT4RoI [69] introduces spatial box inputs for finer alignment. However, these models typically lack end-to-end segmentation capabilities. To address this limitation, LISA [34] combines vision foundation segmentation models like SAM [30] with multimodal embeddings, enabling language-guided segmentation. PARIS3D [12] extends this approach to referential 3D segmentation by processing multi-view object renders through both SAM and LLaVA [42] for spatially aware segmentation.

Taking inspiration from these approaches, we exploit the language-guided segmentation model for the task of predicting human and object contact in 3D. However, unlike PARIS3D, we process a single RGB image with LLaVA and multi-view renders of the human-object mesh with SAM. We introduce a feature-lifting technique that extends LLaVA's 2D features into 3D using camera parameters, guiding SAM's segmentation across views with cues from a single image. This approach ensures multi-view consistency and efficiently predicts contact affordances, extending the use of multimodal models in 3D reasoning tasks focused on human-object interaction.

2.2. 3D Human and Object from Single Image

Estimating 3D human pose and shape from a single image has evolved from optimization-based methods to learning-based approaches. Optimization methods fit parametric body models like SMPL [46], SMPL-X [50], or GHUM [65] to 2D cues such as keypoints [8], silhouettes [49], or segmentation masks [49]. Learning-based methods either regress body parameters from images or videos [4, 13, 14, 28, 32, 33, 37] or estimate non-parametric bodies as vertices [35, 41], implicit surfaces [47, 55], or dense points [57]. Transformer-based methods [15, 20, 40, 56] have further improved robustness. For 3D object reconstruction from a single image, regression-based methods predict object geometry using meshes, voxels, or point clouds. Diffusion based models [24] utilize large-scale 3D datasets like Objaverse [10] or 2D diffusion models [38, 44, 45, 52] to guide reconstructions. Retrieval-based methods, such as OpenShape [43] and Uni3D [70], have demonstrated robustness in cases with occlusions.

2.3. 3D Human-Object Interactions

Understanding 3D human-object interactions is essential for modeling realistic scenes. Early works focused on hand-object interactions, such as ObMan [23] and FPHA [19], with more recent studies like ARCTIC [16] and HOLD [17] providing more detailed data and reconstruction for hands. For full-body interactions, initial studies involve interactions with scenes, as in PROX [22], and full-body and object interactions, as in BEHAVE [6], GRAB [58], and InterCap [27]. However, these methods often rely on proxy approximations of 3D interactions and depend on capturing devices; while motion capture provides high accuracy, it is not scalable, and multi-camera setups are prone to errors.

Recent methods like DECO [59] capture approximate 3D interactions as contact vertices on meshes by crowd-sourcing annotations through mesh painting. Predicting contact on objects is challenging due to varying object shapes, and currently, there is no dataset that captures object contact in the wild. Therefore, we consider 3D object affordance prediction as a proxy for object contact estimation. While 3D-AffordanceNet [11] introduces affordances not grounded in images, capturing the likelihood of humans interacting with specific parts of an object for a given affordance (e.g., "sit on chair"), PIAD [66] curates RGB images depicting object affordances and designs a network to estimate them. LEMON [67] extends object affordance prediction to include human contact vertices. However, these methods require human contact vertices and corresponding object affordances for training, limiting the number of categories they handle (trained on 21 categories). In contrast, we learn from unpaired human and object interaction data and perform joint interaction reasoning via a cross-attention layer, enabling human interaction reasoning for 80 categories and object affordance prediction for 32 categories.

2.4. Joint 3D Human-Object Reconstruction

Joint reconstruction of humans and objects in 3D has been approached using both regression and optimization methods. Regression-based methods directly predict 3D human-object meshes, as in HDM [63] and CONTHO [48], while others first predict contact points and use test-time optimization to fit the human and object, as in CHORE [62] and PHOSA [68]. Since regression methods rely on limited training data, optimization methods are preferred for in-the-wild scenarios, such as PHOSA [68]. Optimization methods either assume known contacts or infer contacts to fit meshes to the image, but their success heavily relies on initial conditions like accurate contact estimates.

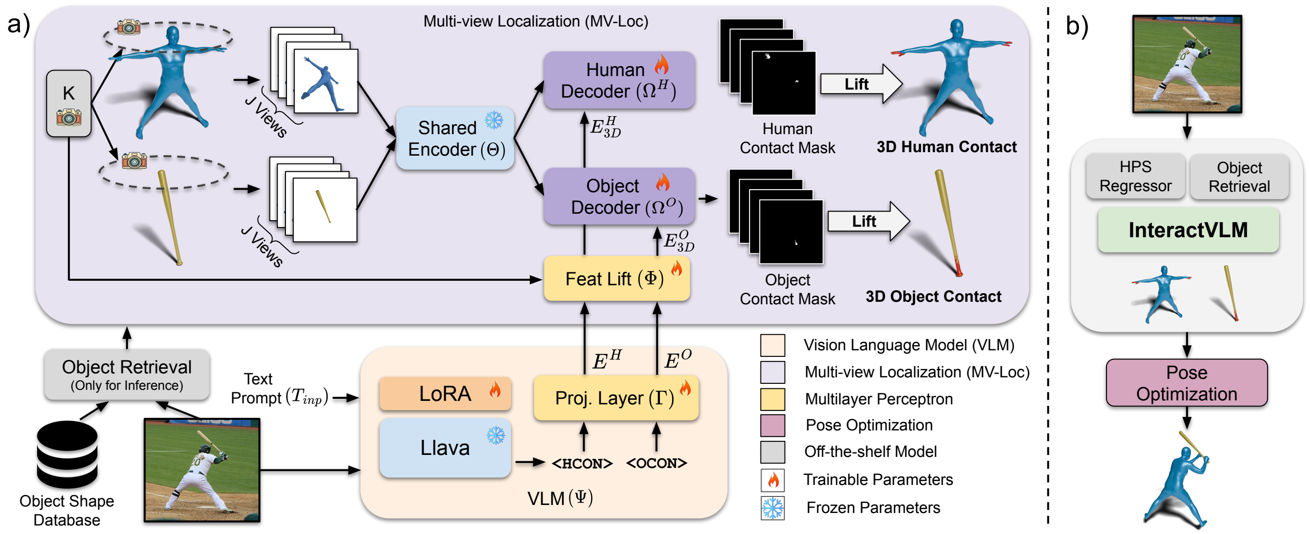

Figure 3. **Method overview.** Given a single in-the-wild color image, our novel InteractVLM method estimates 3D contact points on both humans and objects (a). Then, we reconstruct a 3D human and object in interaction by exploiting these contacts (b). More specifically: **(a) Contact estimation.** Given an image, I, and prompt text, T_{inp}, our VLM, Ψ, produces contact tokens for humans and objects, <HCON> and <OCON>, which are projected (Γ) into feature embeddings, E^H and E^O. These guide a "Multi-View [contact] Localization" model. This renders the 3D human and object geometry via cameras, K, into multi-view 2D renders and passes these to encoder, Θ, while decoders, Ω^H, Ω^O, estimate and highlight 2D contacts in these renders. Then, via camera parameters, K, it transforms via the FeatLift module, Φ, the VLM's features (E^H, E^O) to become 3D-aware (E^H_{3D}, E^O_{3D}) . A final module lifts the detected 2D contacts to 3D. **(b) 3D HOI reconstruction.** For joint human object reconstruction, we use InteractVLM's inferred contacts in an optimization framework.

Our method improves upon these by providing more accurate contact predictions, which in turn facilitate better fitting of human and object meshes to image evidence, improving the realism and accuracy of 3D human-object reconstructions from single images.

3. Method

3.1. Input Representation

Given an image, $I \in \mathbb{R}^{H \times W \times 3}$, InteractVLM estimates 3D contacts for both human bodies and objects.

The human is represented by a SMPL+H [54] 3D body mesh, $\mathcal{H}$, with vertices $V \in \mathbb{R}^{10475 \times 3}$. The body is posed in a canonical star-shape (see details in Sec. 3.4). The human contacts are binary per-vertex labels, $C^H \in \{0, 1\}$.

The object is represented by a 3D point cloud (or mesh), $\mathcal{O} \in \mathbb{R}^{M \times 3}$, with M points. As there are no datasets of natural images paired with 3D contacts for objects, we use a large-scale 3D affordance dataset [66], instead, as affordances are closely related to contact. Specifically, they represent the likelihood of contact on 3D object areas for various purposes. Therefore, for objects we use the terms "affordance" and "contact" interchangeably. The object contacts are continuous per-point values, $C^O \in [0, 1]$. During inference, we retrieve a 3D object shape from a large database [10] via OpenShape [43] conditioned on image I.

