InteractVLM: 3D Interaction Reasoning from 2D Foundational Models

Supplementary Material

S.1. Human Contact Prediction

S.1.1. Evaluation on the 3DIR Dataset

We evaluate our method against several state-of-the-art approaches for human contact prediction on the 3DIR dataset [7] as shown in Tab. S.1. 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.

S.1.2. Contact Estimation Across Body Parts

We extend our binary contact estimation's evaluation to measure our method's performance across different human body parts to ensure it captures nuanced interactions effectively. As shown in Tab. S.2, our method (InteractVLM) significantly outperforms DECO [6] across all body parts, including the head, torso, hands, arms, feet, and legs. The results demonstrate our method excels at detecting contacts across diverse body parts, making it well-suited for real-world scenarios.

S.1.3. Semantic Human Contact per Object Class

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. S.3. 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 [6] 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).

Method	3D Supervision		F1	Prec.	Rec.	Geo.
	Human	Obj.	(%) ↑	(%) ↑	(%) ↑	(cm) ↓
BSTRO [2]	✓	✗	55.0	57.0	58.0	28.58
DECO [6]	✓	✗	69.0	70.0	72.0	15.25
LEMON-P [8]	✓	✓	77.0	76.0	81.0	9.02
LEMON-D [8]	✓	✓	78.0	78.0	**82.0**	7.55
InteractVLM	✓	✗	**78.4**	**82.5**	76.3	**6.73**

Table S.1. Evaluation for "Binary Human Contact" prediction on the 3DIR dataset [8]. 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.

Method	Head	Torso	Hips	Hands	Arms	Feet	Legs
DECO [6]	20.0	46.1	66.6	74.3	22.2	94.4	66.6
Ours	**56.0**	**87.2**	**95.7**	**93.5**	**71.5**	**96.9**	**68.3**

Table S.2. F1 scores for human contact estimation w.r.t. body parts

S.2. Ablation Studies

We conduct extensive ablation studies to evaluate the contribution of InteractVLM's main components, including the influence of the VLM and prompt design choices. The results are summarized in Tab. S.4, where we use alphabet numbering to refer to each variant for clarity. Below, we discuss the key findings from these experiments.

Mask Resolution and MV-Loc Components. Increasing the mask resolution from 512 × 512 (variant (a)) to 1024 × 1024 (variant (b)) yields a significant improvement of 4.9% in F1 score, highlighting the importance of fine-grained spatial information for contact detection. For the MV-Loc feature embedding, using our FeatLift network (variant (e)) outperforms simply concatenating camera parameters (variant (d)) by 3.7%, demonstrating its effectiveness in incorporating viewpoint information. Removing camera parameters entirely (variant (c)) further degrades performance, emphasizing their role in the pipeline. However, the performance significantly drops when we replace MV-Loc with a 2 layer MLP (variant (f)).

Loss Functions. Using only the valid mask regions for training (variant (h)) improves performance by 3.3% compared to using the whole mask (variant (g)). The addition of our 3D contact loss (variant (i)) further boosts the F1 score by 3%, underscoring the importance of explicitly modeling 3D contact cues during training.

Data and VLM Influence. The choice of training data significantly impacts performance. Using only 3D contact datasets (variant (j)) results in a relatively low F1 score of 65.9%. Adding contact parts in text form (variant (k)) improves performance by 8.9%, while further incorporating HOI-VQA data (variant (l)) achieves the best results. This demonstrates the value of leveraging textual and contextual cues for contact localization.

The VLM plays a critical role in the pipeline. Removing the VLM entirely (variant (m)) drastically reduces performance, while using a VLM with only image input (variant (n)) serves as a strong baseline. Fine-tuning the VLM (variant (b)) is crucial, as the non-fine-tuned version (variant (o)) shows a significant drop in performance. Interestingly,

Object	#	Semantic-DECO [6]				**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 S.3. Evaluation for "Semantic Human Contact" prediction on the DAMON [6] 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 [6] model for this new task. Zero metrics indicate no correct predictions for the class.

