CARI4D: Category-Agnostic 4D Reconstruction of Human-Object Interaction

Anonymous CVPR submission

Paper ID ***

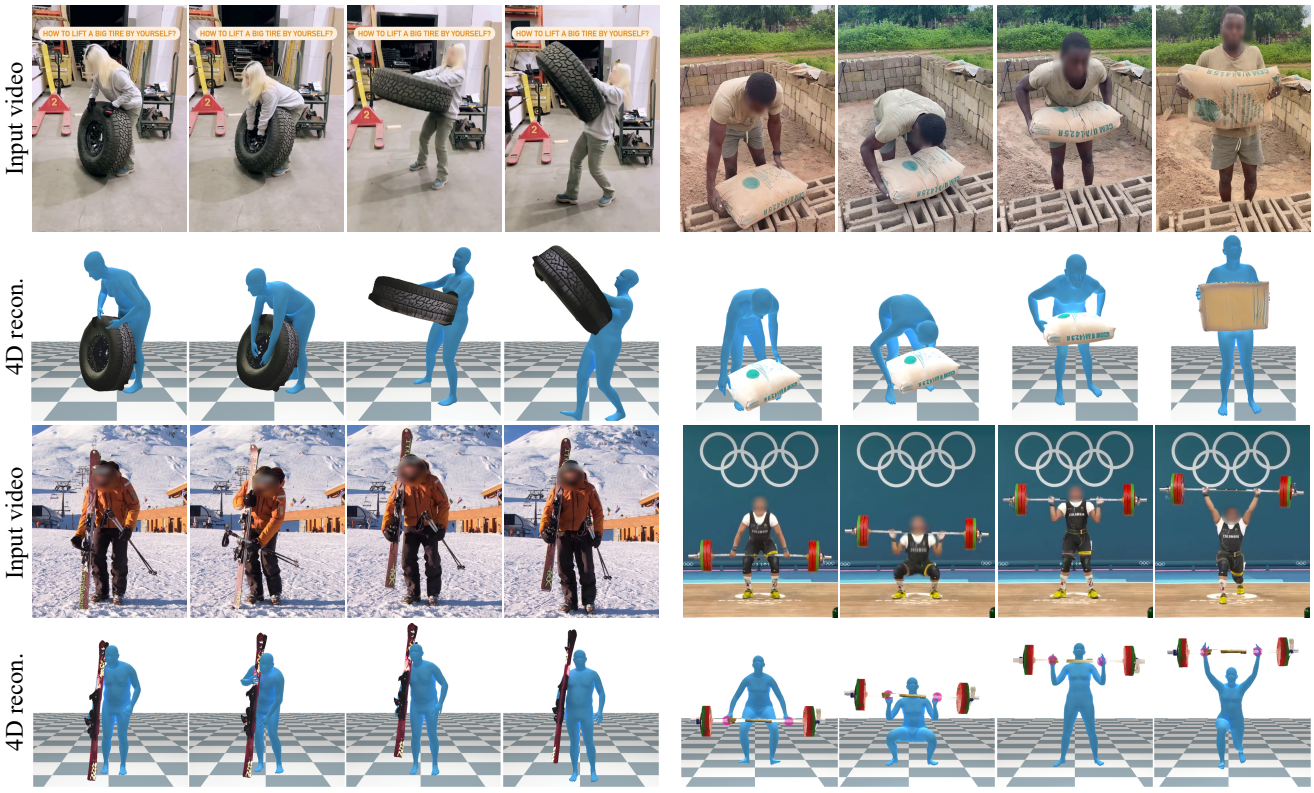

Figure 1. **Category-agnostic 4D interaction reconstruction.** From a monocular RGB video, CARI4D reconstructs a metric-scale object mesh, tracks human and object motion over time, and predicts coherent hand-object contacts for categories unseen during training.

Abstract

Reconstructing 4D human-object interaction from monocular RGB video would make large-scale interaction capture possible outside controlled studios, but the task is ill posed: object geometry is unknown, depth and scale are ambiguous, and contacts are often heavily occluded. We present CARI4D, a category-agnostic framework that reconstructs a metric-scale object, estimates temporally consistent human and object motion, and reasons about hand-object contact from a single RGB video. CARI4D combines foundation models for object generation, metric depth, human pose, and object pose, but explicitly aligns their outputs through coarse-to-fine scale selection, pose-hypothesis filtering, learned render-and-compare contact reasoning, and contact-aware joint optimization. On BEHAVE and zero-shot InterCap evaluations, CARI4D improves combined reconstruction Chamfer distance by more than 35% over prior video-based baselines, while also generalizing to in-the-wild videos with previously unseen object categories.