3.2. Overview of InteractVLM

The biggest challenge for learning 3D contact prediction in the wild is the limited 3D contact data for humans and objects. To go beyond existing limited datasets, we introduce a novel method, called InteractVLM, that harnesses the commonsense knowledge of large language models.

Specifically, InteractVLM (Fig. 3) has two main components: a Vision Language Model (VLM) and a novel Multi-View contact Localization model (MV-Loc). MV-Loc highlights parts that are in contact for both humans and objects with the VLM's guidance. The input to the VLM (Sec. 3.3) is an image, I, and a prompt that asks the VLM to perform contact detection. The input to MV-Loc (Sec. 3.4) is the 3D geometry of humans and objects, $\mathcal{H}$ and $\mathcal{O}$, respectively.

3.3. Interaction Reasoning through VLM

The VLM module, Ψ, conducts the core interaction reasoning. It takes an input image, I, and prompt text, T_{inp}, and outputs text, $T_{out} = \Psi(I, T_{inp})$. Inspired by the recent LISA [34] model, we expand the VLM's vocabulary with two specialized tokens, <HCON> and <OCON>, for contact information for humans and objects, respectively.

To denote contact, Ψ produces a prompt that includes the above tokens. To aid MV-Loc in localizing contact, we extract the last-layer embeddings of the VLM corresponding to these tokens and pass them through a projection layer, Γ, to obtain the feature embeddings, E^H or E^O. Let T_{gt} be the ground-truth text, and T_{pred} be the predicted one. Then, our token-prediction loss is defined as a cross-entropy loss:

$$\mathcal{L}_{token} = -\sum_{i=1}^{N}(T_{\text{gt}}^{(i)} \cdot \log(T_{\text{pred}}^{(i)})). \quad (1)$$

3.4. Interaction Localization through MV-Loc

We develop a novel MV-Loc module that has a shared image encoder, Θ, and separate decoders, Ω^H and Ω^O, for humans and objects. MV-Loc performs contact localization in using a novel "Render-Localize-Lift" (RLL) framework. To this end, it takes three steps; (1) rendering the 3D shape of humans and objects in 2D, (2) predicting 2D contact maps for both of these, and (3) lifting the 2D contact maps to 3D.

RLL step #1: Render 3D→2D. The input is human and object geometry, namely $\mathcal{H}$ and $\mathcal{O}$, respectively. Both serve as a "canvas" for "painting" detected contacts on these. The body has a default SMPL+H shape in a canonical star-shape pose to minimize self-occlusions when rendering. The object geometry (initialized in Sec. 3.1) is normalized to a unit sphere. Each geometry is rendered from J fixed views with camera parameters, K, to form multi-view renderings, $R^{H,O} = \{R_j\}_{j=1}^{J}$, so that the entire 3D geometry is captured. Since our geometries do not have texture, we color the meshes with normals and point clouds using the NOCS map [60]. This enhances cross-view correspondence, making renderings resemble real images to Θ.

RLL step #2: Localize in 2D. The rendered geometry, $R^{H,O}$, is first sent to the image encoder, Θ, and then passed to decoders, so that the final contact masks, M^H and M^O, get highlighted on it. However, MV-Loc requires spatial and contextual cues for highlighting the contact region. To this end, we use the feature embeddings (Sec. 3.3), i.e. E^H and E^O to guide the contact localization.

However, since the VLM reasons in 2D, these features are not 3D aware. This is a problem, because MV-Loc needs 3D awareness to localize contact consistently across multi-view renderings, We transform the features to "lift" them to "3D" to better guide multi-view localization.

In detail, we design a lifting network, Φ, which takes the 2D $E^{H,O}$, and camera parameters, K, and lifts these to 3D as $E_{3D}^{H,O} = \Phi(E^{H,O}, K)$. Contact masks are defined as:

$$M^{H,O} = \Omega^{H,O}(\Theta(R^{H,O}), \Phi(E^{H,O}, K)). \quad (2)$$

Below, the $\mathcal{H}$ and $\mathcal{O}$ superscripts are dropped for brevity. We calculate loss only on "valid" regions–areas within the outline of rendered geometry; we refer this as M. To encourage overlap between the predicted masks, M, and the ground-truth ones, $\widehat{M}$, particularly in sparse contact regions, we use a focal-weighted BCE loss and a Dice loss:

$$\mathcal{L}_{BCE} = -\alpha(1-p_M)^{\gamma}\log(p_M) - (1-\alpha)p_M^{\gamma}\log(1-p_M), \quad (3)$$

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum M \cdot \widehat{M} + \epsilon}{\sum M + \sum \widehat{M} + \epsilon}, \quad (4)$$

where p_M is the predicted mask probability, α controls class balance, γ adjusts the focus on hard examples, and ϵ is a residual term to prevent division by zero.

RLL step #3: Lift 2D→3D. To lift the inferred 2D contact points to 3D points, we follow the inverse of step #1. Normally, 2D points backproject to 3D lines due to depth ambiguities. In our case the lines intersect with the known 3D geometry that produced the multi-view renders. So, 2D points are lifted to exact 3D points, and by extension 2D contacts, $M^{H,O}$, are lifted to exact 3D contact areas, $C^{H,O}$.

We use a human contact loss $\mathcal{L}_C^H$ that combines a focal loss with sparsity regularization, so that inferred contacts encourage precise predictions in contact regions while discouraging false positives in non-contact areas:

$$\mathcal{L}_C^H = \alpha(1 - p_{hC})^{\gamma}\log(p_{hC}) + \lambda\|C^H\|_1, \quad (5)$$

where p_{hC}, is the contact probability, λ, α and γ are scalar weights. We also use an object contact loss $\mathcal{L}_C^O$ that combines a Dice and Mean Squared Error (MSE) loss:

$$\mathcal{L}_C^O = \mathcal{L}_{Dice}(C^O, \widehat{C}^O) + \beta\|C^O - \widehat{C}^O\|_2^2, \quad (6)$$

where β is a weighting factor, and $\widehat{C}^O$ denotes the ground-truth 3D contacts for objects.

3.5. Implementation Details

Architecture. We use LLaVA [42] as our VLM and SAM [31] for our MV-Loc, with weights pre-trained by LISA [34] for segmentation [34]. The feature lifting network, Φ, contains a spatial-understanding network (two fully-connected layers of size 128 with ReLU activation) followed by view-specific (256-dimensional) transformations and a sigmoid activation. For 3D contact prediction, our MV-Loc model converts 2D segmentation masks to 3D contact points via 2D-to-3D pixel-to-vertex mappings that are precomputed during MV-Loc's rendering step. For details, see ***Sup. Mat.***

Training. To efficiently fine-tune our VLM, we employ LoRA [25] with rank 8. The separate decoders for human and object contact prediction are trained without LoRA, while keeping the image encoder frozen. For training, we use DeepSpeed [3] with mixed precision training (bfloat16), batch size of 8. We train on 4 Nvidia-A100 GPUs for 30 epochs. For more details, please refer to ***Sup. Mat.***

Datasets. We focus on two tasks, i.e., 3D human contact and 3D object affordance prediction, using two in-the-wild datasets, i.e., DAMON [59] and PIAD [66], respectively. For human contact we train and evaluate on DAMON [59]. For 3D object affordances, we train and evaluate on PIAD [66]. We find that exploiting textual descriptions for the contacting body parts, and the object type in contact, helps training. Similarly, adding Visual Question-Answering (VQA) data generated by GPT4o for training images also helps. For details, see ***Sup. Mat.***

Method	F1 (%) ↑	Precision (%) ↑	Recall (%) ↑	Geodesic (cm) ↓
POSA$^{\text{PIXIE}}$ [21] [18]	31.0	42.0	34.0	33.00
BSTRO [26]	46.0	51.0	53.0	38.06
DECO [59]	55.0	65.0	57.0	21.32
InteractVLM	**75.6**	**75.2**	**76.0**	**2.89**

Table 1. Evaluation for “Binary Human Contact” prediction on the DAMON dataset [59]. We compare our InteractVLM model (trained only on this task) with the state of the art.

	Semantic-DECO [59] (Baseline)				**InteractVLM**			
Object Categories	F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑	Geo (cm) ↓	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**

Table 2. Evaluation for “Semantic Human Contact” prediction on the DAMON [59] dataset. For results on each class, see ***Sup. Mat.*** The “Semantic-DECO” baseline extends DECO for our new task.