	Variants	F1 (%) ↑	Prec. (%) ↑	Rec. (%) ↑
Masks	(a) Size 512×512	70.7	70.1	71.4
	(b) Size 1024×1024	75.6	75.2	76.0
MV-Loc	(c) No CamParams	69.4	68.0	71.1
	(d) Concat CamParams	71.9	72.0	71.8
	(e) FeatLift (Φ)	75.6	75.2	76.0
	(f) No MV-Loc	62.3	60.8	63.9
Losses	(g) Whole Mask	69.3	68.7	70.0
	(h) Valid Mask	72.6	71.2	74.0
	(i) + 3D Contact Loss	75.6	75.2	76.0
Data	(j) 3D Contact Datasets	65.9	64.8	67.0
	(k) + Contact Parts (text)	74.8	74.5	75.1
	(l) + HOI-VQA	75.6	75.2	76.0
VLM	(m) No VLM	32.3	30.8	43.0
	(n) VLM-13B-Img	67.2	68.5	66.0
	(o) VLM-13B-NoFT	64.8	65.3	64.2
	(p) VLM-7B	73.3	76.8	73.5
Prompt	(q) Contact parts (fine)	74.8	74.5	75.1
	(r) Contact parts (coarse)	68.4	69.0	67.8
	(s) No object name	71.5	72.1	70.9

Table S.4. Ablation study for the effect of different InteractVLM components. We evaluate for "Binary Human Contact" prediction on the DAMON dataset [6].

reducing the VLM size from 13B to 7B parameters (variant (p)) has minimal impact, suggesting that the model can maintain strong performance even with fewer parameters.

Prompt Design. The design of the text prompt significantly influences the results. Using fine-grained contact parts (variant (q)) outperforms a coarse segmentation (variant (r)) by 6.4% in F1 score, indicating that finer granularity in body part labeling is beneficial. Removing the object name from the prompt (variant (s)) also degrades accuracy, highlighting the importance of explicit object context in guiding the VLM's predictions.

Our ablation studies demonstrate the importance of fine-grained spatial information, effective feature embedding, 3D contact modeling, and well-designed prompts in achieving robust contact localization. The VLM's role, particularly its fine-tuning and input modalities, is also critical to the overall performance.

S.3. Impact of RLL

"Render-Localize-Lift" (RLL) is central to our method. Traditional approaches, like DECO [6] and RICH [2], rely on fully supervised learning with limited 3D GT data to predict 3D contacts. While effective for scenarios encountered during training, these methods fail to generalize to in-the-wild cases. To address this limitation, we leverage the broad visual knowledge of VLM to learn from the limited data. However, effectively utilizing VLM requires reformulating our 3D problem into a 2D representation, making it compatible with VLM, which we achieve through RLL. As demonstrated in the main paper, by using RLL with VLM, we surpass the state-of-the-art method for human contact estimation while training on only 5% of the data.

S.4. Implementation Details

S.4.1. Architecture

InteractVLM has two major blocks; a reasoning module, Ψ, based on LLaVA-v1 [4] and a novel multi-view localization model, MV-Loc, based on SAM [3]. 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.

S.4.2. 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.

S.4.2.1. Human Mesh Rendering

The human mesh rendering pipeline uses a comprehensive multi-view approach using the SMPL+H [5] 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 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.

S.4.2.2. 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.

S.4.3. Additional Text Data for Training

S.4.3.1. 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 [6], LEMON [8], and PIAD [7], 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.

S.4.3.2. 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

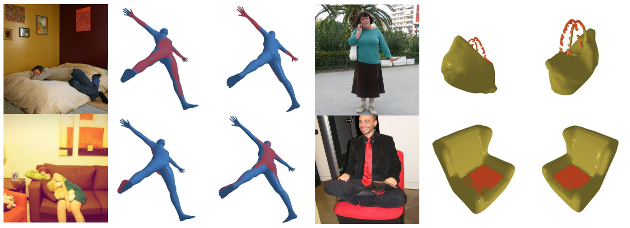

Figure S.1. **Contact Estimation Failure Cases.** Our method struggles with unusual human poses (left). For objects (right), training on affordances rather than actual contacts can sometimes lead to ambiguous contact predictions, especially for large objects like chairs. However, no dataset exists for 3D object contacts for in-the-wild images.

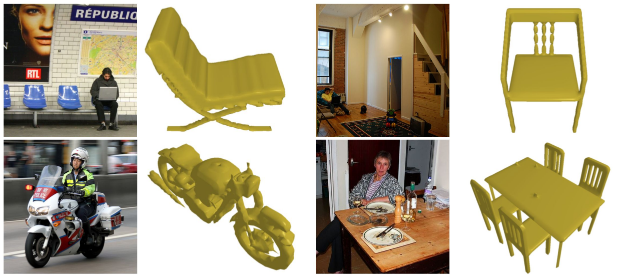

Figure S.2. **Object Retrieval Failure Cases.** The retrieved object meshes (right) differ notably from the actual objects in the input images (left), especially in cases of significant occlusion, atypical object instances, or limited database coverage. Despite these inaccuracies, the retrieval consistently selects objects within the correct semantic category.