1. Introduction

Human-object interaction capture is central to human understanding, AR/VR, gaming, and robot learning. Existing

high-fidelity pipelines require calibrated multi-view studios or RGBD rigs [2, 3, 7], which limits scale. Monocular RGB videos are far easier to obtain, but metric 4D reconstruction from them is difficult because object geometry is unknown, depth is ambiguous, and contact-rich motion creates severe occlusion.

Prior RGB-based video methods assume known object templates or fixed training categories [17–19]. Foundation models now provide strong object generation, pose estimation, human reconstruction, and metric-depth priors [10, 11, 13, 14], but directly composing them is insufficient: their outputs lie in different coordinate systems and do not satisfy interaction contacts.

We introduce CARI4D, a category-agnostic monocular RGB framework for full-body 4D human-object interaction reconstruction. CARI4D reconstructs the object and metric scale, initializes human and object poses in a shared space, refines them with the learned contact reasoning network CoCoNet, and finally performs contact-aware joint optimization. Our contributions are a template-free metric 4D pipeline, robust foundation-model integration, and contact reasoning that improves zero-shot reconstruction and contact coherence.

2. Related Works

Foundation models for 3D reasoning. Recent models reconstruct object shape [13, 15, 20], estimate object or human pose [11, 14], and predict metric scene depth [9, 10]. They provide strong priors but reason about humans, objects, and scenes separately.

Interaction reconstruction. Image-based interaction methods estimate hand-object or full-body contacts from a single frame [4, 5, 16, 21], while video trackers improve temporal consistency but typically require object templates or category-specific training [17, 19]. CARI4D instead reconstructs the object, tracks human and object motion, and reasons about contacts from monocular RGB video without assuming a known category.

3. Method

Given RGB frames $\{\mathbf{I}_i\}_{i=1}^N$, our goal is to recover an object mesh $\mathbf{O}$ with per-frame 6DoF poses $\{\mathcal{O}_i = (\mathbf{R}_i^o, \mathbf{t}_i^o)\}_{i=1}^N$ and SMPL-H human parameters $\{\mathcal{H}_i = (\boldsymbol{\theta}_i, \boldsymbol{\beta}, \mathbf{t}_i^h)\}_{i=1}^N$ in a single metric coordinate system. Figure 2 summarizes the pipeline.

3.1. Metric-scale Object Reconstruction

We assume the object is mostly visible in the first frame, a common setup for interaction videos. We segment it with f-BRS [12] and reconstruct a normalized object mesh using Hunyuan3D-2 [13]. To place the mesh in metric scale, we estimate depth and intrinsics with UniDepth [10] and run FoundationPose [14] over candidate object scales. For each scale, the posed mesh is compared with the segmented depth point cloud using a one-way Chamfer distance. We perform a coarse search, refine around the top candidates, and keep the scale with the lowest distance.

3.2. Human and Object Pose Initialization

With the metric object mesh fixed, FoundationPose can provide object pose candidates per frame, but its top-ranked prediction is unreliable under monocular depth noise, imperfect generated geometry, and hand/body occlusion. We therefore select among its K candidates using two filters. First, each candidate is rendered and scored by mask IoU against the observed object mask after removing human occlusion. Second, candidates whose rotation changes too abruptly from the previous accepted frame are rejected. If no candidate survives during an occluded interval, we jump forward to a reliable frame and track backward, applying the same filters. For the human, we run NLF [11] and align its global translation and scale to UniDepth point clouds so the human and object share metric depth.

3.3. Contact Reasoning and Refinement

The initialization treats human and object separately, so contacts can float, penetrate, or drift over time. We train CoCoNet, a category-agnostic contact reasoning model, to refine poses and predict binary left/right hand contact labels. For a short frame window, we render the initialized human and object into RGB, depth, and mask channels, then compare these renderings against the observed RGB, depth, and segmentation cues. A DINOv2-based encoder [8] extracts visual and geometric features, spatiotemporal attention aggregates them across frames, and lightweight heads predict pose updates and contacts. During training, estimated depth is aligned to ground-truth depth before generating initializations, which prevents the network from learning dataset-specific depth biases instead of interaction corrections.

3.4. Contact-based Joint Optimization

The final stage optimizes the full sequence by minimizing a weighted objective

$$L = \lambda_c L_c + \lambda_j L_{\text{j2d}} + \lambda_m L_m + \lambda_p L_{\text{pen}} + \lambda_a L_{\text{acc}}. \quad (1)$$

L_c attracts contacted hand joints to the object when CoCoNet predicts contact, L_{j2d} and L_m preserve image evidence, L_{pen} discourages human-object penetration, and L_{acc} smooths motion.