Unlike LEMON [67], which requires paired human-object geometry for training using the 3DIR dataset [67], we use unpaired data. This enables us to scale for many human and object categories not addressed by prior methods. For the final joint human-object reconstruction task, we combine DAMON [59], PIAD [66], 3DIR [67] and all textual descriptions about body parts, contact type and HOI.

Evaluation metrics. For human contact prediction, following Tripathi et al. [59], we report the F1, precision, and recall scores using a threshold of 0.5, and a geodesic distance measuring spatial accuracy. For object contact prediction, following Yang et al. [66], we report the Similarity (SIM), Mean Absolute Error (MAE), Area Under ROC Curve (AUC), and average Intersection over Union (IOU).

4. Experiments

4.1. “Binary Human Contact” Estimation

This task refers to estimating contact areas on the body via binary classification of its vertices, ignoring the number or type of objects involved. We train and evaluate on the DAMON [59] dataset, and report results in Tab. 1.

InteractVLM significantly outperforms all previous methods achieving a 20.6% improvement in F1 score. Although, here, InteractVLM is trained on the same data as DECO, it goes beyond this by harnessing the commonsense knowledge of a large foundation model.

In ***Sup. Mat.*** we also evaluate on the 3DIR [67] dataset and compare with the LEMON method [67]. Even though LEMON uses paired human-object data, InteractVLM performs on-par with it despite training on human-only data.

	Variants	F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑
Masks $R^{H,O}$	Size 512 × 512	70.7	70.1	71.4
	Size 1024 × 1024	75.6	75.2	76.0
VLM Emb.	No CamParams	69.4	68.0	71.1
	Concat CamParams	71.9	72.0	71.8
	FeatLift (Φ)	75.6	75.2	76.0
Losses	Whole Mask	69.3	68.7	70.0
	Valid Mask	72.6	71.2	74.0
	+ 3D Contact Loss	75.6	75.2	76.0
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

Table 3. Ablation study for the effect of different InteractVLM components and design choices. We evaluate for “Binary Human Contact” prediction on the DAMON dataset [59]. All baselines are trained only on this task and dataset.

	PIAD-Seen [66]				PIAD-Unseen [66]			
Methods	SIM (%) ↑	AUC (%) ↑	aIOU (%) ↑	MAE ↓	SIM (%) ↑	AUC (%) ↑	aIOU (%) ↑	MAE ↓
PMF [71]	42.5	75.05	10.13	1.41	33.0	60.25	4.67	2.11
ILN [9]	42.7	75.84	11.52	1.37	32.5	59.69	4.71	2.07
PFusion [64]	43.2	77.50	12.31	1.35	33.0	61.87	5.33	1.93
XMF [1]	44.1	78.24	12.94	1.27	34.2	62.58	5.68	1.88
IAGNet [66]	54.5	84.85	20.51	0.98	35.2	71.84	7.95	1.27
InteractVLM	**62.7**	**86.47**	**21.20**	**0.81**	**41.4**	**75.45**	**8.50**	**0.99**

Table 4. Evaluation for “Object Affordance Prediction” on the PIAD [66] dataset. We compare our InteractVLM model (trained only on this task) with the state of the art.

4.2. “Semantic Human Contact” Estimation

In real-life interactions, multiple objects can be contacted by different body areas concurrently. Thus, we introduce a novel task called “Semantic Human Contact” prediction. We evaluate on DAMON [59] and report results in Tab. 2; for a finer-grained version of this table, see ***Sup. Mat.***

To establish a baseline, we adapt the DECO model [59] that detects binary contacts and turn it into a multi-class prediction model, called Semantic-DECO. Due to DAMON’s limited training data, this has poor performance. Instead, as discussed in Sec. 4.1, InteractVLM learns effectively from this data, by also leveraging the commonsense of foundation models. Thus, it significantly outperforms Semantic-DECO. Qualitative results reflect this finding; see Fig. 4. Our model captures detailed, accurate contact regions, whereas Semantic-DECO often highlights false-positive areas that differ from the actual contact regions.

4.3. Ablations

We conduct extensive ablation studies to evaluate the contribution of InteractVLM’s main components. Specifically, we evaluate for “Binary Human Contact” prediction on the DAMON [59] dataset, and report results in Tab. 3.

Increasing the mask resolution (first row) from 512 to

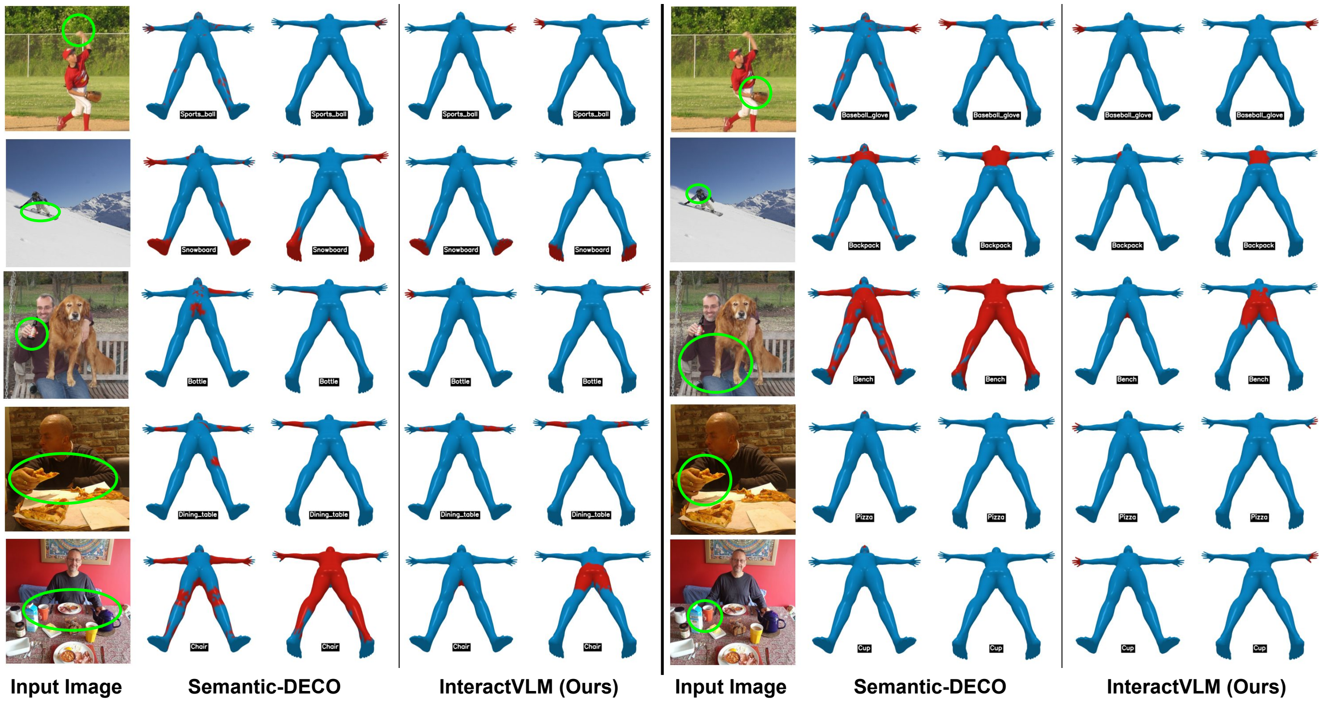

Figure 4. **Semantic Human Contact estimation (Sec. 4.2).** Given a single image and an object label, our InteractVLM method infers body contacts for this object. InteractVLM outperforms a Semantic-DECO [59] baseline. Objects are circled in green; contacts are red.

1024 yields a significant improvement of 4.9% in F1 score, highlighting the importance of fine-grained spatial information for contact detection. For VLM-feature embedding (second row), our FeatLift (Φ) network outperforms simply concatenating camera parameters by 3.7%, demonstrating its effectiveness in incorporating viewpoint information. Using the "valid mask" regions (third row) for training instead of "whole masks" improves performance by 3.3%, while the addition of our 3D contact loss further boosts F1 score by 3%. The biggest contribution comes from forcing the VLM to predict body-part names (fourth row) in text form (+8.9%). This helps the VLM to attend to contact regions in the image and pass relevant spatial and contextual cues to MV-Loc for contact localization. Adding VQA data generated from GPT4o further improves performance.

4.4. Object Affordance Prediction

We evaluate the performance of our InteractVLM model on predicting object affordances. This involves identifying regions on objects of possible contact for a certain interaction intent, such as "sitting" on a chair or "moving" it. We train and evaluate on the PIAD [66] dataset. We compare against SotA methods, and report results in Tab. 4.