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. 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.

S.5. 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. S.1 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. S.1.

In highly occluded or visually ambiguous scenarios, our approach can face challenges due to object lookup failure. The object lookup is also limited by the richness and diversity of underlying object database. However, since our method performs retrieval within predefined object categories, it consistently retrieves an object instance belonging to the correct semantic category, even if exact geometric matches are not always guaranteed. We provide qualitative examples highlighting these limitations in Fig. S.2.

S.6. Qualitative Results

We present qualitative results for our InteractVLM method for three different tasks. First, in Fig. S.3 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. S.4, where our method successfully identifies contact regions on human bodies specific to different object categories, even in complex scenarios. Finally, in Fig. S.5, 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.

S.7. 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.

Moreover, currently our approach learns on disjoint image datasets of body contacts and object affordances. In the future, we will also exploit recent datasets of images paired with contact annotations for both bodies and objects [1].

References

[1] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S. Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. S.4

[2] 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. S.1, S.2

[3] 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. S.2

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. S.2

[5] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6), 2022. S.3

[6] 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. S.1, S.2, S.3

[7] 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. S.1, S.3, S.6

[8] 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. S.1, S.3

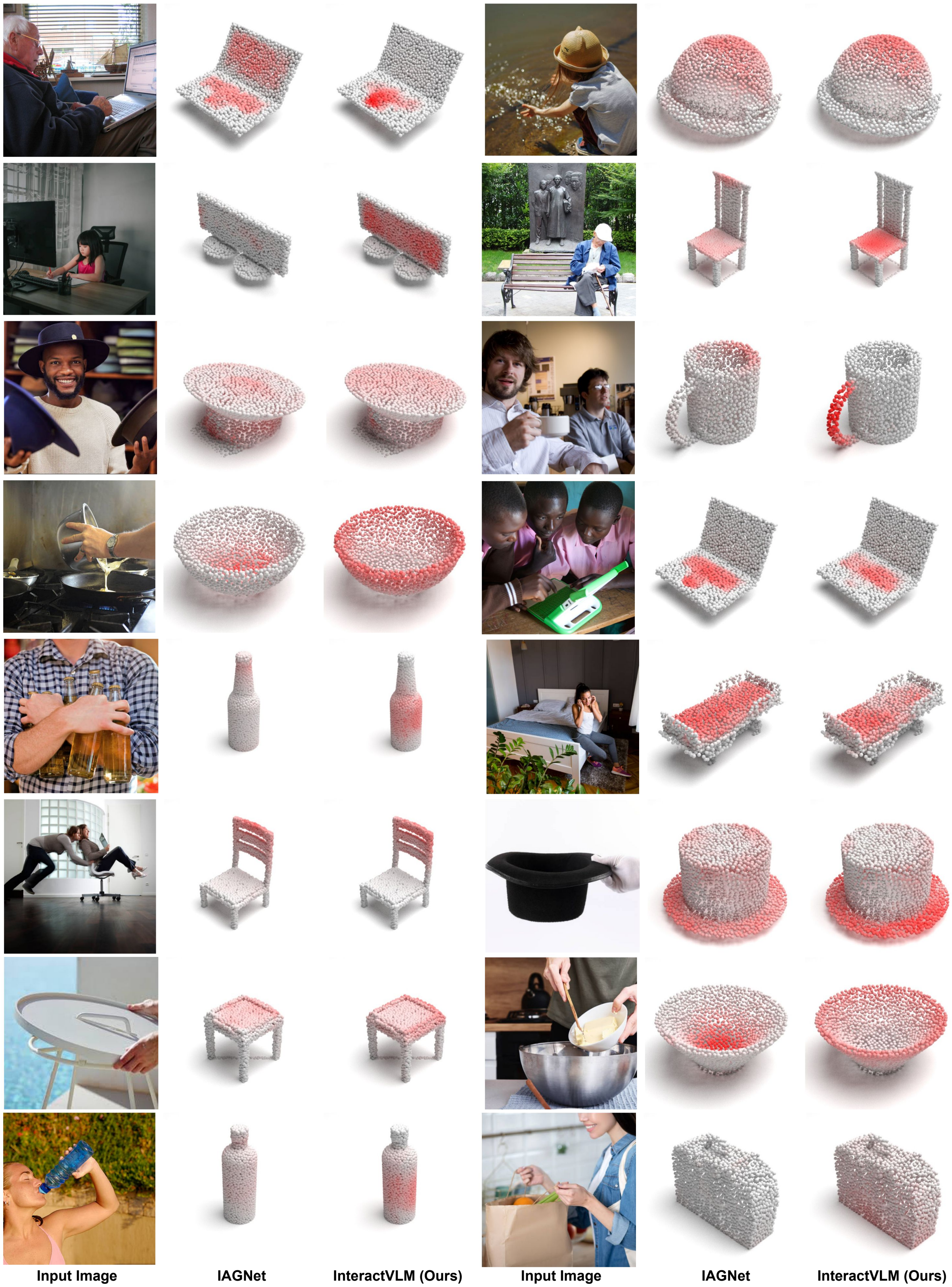

Figure S.3. **Object Affordance Prediction.** Here we compare our InteractVLM method trained for object affordance prediction on PIAD [7] dataset with the state-of-the-art IAGNet method. We train for affordance detection because there exists no dataset of in-the-wild images paired with ground-truth 3D 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.

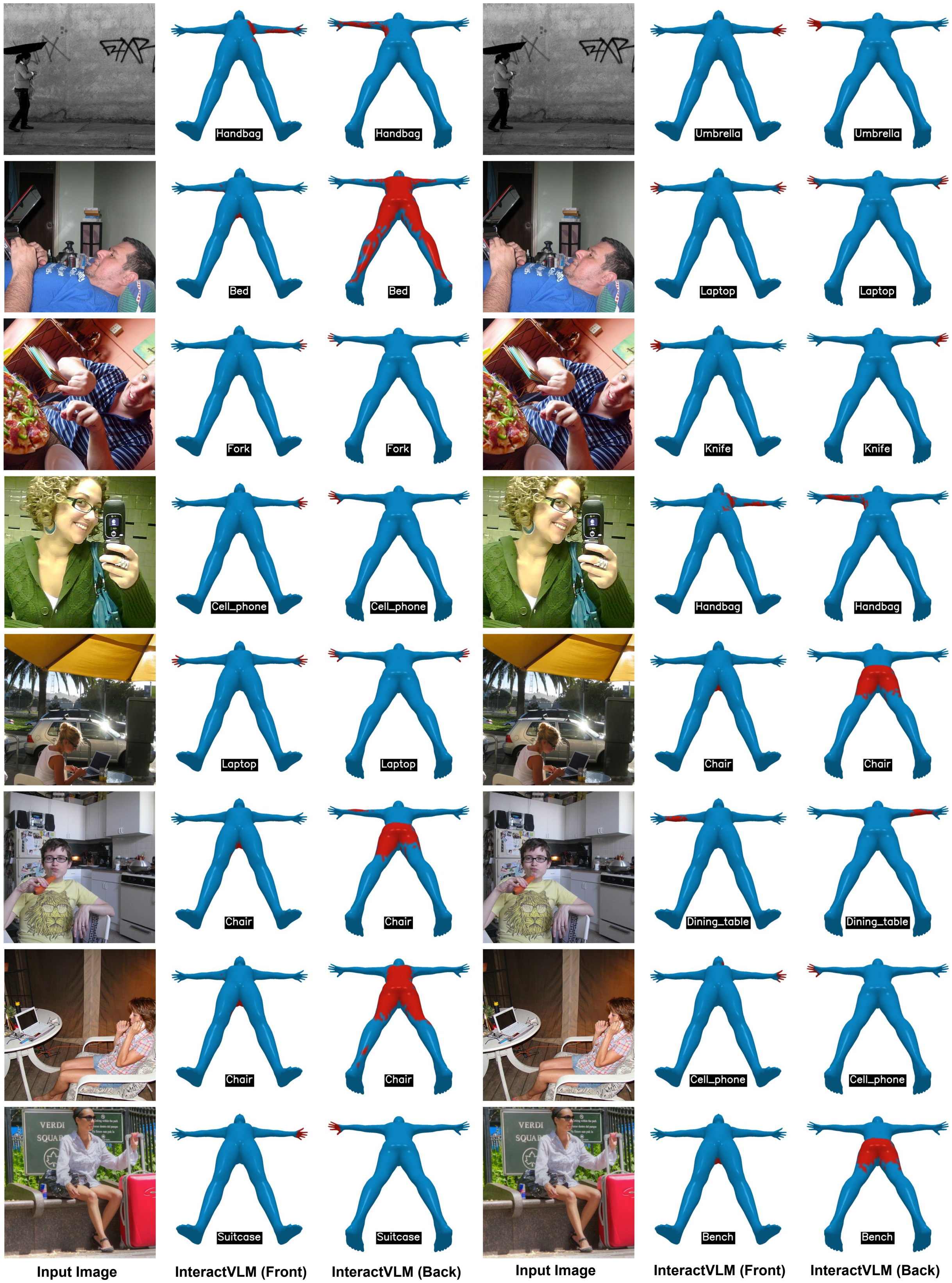

Figure S.4. **Semantic Human Contact estimation.** Here we show results for “semantic human contact” estimation from in-the-wild images. Each row shows a person in contact with multiple objects. Note how InteractVLM estimates contact on bodies that is specific to the object.

