Articulation Perception from In-the-Wild Egocentric Hand–Object Interactions

Yihao Wang[1,6,*], Yang Miao[2,*], Wenshuai Zhao[1,6], Wenyan Yang[1]
Zihan Wang[1], Joni Pajarinen[1], Luc Van Gool[2,3], Danda Pani Paudel[2]
Juho Kannala[1,7], Xi Wang[3,4,5,†], and Arno Solin[1,6]

[1]Aalto University [2]INSAIT, Sofia University [3]ETH Zurich [4]TU Munich
[5]MCML [6]ELLIS Institute Finland [7]University of Oulu
{firstname.lastname}@aalto.fi, {firstname.lastname}@insait.ai,
vangool@vision.ee.ethz.ch, xi.wang@inf.ethz.ch
[*]Equal contribution. [†]Co-advisor.

Abstract. Articulation perception recovers the motion and mechanical structure of articulated objects (*e.g.* how a drawer opens). Existing methods rely on supervised training with high-quality 3D data and manual annotations, limiting scalability, whereas large-scale egocentric videos naturally capture realistic human interactions with articulated objects. We propose a training-free framework that extracts object- and scene-level articulations directly from these hand-object interactions, using natural 3D hand and geometric scene cues. Experiments on public data sets and downstream tasks further demonstrate its effectiveness.

1 Introduction

Interactions with articulated objects pervade daily life, from opening a fridge to closing a bedside drawer. For intelligent agents to operate in human environments, they must not only recognize objects [6, 13, 39, 56] but also perceive how these objects move, *i.e.*, their motion type, axis, and pivot. Such articulation perception is fundamental to 3D scene understanding and central to robotic manipulation [1, 22, 35], animation and simulation [53], and interactive content generation [4, 17, 25, 27, 34].

A large body of work models articulation from *static* input, predicting part segmentation and motion from point clouds [10, 15, 26, 31, 54] or single RGB(D) images [18, 46]. But they rely on costly annotations or high-fidelity 3D structures and, lacking motion cues, struggle under complex geometry, occlusion, or novel categories. Another line recovers articulation from *dynamic* input, using 3D structures [28], RGB [16, 20, 23, 42, 45], or RGB-D [3, 9, 14, 51, 52, 55]. Yet tracking-based methods [14, 32, 52, 55] degrade under occlusion, low texture, and motion blur, while approaches built on parametric hand or body models [16, 20, 21, 33] suffer from self-occlusion and often require multi-view or depth input. In contrast, large-scale egocentric video [11, 12, 40, 49] naturally records humans manipulating doors, drawers, cabinets, and appliances, whose 3D hand trajectories, recoverable with recent hand reconstruction methods [38, 41, 44, 57], implicitly reveal how objects move. Such video is monocular, noisy, and unstructured, however, rendering conventional multi-view or temporal methods inapplicable.

We address these limitations with a training-free framework that infers scene-level articulation directly from monocular egocentric video. Our key insight is that natural 3D hand trajectories carry strong cues about object motion and articulation structure. Combining reconstructed hand motion with geometric cues from the 3D scene, our method estimates articulation types and axes and localizes them in complex scenes, without manual annotations or category-specific training. We further introduce a foundation-model-driven component that leverages Vision-Language Model (VLM) priors to infer plausible articulation motion and structure. On HD-EPIC [40] and Arti4D [52], our method is on par with state-of-the-art baselines and surpasses them on most metrics, and the recovered articulations benefit downstream fine-tuning of articulation prediction models and real-world robotic manipulation.

Our main contributions are: *(1)* a fully automatic, training-free pipeline for scene-level articulation perception from in-the-wild monocular RGB egocentric videos; *(2)* a foundation-model-driven approach leveraging VLM priors to infer articulation motion and structure without task-specific training; *(3)* extensive experiments showing performance on par with, and surpassing on most metrics, strong state-of-the-art baselines; and *(4)* validation of the pipeline's generalizability across downstream 3D articulation prediction and real-world robotic manipulation.

2 Method

Problem Formulation. We target human-made objects and model each joint ϕ_i as a single-DoF articulation with motion type c_i, axis a_i, and origin o_i [8, 15, 36]. Given an in-the-wild untrimmed RGB video $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$ and action descriptions $\mathcal{L} = \{l_1, \ldots, l_M\}$ (each l_j specifying an action, *e.g.* open the door, and its temporal span), we estimate object-level $\phi_{obj} = \Phi(\mathbf{V}, l_j)$ and aggregated scene-level $\phi_{scene} = \Phi(\mathbf{V}, \mathcal{L})$ articulations.