Note that we evaluate on object instances that are both seen during training ("PIAD-Seen" column) and unseen ("PIAD-Unseen" column). Our model demonstrates a significant improvement over SotA methods in terms of similarity (SIM), area under the curve (AUC), and mean absolute error (MAE) for both seen and unseen objects. Thus, our method is effective not only for estimating human contact (Secs. 4.1 and 4.2), but also for object affordances.

5. Joint Human-object Reconstruction

We demonstrate the usefulness of 3D contacts inferred by our InteractVLM for reconstructing a 3D human and object in interaction from a single in-the-wild image, I.

Initializing 3D body pose & shape, object shape. We use OSX [39] to estimate a 3D SMPL-X body mesh, $M^{\mathcal{H}}$, and OpenShape [43] to retrieve a 3D object mesh, M^{O}, from the Objaverse [10] database that best matches the image.

Initializing 3D object pose. We apply InteractVLM on image I to predict 3D contact body and object vertices, $C^{\mathcal{H}}, C^{O}$. Then, we solve for object pose, $\{R^{O}, t^{O}\}$ by applying the ICP [5] algorithm to snap the 3D object contact points, C^{O}, onto body ones, $C^{\mathcal{H}}$. To avoid false correspondences, the 3D normals of contact points must be compatible; they should have similar angles but opposite directions.

Optimizing 3D object pose. Given the above initialization, we optimize over object rotation, R^{O}, translation, t^{O}, and scale, s^{O}, via render-and-compare by minimizing:

$$E = E_M + \lambda_C E_C, \tag{7}$$

$$E_M = IoU(\widehat{m}, m) + \|\widehat{m}^c - m^c\|_2, \tag{8}$$

$$E_C = \frac{1}{|C^{\mathcal{H}}||C^{O}|} \sum_{i\in|M^{\mathcal{H}}|} \sum_{j\in|M^{O}|} C_i^{\mathcal{H}} C_j^{O} \|v_i - v_j\|_2, \tag{9}$$

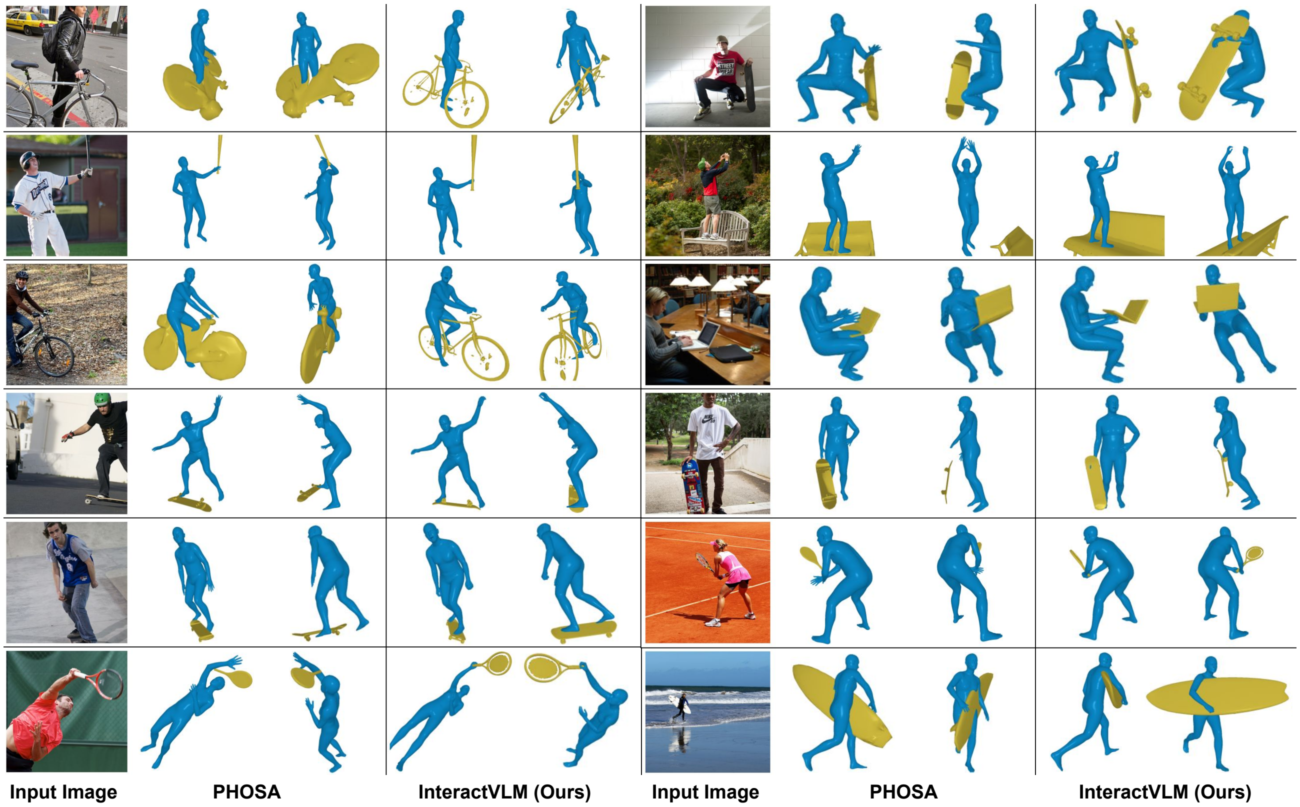

Figure 5. **3D HOI reconstruction (Sec. 5).** We build an optimization method that fits a SMPL-X body and OpenShape-retrieved object to an in-the-wild image. We evaluate against the SotA method PHOSA [68]. Reconstruction is guided by InteractVLM-inferred contacts.

where E_M is a mask loss, E_C is a contact losses, IoU is intersection-over-union, $\widehat{m}, m$ are hypothesis and ground-truth masks, respectively, $\widehat{m}^c, m^c$ are the corresponding mean mask pixels, $|C^{\mathcal{H}}|$, $|C^{O}|$ are the number of contact vertices on the human and object, $|M^{\mathcal{H}}|$, $|M^{O}|$ are the number of mesh vertices, v_i, v_j are the i-th and j-th human and object vertices, while C^{O}_i, $C^{\mathcal{H}}_j$ indicate whether vertices are in contact or not. Given the image I, we extract the object mask, m, via SAM [31], and a depth map, D, via Depth Pro [7].

Intuitively, we use the E_M loss to align the 3D object to the image, while in E_C the predicted 3D contacts "anchor" the 3D object onto the body so they interact realistically. We perform iterative optimization via Adam [29]. OSX produces reasonable bodies, so we keep these fixed. The object is updated in each iteration – we render depth, $\widehat{D}$, and masks, $\widehat{m}$, via a PyTorch3D [53] differentiable renderer.

Qualitative Results. We reconstruct 3D human-object interaction (HOI) from an image. We show results in Fig. 10, and compare with PHOSA [68], the most related SotA method. InteractVLM's reconstructions look more realistic. Note that PHOSA uses handcrafted contacts on humans and bodies. Instead, InteractVLM infers 3D contacts on both of these from the image. These play a crucial role for guiding 3D reconstruction under occlusions and depth ambiguities.

Perceptual Study. There exists no in-the-wild dataset with 3D ground truth for HOI, so we conduct a perceptual study via Amazon Mechanical Turk for evaluation. Specifically, we evaluate the realism of our reconstructions against ones of PHOSA. We randomly select 55 images for which PHOSA has handcrafted contact annotations. For each image, participants are shown (with random swapping) reconstructions generated by our method and by PHOSA, and are asked to select the one that best represents the image. Our reconstructions are preferred 62% of the time.

6. Conclusion

We present InteractVLM, a novel method for estimating 3D human-object interactions from single in-the-wild images by inferring contact points on both humans and objects. We leverage the broad visual knowledge of Vision-Language Models to minimize the reliance on expensive 3D contact annotations. Specifically, we introduce a novel "Render-Localize-Lift" (RLL) framework and a novel multi-view localization model (MV-Loc) to adapt 2D foundation models for 3D contact estimation. We outperform existing work on contact estimation and introduce a new "Semantic Human Contact" estimation task for inferring body contacts conditioned on object labels. This goes beyond traditional binary contact estimation, which fails to capture rich semantic relationships of multi-object interactions. InteractVLM is the first approach to use inferred contact points on both bodies and objects for joint 3D reconstruction from single in-the-wild images.