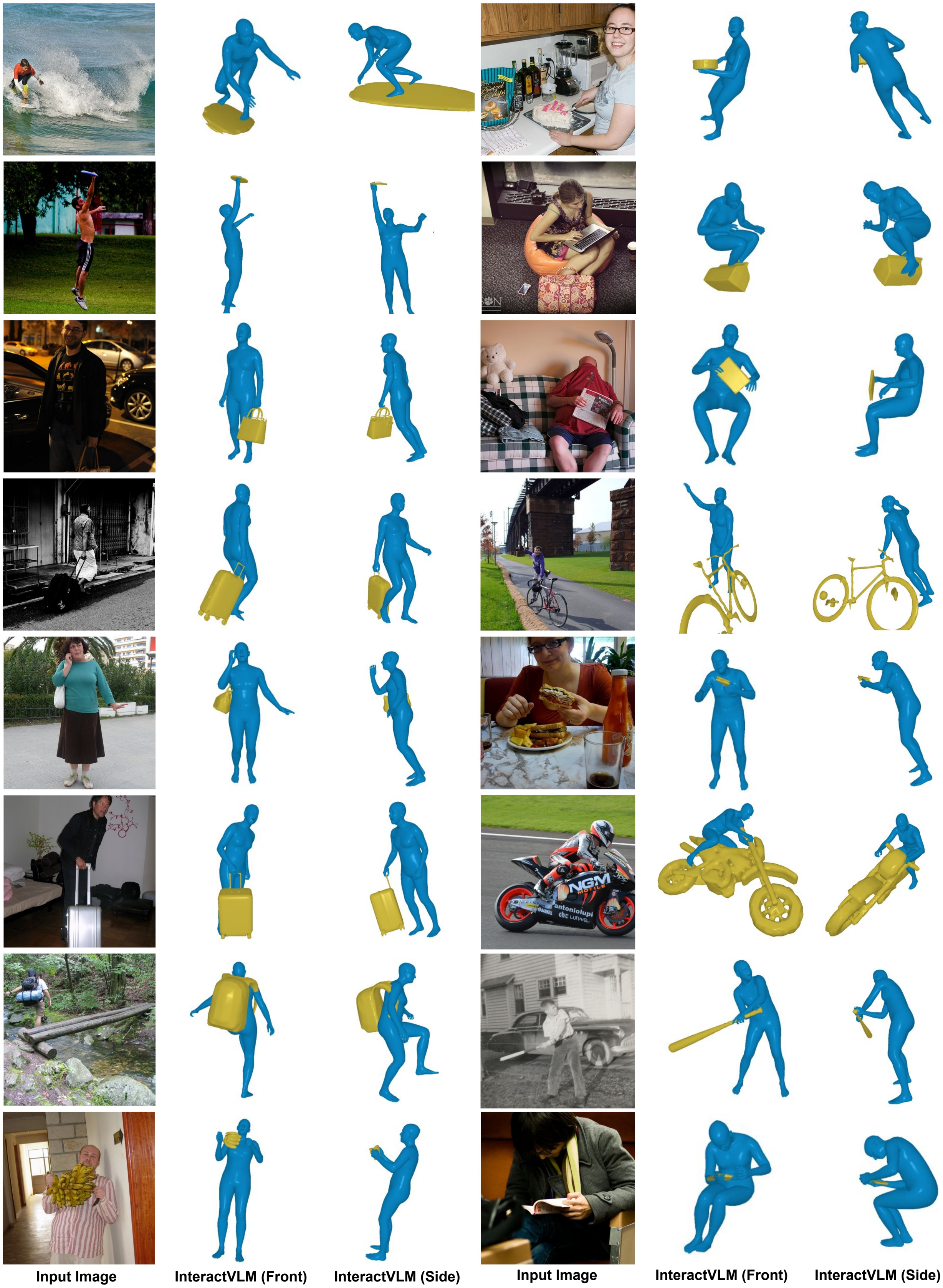

Figure S.5. **3D HOI reconstruction.** Here we show results of our InteractVLM method for 3D HOI reconstruction from in-the-wild images. We use the InteractVLM's inferred contacts on both bodies and objects for joint 3D reconstruction.