Dynamic Interaction Perception Given the video $\mathbf{V}$ and an action description l_j, we reconstruct per-frame MANO [44] hand poses [41, 57] and, following grasp statistics [2, 47], take the thumb, index, and middle fingertip landmarks to form 3D hand trajectories $\{\mathbf{z}_t\}$; we estimate poses directly in metric space and apply Gaussian-Process smoothing to obtain refined articulation paths $\{\hat{\mathbf{p}}_t \in \mathbb{R}^3\}$. In parallel, from a coarse scene reconstruction [19] we build a voxel grid $\mathcal{V}$ and localize the interacted object as a high-confidence region $\mathcal{V}_{\text{conf}}$ by multi-view foreground voting,

$$\mathcal{V}_{\text{conf}} = \{\mathbf{v} \in \mathcal{V} \mid s(\mathbf{v}) \geq \tau_{\text{vox}}\}, \quad s(\mathbf{v}) = \textstyle\sum_f \mathbb{I}[\pi_f(\mathbf{v}) \in \mathcal{M}_f]\, \mathbb{I}[\mathbf{v} \text{ visible in } f], \tag{1}$$

where π_f projects a voxel into frame f and $\mathcal{M}_f$ is its foreground mask, yielding a spatial prior resilient to occlusions and hand-induced motion.

Scene Geometry Recovery We integrate camera view reselection with multi-view reconstruction, using motion-type priors to recover revolute and prismatic axis candidates a_i via different strategies [15]. From the region $\mathcal{V}_{\text{conf}}$, we select N global views whose camera frustums intersect the object via farthest-point

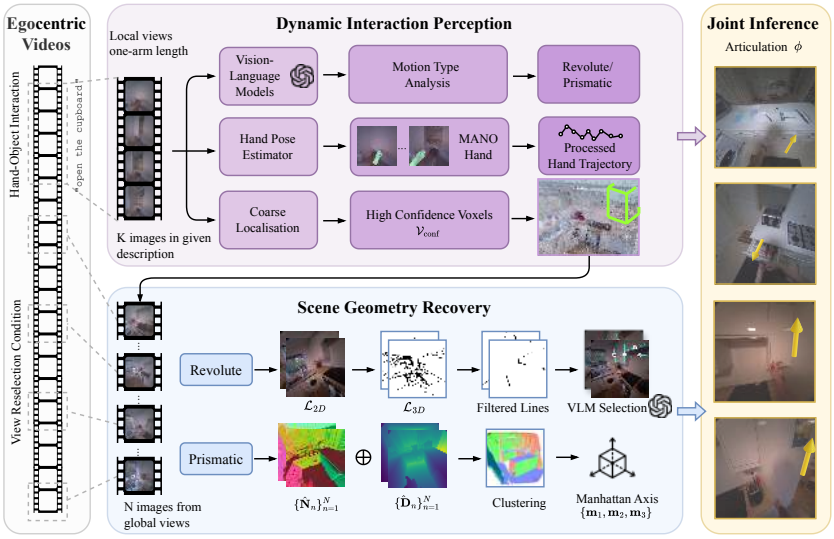

Fig. 1: Overall pipeline. Given a full in-the-wild egocentric video and a language description as input, our pipeline consists of Dynamic Interaction Perception, Scene Geometry Recovery, and Vision-Language Guided Motion Reasoning modules.

sampling over camera positions, and refine them with object semantic masks $\mathcal{M}$ [24, 30, 43] from the description l_j or a foundation model. We then detect 2D line segments [29, 37], unproject them to 3D via metric depth [50] and camera poses, aggregate them into dense 3D line tracks across views, and fit a direction-and-origin model with LO-RANSAC to obtain revolute axis candidates $\mathcal{L}_{3D}$; prismatic directions follow the Manhattan-world assumption [7].

Vision-Language Guided Motion Reasoning We leverage a VLM in a two-stage Visual Question Answering (VQA) pipeline: *(1)* **Motion Type Classification** classifies the joint motion type $c_i \in \{\text{revolute}, \text{prismatic}\}$. We uniformly sample frames from the clip $\mathbf{v}_j$, query the VLM per frame for the interacted object and its frame-level motion type, and aggregate these predictions to have a robust result. *(2)* **Axis Grounding VQA** helps to select line candidates $\mathcal{L}_{3D}$ that correspond to irrelevant static structures (*e.g.*, table corners or appliance edges) in form of a constrained multiple-choice VQA. We project the 3D candidates into a selected view and filter them by the semantic mask $\mathcal{M}$ using the per-view overlap ratio $\alpha_f(\ell) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}[\pi_f(\ell, s) \in \mathcal{M}_f]$, overlay the most prominent survivors as numbered Set-of-Marks prompts, and ask the VLM to pick the index best matching the physical hinge, yielding a verified subset $\mathcal{L}^*_{3D} \subset \mathcal{L}_{3D}$.