References

[1] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:37349–37362, 2022. 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[3] Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, A. A. Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. DeepSpeed inference: Enabling efficient inference of transformer models at unprecedented scale. *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2022. 5

[4] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[5] Paul J. Besl and Neil D. McKay. A method for registration of 3D shapes. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 14:239–256, 1992. 7

[6] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[7] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. *arXiv:2410.02073*, 2024. 8

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[9] Honghua Chen, Zeyong Wei, Yabin Xu, Mingqiang Wei, and Jun Wang. ImLoveNet: Misaligned image-supported registration network for low-overlap point cloud pairs. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 1–9, 2022. 6

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 2, 3, 4, 7

[11] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3D AffordanceNet: A benchmark for visual object affordance understanding. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[12] John Doe, Jane Smith, and Alice Brown. Paris3d: Language-guided 3d segmentation with multimodal models. *arXiv:2406.08394*, 2024. 3

[13] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[14] Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J. Black, and Dimitrios Tzionas. POCO: 3D pose and shape estimation using confidence. In *International Conference on 3D Vision (3DV)*, 2024. 3

[15] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[16] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[17] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[18] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 6

[19] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[20] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[21] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 6

[22] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 3

[23] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated ob-

jects. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[24] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. *arXiv: 2311.04400*, 2023. 3

[25] J. E. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2021. 2, 5

[26] Chun-Hao Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13264–13275, 2022. 6, 12

[27] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision (IJCV)*, 132(7):2551–2566, 2024. 3

[28] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52, 2021. 3

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 8

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 3

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 2, 5, 8, 12

[32] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 3

[33] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 3

[34] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2024. 3, 4, 5

[35] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 3

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742, 2023. 3

[37] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[38] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-resolution text-to-3D content creation. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[39] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168, 2023. 7

[40] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[41] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3, 5, 12

[43] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 4, 7

[44] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[45] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. *International Conference on Computer Vision (ICCV)*, 2023. 3

[46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34 (6):248:1–248:16, 2015. 3

[47] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[48] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3D human and object via contact-based refinement transformer. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[49] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape es-

timation. In *International Conference on 3D Vision (3DV)*, 2018. 3

[50] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3

[51] Bingbing Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023. 3

[52] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[53] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 8

[54] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6), 2022. 2, 4, 12

[55] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[56] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-one-stage expressive human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[57] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3

[58] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[59] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *International Conference on Computer Vision (ICCV)*, pages 8001–8013, 2023. 2, 3, 5, 6, 7, 12, 13, 14

[60] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[61] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[62] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[63] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[64] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. PointFusion: Deep sensor fusion for 3D bounding box estimation. In *International Conference on Computer Vision (ICCV)*, pages 244–253, 2018. 6

[65] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[66] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3D object affordance from 2D interactions in images. In *International Conference on Computer Vision (ICCV)*, pages 10871–10881, 2023. 2, 3, 4, 5, 6, 7, 12, 13, 15

[67] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. LEMON: Learning 3D human-object interaction relation from 2D images. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 6, 12, 13

[68] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 8

[69] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. GPT4RoI: Instruction tuning large language model on region-of-interest. *arXiv:2307.03601*, 2023. 3

[70] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[71] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3D lidar semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 6

InteractVLM: 3D Interaction Reasoning from 2D Foundational Models

Supplementary Material

7. Human Contact Prediction

7.1. Evaluation on the 3DIR Dataset

We evaluate our method against several state-of-the-art approaches for human contact prediction on the 3DIR dataset [66] as shown in Tab. 5. Our method outperforms methods that are trained on 3D training data for only humans, while it is on par with methods that use 3D data for both humans and objects. Moreover, by eliminating the requirement for paired human-object contact training data, our method can be trained on more categories than prior work, as unpaired datasets are more varied. This makes our method more practical for real-world applications.

7.2. Semantic Human Contact per Object Category

We evaluate our method's performance on "semantic human contact" prediction across a diverse set of object categories from the DAMON dataset, as shown in Tab. 6. Results for high-level categories are presented in the main paper. We compare our method against "Semantic-DECO", which is our extension of the existing DECO [59] model for this new task. Our method significantly outperforms Semantic-DECO in terms of F1-score for all categories. It also demonstrates strong performance across a wide range of object categories, from large objects like furniture (couch: 62.1% F1, chair: 70.3% F1) to small objects for sports (baseball glove: 93.6% F1, tennis racket: 82.3% F1).

8. Analysis of reliance on 3D annotations

To analyze our model's efficiency in utilizing 3D supervision data, we conduct experiments with varying amounts of training data from the DAMON dataset. Fig. 6 illustrates our model's performance across different percentages of DAMON training data, ranging from 1% to 100%. Remarkably, our method achieves an F1 score of 0.53 with just 1% of the training data, nearly matching DECO's performance (0.55 F1) that requires the full dataset. When trained on 5% of the data, our approach reaches an F1 score of 0.58, already surpassing DECO [59]. This performance gap continues to widen as we increase the training data, ultimately achieving an F1 score of 0.75 with 100% of the data, demonstrating a substantial improvement over DECO. These results highlight our method's efficient use of 3D supervision, achieved by leveraging the rich visual understanding of foundation models. The strong performance with limited training data suggests that our approach is particularly valuable since obtaining 3D annotations are very expensive.

Method	3D Supervision		F1	Prec.	Rec.	Geo.
	Human	Obj.	(%) ↑	(%) ↑	(%) ↑	(cm) ↓
BSTRO [26]	✓	✗	55.0	57.0	58.0	28.58
DECO [59]	✓	✗	69.0	70.0	72.0	15.25
LEMON-P [67]	✓	✓	77.0	76.0	81.0	9.02
LEMON-D [67]	✓	✓	78.0	78.0	**82.0**	7.55
InteractVLM	✓	✗	**78.4**	**82.5**	76.3	**6.73**

Table 5. Evaluation for "Binary Human Contact" prediction on the 3DIR dataset [67]. Note that LEMON is trained with paired human-object contact data from 3DIR dataset. Instead, for this task, InteractVLM is only trained with human contact data from the same dataset.

9. Implementation Details

9.1. Architecture

InteractVLM has two major blocks; a reasoning module, Ψ, based on LLaVA-v1 [42] and a novel multi-view localization model, MV-Loc, based on SAM [31]. MV-Loc has 2 components; a shared encoder, Θ and two separate 2D contact decoders, Ω^H and Ω^O, for humans and objects respectively. Θ, Ω^H, and Ω^O have the same architecture as SAM.

Given an RGB image, I, and prompt text, T_{inp}, the VLM produces contact tokens, <HCON> and <OCON>, for humans and objects, respectively. To aid MV-Loc in localizing contact, we extract the last-layer embeddings of the VLM corresponding to these tokens and pass them through a projection layer, Γ. The latter, Γ, is a multi-layer perceptron with 2 layers each of size 256 and a ReLU activation.

9.2. Training

Before the start of training, we render multiple views of the human mesh and object point cloud. We also compute the ground-truth contact mask.

9.2.1. Human Mesh Rendering

The human mesh rendering pipeline uses a comprehensive multi-view approach using the SMPL+H [54] parametric body model. We initialize the model in a neutral shape, positioning the body in a Vitruvian pose. This specific pose ensures optimal visibility of potential contact surfaces. We use PyTorch3D for rendering. We select 4 camera viewpoints to capture the complete body geometry: top-front (elevation 45°, azimuth 315°), top-back (45°, 135°), bottom-front (315°, 315°), and bottom-back (315°, 135°). Each viewpoint is positioned at a distance of 2 units from the subject with slight horizontal translations to optimize coverage. We use a FoV-Perspective projection model rendered

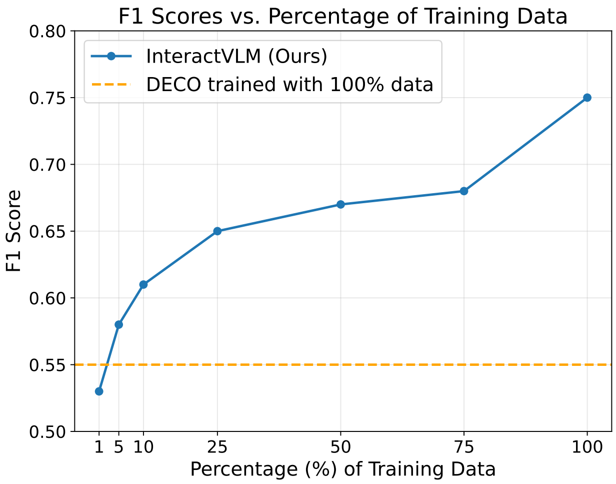

Figure 6. **Analysis of reliance on 3D annotations.** Performance evaluation for "binary human contact" (F1 score) for models trained on a varying percentage of DAMON [59] training data. The DECO baseline trains on 100% of DAMON. Instead, InteractVLM trains on a varying (smaller) portion of this dataset. Yet, it achieves a significantly higher performance, by leveraging the broad visual knowledge of foundation models.

at 1024×1024 resolution, with "blur-radius" and "faces-per-pixel" settings set as 0.0 and 1, respectively. For realistic appearance, we use point lights positioned at [0, 0, ±3] coordinates relative to the mesh. The lighting settings such as "ambient", "diffuse", and "specular" are set at 0.5, 0.3, 0.2, respectively. This creates a balanced illumination that highlights surface details. Surface normals are computed per vertex and are used as vertex colors.