4. Experiments

Experimental setup and metrics. We train on BEHAVE [1] and HODome [22], then evaluate on BEHAVE and the unseen InterCap dataset [6]. Following prior work [4, 17], we

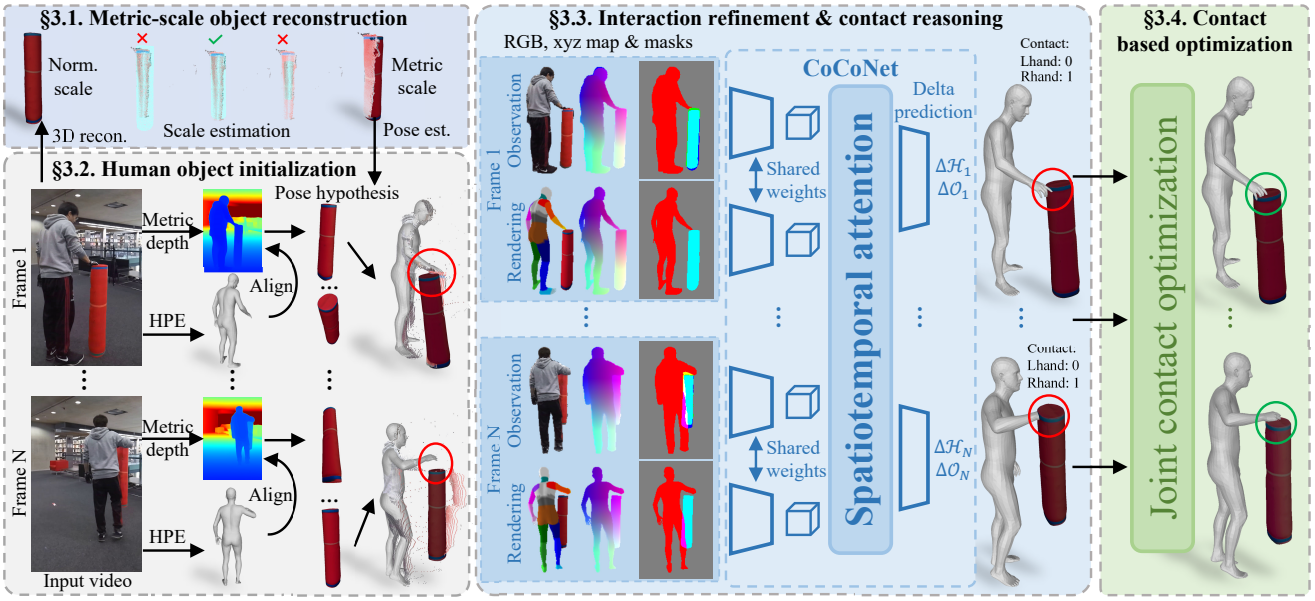

Figure 2. **CARI4D overview.** We reconstruct a metric object mesh, initialize human and object poses in a shared coordinate system, refine them with CoCoNet using render-and-compare contact reasoning, and perform contact-aware joint optimization.

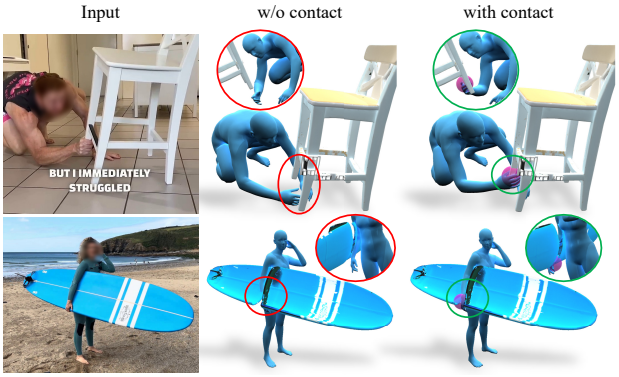

Figure 3. **Contact-aware refinement.** Predicted contacts guide optimization away from floating objects and penetration artifacts.

report Chamfer distance in centimeters for the human (CD-h), object (CD-o), and combined mesh (CD-c), along with human and object acceleration errors. To preserve the metric 4D setting, we align only the first frame to ground truth and apply that single transformation to the full video.