Joint Inference. Given the motion type, we fuse the dynamic hand cues with the static geometry and apply PCA and RANSAC to obtain the final parameters $\phi_i = \{c_i, a_i, o_i\}$, either from a single clip or aggregated across clips in a scene.

3 Experiments

We evaluate on the curated HD-EPIC [40] and Arti4D [52] data sets. Following [8, 15, 18, 46], we match each prediction to its ground-truth instance and

Table 1: Quantitative results.

Methods	Recall			Precision		
	M	MA	MAO	M	MA	MAO
Articulation3D* [42]	0.28	0.08	0.05	0.83	0.28	0.20
Articulation3D** [42]	0.44	0.11	0.07	0.72	0.22	0.16
ArtiPoint [52]	**0.55**	0.29	0.24	0.74	0.35	0.29
Ours	**0.55**	**0.43**	**0.32**	**0.98**	**0.77**	**0.56**
Ours-scene	0.72	0.58	0.53	0.99	0.80	0.72

Table 2: Ablation on components.

Ablation		Precision		Recall	
		MA	MAO	MA	MAO
Hand Cues	✗	0.45	0.18	0.35	0.15
	✓	0.78$_{+.32}$	0.56$_{+.38}$	0.42$_{+.07}$	0.32$_{+.17}$
VLM Axis Sel.	✗	0.68	0.52	0.44	0.34
	✓	0.75$_{+.07}$	0.53$_{+.01}$	0.47$_{+.03}$	0.33$_{-.01}$

report accuracy on three criteria: motion type (**M**), type+axis (**MA**), and type+axis+origin (**MAO**) averaged across scenes.

Experimental Results As shown in Tab. 1, our method is on par with prior baselines and surpasses them on most metrics. Both pre-trained Articulation3D variants [42] succeed on few clips due to the domain gap. ArtiPoint [52] works when tracking succeeds, but rapid actions, viewpoint changes, low-texture surfaces, and depth errors [50] leave its trajectories inaccurate, whereas our 3D hand estimation and global geometric constraints remain robust. Notably, our scene-level precision (Ours-scene) is nearly 100%, providing highly reliable articulation labels for downstream applications.

Ablation and downstream applications Table 2 ablates our two key components. Removing hand trajectories (random axis proposals instead) drops Prec.MAO by 0.38, confirming hand cues provide critical dynamic information. VLM-guided axis selection further improves accuracy over geometry-only inference by filtering implausible candidates. For downstream uses, finetuning USDNet [15] on our approach-extracted HD-EPIC labels significantly outperforms the zero-shot baseline both in-domain and cross-dataset (Arti4D), and, as a proof-of-concept, our extracted parameters drive a robot to open and close cupboards and drawers from human demonstrations.

4 Conclusion

We introduced a training-free framework that extracts object- and scene-level articulation from egocentric videos via hand cues, scene geometry, and VLM reasoning. It surpasses prior methods on most metrics across public data sets and transfers to downstream articulation prediction and robotic manipulation without high-fidelity 3D reconstruction or task-specific supervision.

Acknowledgements

We acknowledge funding by the Research Council of Finland (projects 339730, 362407, 362408, 373780, 373999), and CSC – IT Center for Science and the Aalto Science-IT project for computational resources. This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