Crucially, InteractVLM maintains precise correspondence between 2D rendered pixels and 3D mesh vertices. For each rendered view, it generates: (1) A pixel-to-vertex mapping matrix storing the indices of mesh vertices visible at each pixel. (2) Barycentric coordinates for accurate interpolation within mesh faces. (3) Binary contact masks for regions with at least three neighboring vertices in contact.

This comprehensive multi-view representation, combined with precise pixel-to-vertex correspondences, enables accurate lifting of 2D contact predictions back to the 3D mesh space. Our model processes each view as separate channels in a $B \times V \times 3 \times H \times W$ tensor shape during training, where B is the batch size and V is the number of views.

9.2.2. Object Point Cloud Rendering

The object rendering pipeline uses point clouds to capture object affordances in multiple views. The point cloud preprocessing begins with normalization, where each object is centered at its geometric centroid and scaled to fit within a unit sphere, ensuring consistent scale across different objects. Since the point clouds do not have color, we use the NOCS representation for coloring, namely for every point we assign a color derived from its normalized spatial NOCS coordinates (scaled to [0.1, 0.9] for better contrast).

Our rendering pipeline uses PyTorch3D with four viewpoints: front-left (elevation 45°, azimuth 315°), front-right (45°, 45°), back-left (330°, 135°), and back-right (330°, 225°). Each view is rendered at 1024×1024 resolution using a FoVPerspective camera positioned at a distance of 2 units from the object center. We use a fixed point cloud radius of 0.05. For the rasterization settings: we use 10 points per pixel and 50,000 points per bin to handle dense point clouds effectively. An alpha compositor is used for the final rendering. For affordance heatmaps, we generate a rendered view with continuous values, [0, 1], representing the affordance likelihood. For each view, we create a pixel-to-point mapping for lifting 2D affordance heatmaps to 3D affordance points.

9.3. Additional Text Data for Training

9.3.1. Data from GPT4o

To enhance our model's understanding of human-object interactions (HOI), we build a comprehensive Visual Question-Answering (VQA) data generation pipeline using GPT-4V (GPT4o). The pipeline processes images from three datasets, namely DAMON [59], LEMON [67], and PIAD [66], generating structured textual descriptions that capture multiple aspects of HOI.

For each image, we query GPT-4V to describe five key aspects: (1) the human's visual appearance including clothing and distinctive features, (2) specific body parts in contact with the object, (3) the nature of the interaction, (4) the object's physical characteristics, and (5) the specific parts of the object in contact with the human. To ensure efficient processing while maintaining visual fidelity, images are resized to 256×256 pixels.

These generated VQA data enrich the training signal with detailed descriptions of interactions. This additional supervision helps our model develop a more nuanced understanding of the relationship between visual features and contact regions, ultimately contributing to improved performance in contact prediction tasks. We format the collected data as JSON files to seamlessly integrate these with our VLM training pipeline, allowing the model to leverage these rich textual descriptions during the learning process.

9.3.2. Converting 3D contact vertices to text

To establish a precise mapping between 3D contact vertices and natural-language descriptions, we leverage the SMPL body model's semantic segmentation. The body is divided into 15 semantically meaningful parts including the torso, head, hands, feet, arms, legs, thigh and forearm. For training our VLM, we employ a diverse set of natural-language prompts that query about body part contacts with objects.

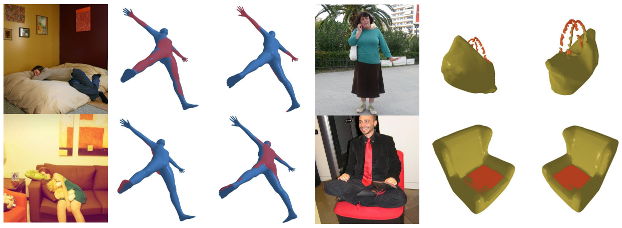

Figure 7. **Failure Cases.** Our method struggles with unusual human poses (left). For objects (right), training on affordances rather than actual contacts can sometimes lead to ambiguous contact predictions, especially for large objects like chairs. However, no dataset exists for 3D object contacts for in-the-wild images.

This structured approach creates a strong bridge between geometric contact information and natural-language understanding, enabling the model to learn the relationship between visual features, contact regions, and their semantic descriptions.

10. Failure Cases

Despite the overall strong performance, our method has certain limitations. For human contact prediction, our method occasionally struggles with unusual or ambiguous poses that deviate significantly from common interaction patterns. For example, in Fig. 7 the person is sleeping in an unusual pose on the bed.

Regarding objects, our method faces challenges inherent to the training paradigm. Since there exists no dataset of in-the-wild images with ground-truth 3D contact annotations for objects, we train on affordance data, which represents likelihood of contact rather than actual contact points. However, the distinction between actual contacts and affordances can be ambiguous, particularly for large objects like chairs, as shown in Fig. 7.

11. Qualitative Results

We present qualitative results for our InteractVLM method for three different tasks. First, in Fig. 8 we show the object affordance prediction results, where our method more accurately identifies plausible contact regions on objects compared to the state-of-the-art IAGNet method. Second, we show "semantic human contact" prediction results in Fig. 9, where our method successfully identifies contact regions on human bodies specific to different object categories, even in complex scenarios. Finally, in Fig. 10, we demonstrate 3D HOI reconstruction from in-the-wild images, where we leverage the *inferred* contacts on *both* human bodies and objects to generate physically plausible 3D reconstructions; this is done for the first time for in-the-wild images.