4.1. Baseline Comparison

Table 1 compares against InterTrack and VisTracker. InterTrack reconstructs category-specific point clouds but produces inconsistent global translation, yielding high CD-c. VisTracker has temporally smoother motion but requires known object templates and struggles when adapted with reconstructed meshes. CARI4D reconstructs the object directly from video, recovers coherent motion, and reduces

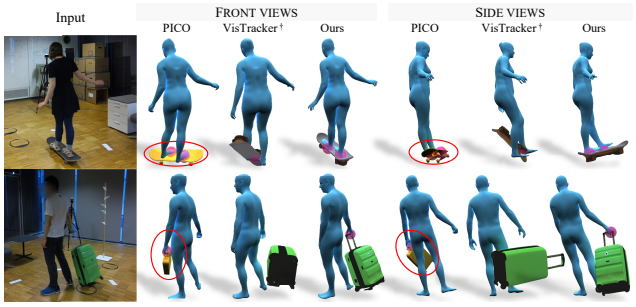

Figure 4. **Zero-shot InterCap examples.** CARI4D reconstructs unseen object geometry and temporally consistent poses without category-specific training or a provided object template.

CD-c from 14.22 to 9.23 on BEHAVE. On unseen InterCap objects, CARI4D improves CD-c from 20.17 to 12.88 over the strongest video baseline.

4.2. Zero-shot Generalization

Compared with image-based PICO [4] on InterCap key frames, CARI4D reduces CD-c from 87.73 to 5.90 when both are evaluated on the same sampled frames. Qualitatively, CARI4D also handles in-the-wild internet videos with object categories not present in training, as shown in Figure 1.

4.3. Ablation Studies

Table 2 shows that vanilla NLF plus FoundationPose is not sufficient in our RGB-only setting. Our pose-selection and alignment stage yields a much stronger initialization,

Method	BEHAVE					InterCap zero-shot				
	CD-h	CD-o	CD-c	Acc-h	Acc-o	CD-h	CD-o	CD-c	Acc-h	Acc-o
InterTrack [19]	25.71	47.66	30.20	5.23	5.64	34.79	40.37	33.53	4.34	5.26
VisTracker† [17]	13.52	18.29	14.22	**0.54**	0.77	16.12	27.41	20.17	**0.98**	1.25
CARI4D	**7.74**	**12.05**	**9.23**	1.14	**0.35**	**11.06**	**15.69**	**12.88**	1.25	**0.82**

Table 1. **Video reconstruction results** (cm, lower is better). †VisTracker requires an object template, so we provide our reconstructed object mesh for this comparison. CARI4D improves combined Chamfer distance on both in-distribution BEHAVE and unseen InterCap.

Figure 5. **In-the-wild qualitative comparison.** Prior approaches can produce noisy object shapes, flipped poses, or incorrect contacts on internet videos. CARI4D maintains more coherent object geometry, pose, and contact predictions across diverse categories.

Variant	CD-h	CD-o	CD-c	Acc-h	Acc-o
Raw NLF + FP tracking	11.53	1565.42	405.13	3.06	4.34
Raw NLF + FP pose est.	11.53	40.54	13.02	3.06	9.26
Our initialization	7.81	16.85	10.79	2.21	8.36
+ CoCoNet	**7.01**	11.59	**8.62**	1.75	3.78
Full model	8.41	**11.57**	9.35	**1.06**	**0.38**

Table 2. **Ablations on BEHAVE subset.** Pose selection and alignment improve initialization; CoCoNet improves reconstruction; joint optimization improves motion smoothness and contact coherence.

CoCoNet further improves Chamfer distance, and the full contact-aware optimization produces the smoothest object motion. Figure 3 illustrates how contact reasoning removes common floating and penetration artifacts.

5. Conclusion

We presented CARI4D, a category-agnostic framework for metric 4D reconstruction of human-object interaction from monocular RGB video. By aligning foundation-model predictions, reasoning about contacts with CoCoNet, and enforcing contact-aware sequence optimization, CARI4D improves reconstruction accuracy on BEHAVE, generalizes zero-shot to InterCap, and applies to in-the-wild videos without a provided object template.

References

[1] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[2] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. *ACM Trans. Graph.*, 27(3):1–9, 2008. 2

[3] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), 2015. 2

[4] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *CVPR (CVPR)*, pages 1783–1794, 2025. 2, 3

[5] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[6] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 2

[7] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2

[9] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR (CVPR)*, 2024. 2

[10] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025. 2

[11] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. 2024. 2

[12] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *CVPR*, pages 8623–8632, 2020. 2

[13] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 2

[14] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024. 2

[15] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR (CVPR)*, 2025. Spotlight. 2

[16] 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)*. Springer, 2022. 2

[17] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4

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

[19] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. 2024. 2, 4

[20] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In *Arxiv*, 2024. 2

[21] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 2

[22] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 2