References

1. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from Human Videos as a Versatile Representation for Robotics. In: CVPR. pp. 13778–13790 (2023)
2. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: ECCV. pp. 361–378 (August 2020)
3. Buchanan, R., Röfer, A., Moura, J., Valada, A., Vijayakumar, S.: Online estimation of articulated objects with factor graphs using vision and proprioceptive sensing. In: ICRA. pp. 16111–16117 (2024)
4. Chen, Z., Walsman, A., Memmel, M., Mo, K., Fang, A., Vemuri, K., Wu, A., Fox, D., Gupta, A.: URDformer: A Pipeline for Constructing Articulated Simulation Environments from Real-World Images. In: Robotics: Science and Systems (RSS) (2024)
5. Cheng, T., Shan, D., Hassen, A., Higgins, R., Fouhey, D.: Towards a Richer 2D Understanding of Hands at Scale. In: NeurIPS. vol. 36, pp. 30453–30465 (2023)
6. Corsetti, J., Giuliari, F., Fasoli, A., Boscaini, D., Poiesi, F.: Functionality understanding and segmentation in 3d scenes. In: CVPR (2025)
7. Coughlan, J., Yuille, A.L.: The Manhattan World Assumption: Regularities in Scene Statistics which Enable Bayesian Inference. In: NeurIPS. vol. 13 (2000)
8. Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., Engelmann, F.: SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In: CVPR. pp. 14531–14542 (2024)
9. Delitzas, A., Zhang, C., Gavryushin, A., Mario, T.D., Sun, B., Dabral, R., Guibas, L., Theobalt, C., Pollefeys, M., Engelmann, F., Barath, D.: Funrec: Reconstructing functional 3d scenes from egocentric interaction videos. In: CVPR (2026)
10. Deng, C., Lei, J., Shen, W.B., Daniilidis, K., Guibas, L.J.: Banana: Banach fixed-point network for pointcloud segmentation with inter-part equivariance. In: NeurIPS. vol. 36, pp. 34139–34152. Curran Associates, Inc. (2023)
11. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4D: Around the World in 3,000 Hours of Egocentric Video. In: CVPR. pp. 18995–19012 (2022)
12. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In: CVPR. pp. 19383–19400 (2024)
13. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R.: ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In: ICRA. pp. 5021–5028 (2024)
14. Gu, Q., Sheng, Y., Yu, J., Tang, J., Shan, X., Shen, Z., Yi, T., Liang, X., Chen, X., Wang, Y.: ArtiSG: Functional 3d scene graph construction via human-demonstrated articulated objects manipulation (2025)
15. Halacheva, A.M., Miao, Y., Zaech, J.N., Wang, X., Gool, L.V., Paudel, D.P.: Articulate3D: Holistic Understanding of 3D Scenes as Universal Scene Description. In: ICCV. pp. 5633–5644 (2025)
16. Haresh, S., Sun, X., Jiang, H., Chang, A.X., Savva, M.: Articulated 3d human-object interactions from rgb videos: An empirical analysis of approaches and challenges. In: 2022 International Conference on 3D Vision (3DV). pp. 312–321. IEEE (2022)

17. Huang, Z., Sun, B., Delitzas, A., Chen, J., Pollefeys, M.: REACT3D: Recovering articulations for interactive physical 3d scenes. IEEE Robotics and Automation Letters (RA-L) (2026)
18. Jiang, H., Mao, Y., Savva, M., Chang, A.X.: OPD: Single-View 3D Openable Part Detection. In: ECCV. pp. 410–426. Springer (2022)
19. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P.: MapAnything: Universal feed-forward metric 3D reconstruction. In: International Conference on 3D Vision (3DV). IEEE (2026)
20. Kerr, J., Kim, C.M., Wu, M., Yi, B., Wang, Q., Goldberg, K., Kanazawa, A.: Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. In: CoRL (2024)
21. Kim, J., Kim, J., Na, J., Joo, H.: ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling of Human-Object Interactions. In: CVPR. pp. 1816–1828 (2025)
22. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P.: OpenVLA: An Open-Source Vision-Language-Action Model. In: CoRL (2024)
23. Kim, S., Ha, J., Kim, Y.H., Lee, Y., Park, F.C.: ScrewSplat: An End-to-End Method for Articulated Object Recognition. In: CoRL (2025)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: ICCV. pp. 3992–4003 (2023)
25. Le, L., Xie, J., Liang, W., Wang, H.J., Yang, Y., Ma, Y.J., Vedder, K., Krishna, A., Jayaraman, D., Eaton, E.: Articulate-Anything: Automatic Modeling of Articulated Objects via a Vision-Language Foundation Model. In: ICLR (2025)
26. Liu, G., Sun, Q., Huang, H., Ma, C., Guo, Y., Yi, L., Huang, H., Hu, R.: Semi-weakly supervised object kinematic motion prediction. In: CVPR. pp. 21726–21735 (2023)
27. Liu, J., Iliash, D., Chang, A.X., Savva, M., Mahdavi-Amiri, A.: Singapo: Single image controlled generation of articulated parts in objects. In: ICLR (2025)
28. Liu, J., Mahdavi-Amiri, A., Savva, M.: Paris: Part-level reconstruction and motion analysis for articulated objects. In: ICCV. pp. 352–363 (2023)
29. Liu, S., Yu, Y., Pautrat, R., Pollefeys, M., Larsson, V.: 3d line mapping revisited. In: CVPR. pp. 21445–21455 (2023)
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In: ECCV. pp. 38–55 (2024)
31. Liu, X., Zhang, J., Hu, R., Huang, H., Wang, H., Yi, L.: Self-Supervised Category-Level Articulated Object Pose Estimation with Part-Level SE(3) Equivariance. In: ICLR (2023)
32. Liu, Y., Jia, B., Lu, R., Gan, C., Chen, H., Ni, J., Zhu, S.C., Huang, S.: Videoartgs: Building digital twins of articulated objects from monocular video. arXiv preprint arXiv:2509.17647 (2025)
33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
34. Lu, R., Liu, Y., Tang, J., Ni, J., Wang, Y., Wan, D., Zeng, G., Chen, Y., Huang, S.: Dreamart: Generating interactable articulated objects from a single image. In: SIGGRAPH Asia Conference Proceedings (2025)

35. Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., Lu, Z.: Being-h0: Vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597 (2025)
36. Mao, Y., Zhang, Y., Jiang, H., Chang, A.X., Savva, M.: MultiScan: Scalable RGBD Scanning for 3D Environments with Articulated Objects. In: NeurIPS (2022)
37. Pautrat, R., Barath, D., Larsson, V., Oswald, M.R., Pollefeys, M.: Deeplsd: Line segment detection and refinement with deep image gradients. In: CVPR. pp. 17327–17336 (2023)
38. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: CVPR. pp. 9826–9836 (2024)
39. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: Openscene: 3d scene understanding with open vocabularies. In: CVPR. pp. 815–824 (2023)
40. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., Chalk, J., Zhu, Z., Guerrier, R., Abdelazim, F., Zhu, B., Moltisanti, D., Wray, M., Doughty, H., Damen, D.: HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In: CVPR (June 2025)
41. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: WiLoR: End-to-End 3D Hand Localization and Reconstruction In-the-Wild. In: CVPR. pp. 12242–12254 (2025)
42. Qian, S., Jin, L., Rockwell, C., Chen, S., Fouhey, D.F.: Understanding 3d object articulation in internet videos. In: CVPR. pp. 1599–1609 (2022)
43. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
44. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
45. Shen, L., Zhang, S., Li, H., Yang, P., Huang, Z., Zhang, Z., Zhao, H.: GaussianArt: Unified Modeling of Geometry and Motion for Articulated Objects. arXiv preprint arXiv:2508.14891 (2025)
46. Sun, X., Jiang, H., Savva, M., Chang, A.X.: OPDmulti: Openable Part Detection for Multiple Objects. arXiv preprint arXiv:2303.14087 (2023)
47. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV. pp. 581–600 (2020), https://grab.is.tue.mpg.de
48. Teed, Z., Deng, J.: DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In: NeurIPS. vol. 34, pp. 16558–16569 (2021)
49. Tschernezki, V., Darkhalil, A., Zhu, Z., Fouhey, D., Larina, I., Larlus, D., Damen, D., Vedaldi, A.: EPIC Fields: Marrying 3D Geometry and Video Understanding. In: NeurIPS (2023)
50. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: NeurIPS (2025)
51. Weng, Y., Wen, B., Tremblay, J., Blukis, V., Fox, D., Guibas, L., Birchfield, S.: Neural implicit representation for building digital twins of unknown articulated objects. In: CVPR (2024)
52. Werby, A., Buechner, M., Roefer, A., Huang, C., Burgard, W., Valada, A.: Articulated Object Estimation in the Wild. In: CoRL. pp. 3828–3849 (2025)
53. Yang, J., Zuo, X., Wang, S., Yu, Z., Li, X., Ni, B., Gong, M., Cheng, L.: Object wake-up: 3d object rigging from a single image. In: ECCV (2022)

54. Yu, Q., Wang, J., Liu, W., Hao, C., Liu, L., Shao, L., Wang, W., Lu, C.: GAMMA: Generalizable Articulation Modeling and Manipulation for Articulated Objects. In: ICRA. pp. 5419–5426 (2024)
55. Yuan, C., Wen, C., Zhang, T., Gao, Y.: General flow as foundation affordance for scalable robot learning. arXiv preprint arXiv:2401.11439 (2024)
56. Zhang, C., Delitzas, A., Wang, F., Zhang, R., Ji, X., Pollefeys, M., Engelmann, F.: Open-Vocabulary Functional 3D Scene Graphs for Real-World Indoor Spaces. In: CVPR (2025)
57. Zhang, J., Deng, J., Ma, C., Potamias, R.A.: HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos. In: CVPR (2025)