Object	#	Semantic-DECO [59]				**InteractVLM (Ours)**			
Categories		F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑	Geo. (cm) ↓	F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑	Geo. (cm) ↓
Skateboard	85	30.3	19.3	**91.3**	99.95	**71.5**	**67.0**	83.5	**0.90**
Surfboard	70	23.1	14.2	**98.4**	101.22	**79.7**	**76.3**	78.9	**0.80**
Snowboard	49	38.2	25.7	**92.2**	108.29	**84.2**	**83.1**	84.0	**0.20**
T. Racket	45	57.0	42.0	**99.6**	64.25	**82.3**	**80.8**	86.3	**0.20**
Cell phone	43	42.4	27.8	**99.6**	51.73	**70.6**	**73.1**	74.3	**7.00**
Couch	38	31.4	19.7	**89.2**	17.07	**62.1**	**62.5**	60.5	**2.10**
Bicycle	37	62.1	48.0	**98.1**	29.89	**81.5**	**84.4**	81.9	**2.50**
Chair	36	23.2	14.6	**87.1**	36.05	**70.3**	**73.6**	68.8	**1.60**
Bench	35	19.0	11.2	**92.1**	29.51	**63.0**	**70.7**	64.4	**4.00**
Motorcycle	33	60.4	45.5	**99.1**	19.24	**76.6**	**78.6**	77.7	**0.90**
Book	27	48.0	33.8	**99.7**	53.59	**74.1**	**75.2**	80.1	**1.10**
Skis	25	36.5	25.0	**93.4**	104.07	**83.0**	**81.4**	83.7	**0.80**
Bed	24	29.1	19.1	**82.9**	20.71	**54.0**	**56.7**	48.8	**2.70**
Laptop	24	36.9	24.9	**94.4**	45.73	**54.0**	**54.0**	68.6	**4.70**
Backpack	24	37.2	24.3	**87.2**	12.10	**59.2**	**71.1**	54.8	**3.50**
Umbrella	23	51.5	36.1	**99.2**	67.20	**82.3**	**83.7**	86.4	**1.00**
Knife	19	63.3	54.0	84.4	31.55	**77.0**	**74.9**	**86.6**	**0.10**
Frisbee	15	33.9	22.0	**99.4**	69.43	**68.7**	**71.5**	84.5	**1.00**
D. Table	11	19.6	14.1	**67.1**	42.56	**35.2**	**44.9**	63.4	**6.60**
B. Glove	10	71.4	63.3	81.9	41.58	**93.6**	**98.6**	**89.1**	**0.10**
Remote	10	0.2	1.0	0.1	82.16	**70.6**	**77.4**	**82.7**	**0.50**
Banana	10	6.1	7.1	6.4	67.19	**76.6**	**74.3**	**81.7**	**2.80**
Kite	9	65.3	51.8	**95.9**	50.50	**85.4**	**86.0**	85.4	**0.30**
Toothbrush	8	2.9	4.7	2.1	56.38	**77.3**	**82.6**	**74.8**	**5.40**
Boat	8	33.5	23.9	**83.7**	46.24	**71.3**	**75.3**	63.1	**1.40**
Sports ball	8	36.0	34.1	39.4	60.54	**64.4**	**74.0**	**83.8**	**5.30**
B. Bat	8	36.7	60.8	27.2	26.00	**82.8**	**81.2**	**87.8**	**1.60**
Apple	7	6.3	17.4	3.9	45.69	**69.3**	**62.9**	**77.7**	**4.20**
Handbag	7	12.1	7.0	**46.2**	26.61	**31.8**	**27.1**	40.4	**4.10**
Tie	6	39.8	28.1	**87.2**	**7.24**	**49.6**	**32.8**	60.8	7.60
Suitcase	6	26.7	24.0	30.7	87.44	**79.2**	**65.9**	**83.4**	**0.80**
Wine glass	5	5.5	8.4	5.0	70.32	**66.4**	**68.5**	**69.4**	**4.40**
Spoon	5	61.1	48.5	**89.9**	15.35	**67.5**	**62.8**	78.5	**5.50**
Fork	5	1.5	1.6	1.3	75.47	**64.9**	**66.2**	**76.5**	**2.20**
Keyboard	5	3.2	6.2	3.1	70.41	**60.8**	**69.1**	**74.0**	**0.50**
Teddy bear	5	17.5	15.7	45.0	24.70	**43.8**	**61.6**	**68.8**	**11.60**
Clock	4	23.3	14.8	58.1	46.42	**37.1**	**68.9**	**75.0**	**3.30**
Cake	4	0.0	0.0	0.0	83.99	**52.4**	**41.9**	**82.2**	**10.60**
Scissors	4	0.2	0.2	0.2	87.88	**28.7**	**21.4**	**73.1**	**40.10**
Cup	4	7.2	11.2	5.4	69.03	**68.6**	**71.4**	**76.2**	**1.70**
Car	4	0.0	0.0	0.0	49.13	**66.7**	**67.7**	**73.3**	**5.30**
Pizza	4	19.4	19.0	35.1	46.43	**44.3**	**44.1**	**71.4**	**29.20**
Carrot	3	0.0	0.0	0.0	90.22	**59.7**	**62.4**	**77.6**	**0.20**
Truck	3	0.0	0.0	0.0	61.65	**81.2**	**84.9**	**77.5**	**3.10**
Bottle	3	0.0	0.0	0.0	91.14	**59.2**	**55.1**	**81.2**	**0.10**
Airplane	2	0.0	0.0	0.0	87.52	**76.4**	**69.3**	**85.2**	**3.60**
Toilet	2	0.0	0.0	0.0	86.55	**32.5**	**35.7**	**71.1**	**3.30**
Hot dog	2	7.0	23.0	4.1	46.32	**81.3**	**84.0**	**78.9**	**4.10**
Donut	2	19.6	30.7	14.8	42.47	**73.6**	**90.1**	**65.6**	**12.00**
Mouse	1	0.0	0.0	0.0	82.03	**40.7**	**27.0**	**82.9**	**0.10**
Vase	1	0.0	0.0	0.0	91.96	**68.5**	**59.3**	**81.0**	**0.20**
F. Hydrant	1	0.0	0.0	0.0	88.18	**85.5**	**82.7**	**88.5**	**0.00**

Table 6. Evaluation for "Semantic Human Contact" prediction on the DAMON [59] dataset for different object categories in the test set. The number of samples for each category is shown in the second column. "Semantic-DECO" is our extension of the existing DECO [59] model for this new task. Zero metrics indicate no correct predictions for the class.

12. Future Work

Our approach follows a two-stage process for 3D HOI: it first predicts human and object contacts, and then uses the inferred contacts in optimization for joint 3D reconstruction. In the future, we will explore learning to perform both 3D contact prediction and 3D reconstruction in an end-to-end fashion. This could lead to more coherent predictions by learning and exploiting direct relationships between contact points and physical constraints.

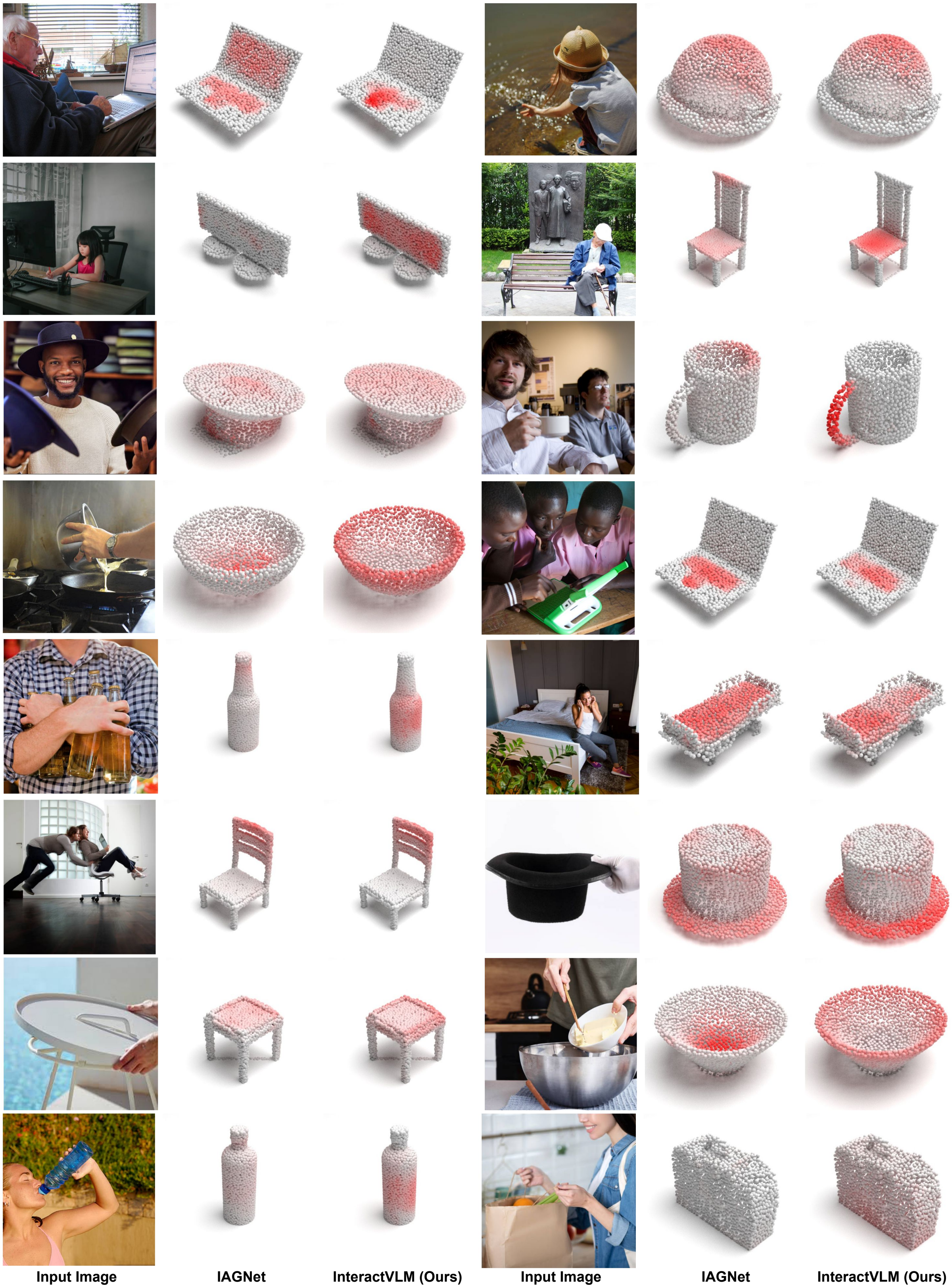

Figure 8. **Object Affordance Prediction.** Here we compare our InteractVLM method trained for object affordance prediction on PIAD [66] dataset with the state-of-the-art IAGNet method. We train for affordance detection because there exists no dataset of in-the-wild images paired with ground-truth 3D contacts for objects. Note that given an image of a person performing an action like "sit" or "grasp", the affordance prediction task estimates "contact possibilities" on the object.

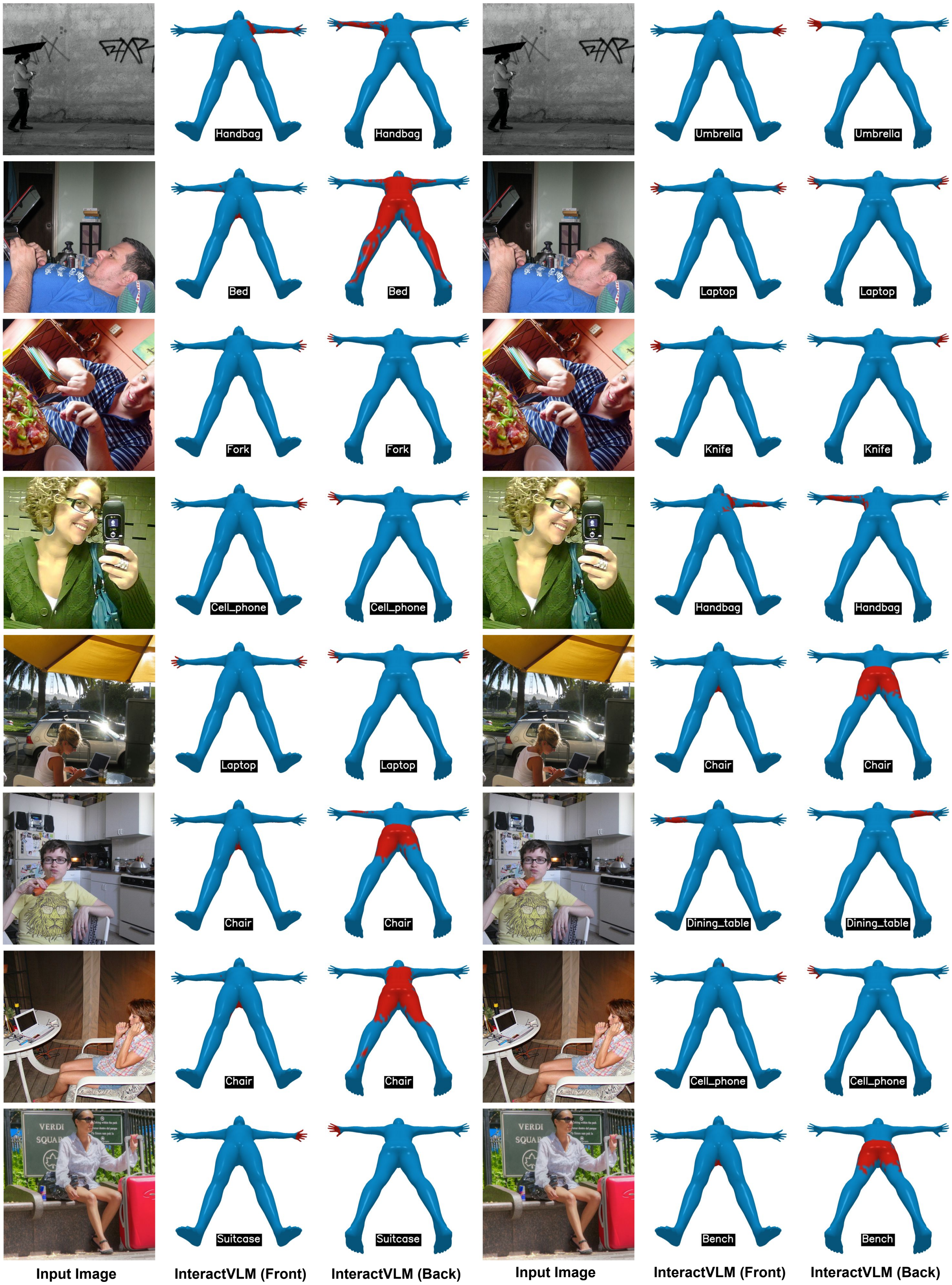

Figure 9. **Semantic Human Contact estimation.** Here we show results for "semantic human contact" estimation from in-the-wild images. Each row shows a person in contact with multiple objects. Note how InteractVLM estimates contact on bodies that is specific to the object.

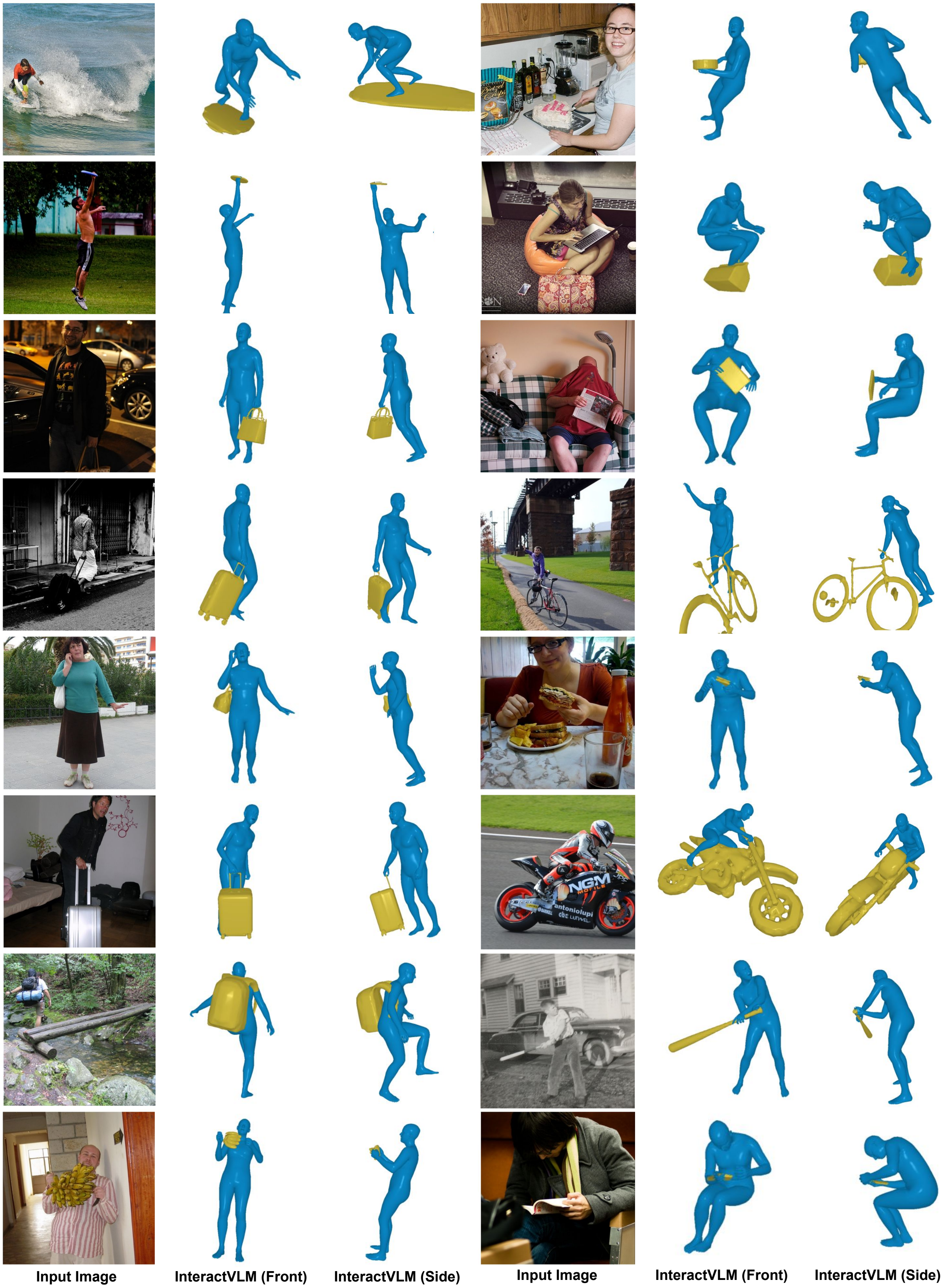

Figure 10. **3D HOI reconstruction.** Here we show results of our InteractVLM method for 3D HOI reconstruction from in-the-wild images. We use the InteractVLM's inferred contacts on both bodies and objects for joint 3D reconstruction.