Reconstructing Objects along Hand Interaction Timelines in Egocentric Video

Supplementary Material

1. Overview

On the project's webpage, we provide qualitative videos showcasing qualitative results and include details describing the video in Sec. 2. Rest of this document is arranged as follows. Section 3 provides additional annotation details of the EPIC-HIT and HOT3D-HIT datasets. We ablate the robustness of COP to the boundaries of segments in HIT in Sec 4. Results on Stable Grasp in the ARCTIC dataset [11] are provided in Sec. 6. Additional implementation details are provided in Sec. 7. We then qualitatively evaluate CAD-agnostic models in Sec. 8. Finally, in Sec. 9, we discuss limitations of our work.

2. Qualitative Video

We include videos showcasing the reconstruction results on the two datasets using our proposed approach COP. The video collection contains examples from both EPIC-HIT and HOT3D-HIT. In each case, we show the original video (left), projected reconstruction in camera frame (middle) and 3D hand-object reconstruction from 2 different views (right). We also show the object and hands in world coordinate frame (bottom) with camera pose as a red prism.

Additionally, we provide examples of Stable Grasp sequences in EPIC-HIT. There are two examples from each object category (bottle, can, mug, glass, bowl, cup, plate, pan, saucepan).

3. Annotating HOT3D-HIT and EPIC-HIT

With the definition of HIT and Stable Grasp in Section 3.1 of the main paper, we annotate Hand-Interaction Timelines in two datasets.

3.1. HOT3D-HIT.

For the HOT3D [1] dataset which has 3D ground truth, we automatically extract stable grasps sequences with threshold $\tau = 0.5$ in Equation 1 in the main paper. We locate $1,239$ stable grasps sequences which we then extend automatically to HIT using the annotations to identify when the object is in-view. In total, we label 113 HITs covering $410,650$ frames across 20 videos, $3,288$ segments (872 Static, 1239 Stable Grasp, 1177 Unstable Contact) and 22 objects.

3.2. EPIC-HIT.

We annotate the temporal segments of HIT from the EPIC-KITCHENS [9] videos. This offers a dataset distinct from prior works, which are collected in lab settings [2, 29, 35] or contain recordings specifically collected to evaluate grasps with no underlying action [4, 14, 18]. Instead, we aim to leverage Stable Grasp definition to identify HIT sequences within unscripted egocentric videos of daily actions. Note that we exclude interactions with non-rigid objects and only focus on interactions with rigid known objects. We next detail our annotation pipeline:

1. Identifying candidate clips. The ultimate goal of hand-object reconstruction is to generalize to any rigid or dynamic objects, including those belonging to novel classes. However, as we show later in Sec. 8, current approaches for reconstruction of unknown objects [12, 17, 27, 28, 33] are still in their infancy. We thus restrict our scope to known object categories and focus instead of high-fidelity hand-object reconstruction. Note that this is distinct from assuming instance-level CAD models – the general CAD model of a bottle might not exactly match all bottles in daily life. We exclude tiny objects and shortlist 9 categories frequently

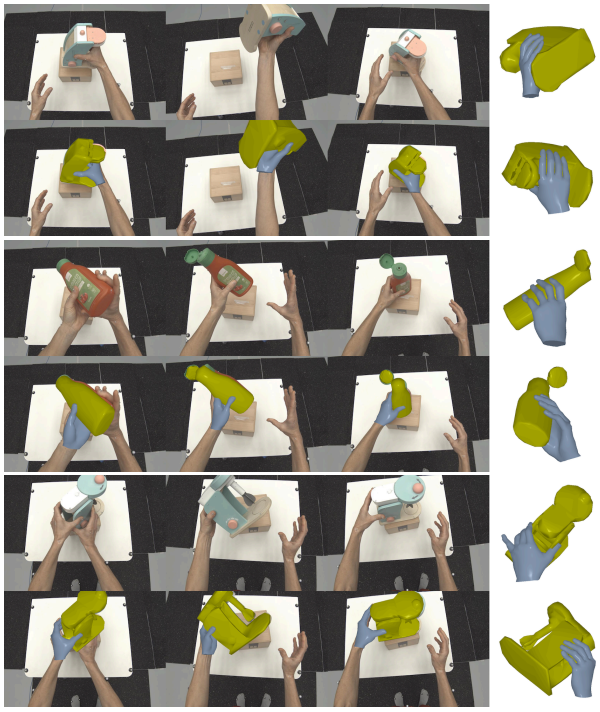

Figure 1. Qualitative results of COP on Stable Grasp from ARCTIC [11]. There are three sequences visualised here. Top row in each sequence contains input frames. Bottom row in each sequence contains frames with reconstructed hand and object. Last column shows the hand and object reconstruction from two different perspectives.

Table 1. Sensitivity to noisy boundary on HOT3D stable grasp subset.

Method	ADD	SCA-ADD
HOMAN [16]	15.0	10.0
COP	70.0	32.9
COP w/ Noisy Stable Grasp Boundaries	60.0	27.4

Figure 2. Eight contact regions: five fingertips V_F + three palm areas. The contact regions serve two purposes: bounding the object inside and attracting the object closer to these regions.

used in kitchens: plate, bowl, bottle, cup, mug, can, pan, saucepan, glass[1]. We use annotations and narrations to find clips where a hand is in contact with one of these categories.

2. Annotating Stable Grasp. Two annotators were asked to label the start-and-end frames following the Stable Grasp definition. We discard segments when, (i) both the hand and object are out-of-view during the sequence or, (ii) the object does not match the category CAD model specified.

In total, we label 2,431 video clips of stable gasps from 141 distinct videos in 31 kitchens [9]. For each clip, we provide a start and end time of the stable grasp, as well as 319,661 segmentation masks for the hand and the object during the stable grasp from the dense VISOR annotations [10]. Of these, 1,446 contain left hand stable grasps and 985 contain right hand stable grasps.

3. Annotating HIT segments. Once we have the stable grasps annotated, we extend them to HIT. We select 42 videos that have verified camera pose estimates from [30] with metric scale and gravity available from [25]. Manual annotations for temporal segments are then added to form consecutive segments labelled with segment type. In total, we label 96 HITs, covering 79,736 frames and 269 segments (135 Static, 106 Stable Grasp, 28 Unstable Contact).

3.3. Dataset Comparison

Table 2 provides a more comprehensive comparison of our datasets with regularly used datasets for hand-object reconstruction. This is an extension of Table 1 in the main paper.

4. Dependency on Accurate Boundaries

COP relies on the provided HIT segment boundaries. One limitation of the method is the need for accurate start-end times of all segments in the hit. These annotations can be relieved if segments are estimated through a localisation model or VLM given a labelled training dataset. While our results in the paper use labelled segments of Hand Interaction Timeline (HIT), we provide an ablation on the need for accurate segment boundaries.

To assess the sensitivity of COP to labelling boundary accuracy, we add random noise—sampled from $\mathrm{Uniform}(10, 30)$ frames—to the ground-truth boundaries for 40 randomly selected Stable Grasp samples from HOT3D. As shown in Table 1, noisy boundaries leads to a performance drop for COP; however, even with such noise, COP still outperforms the baseline [16] by a large margin.

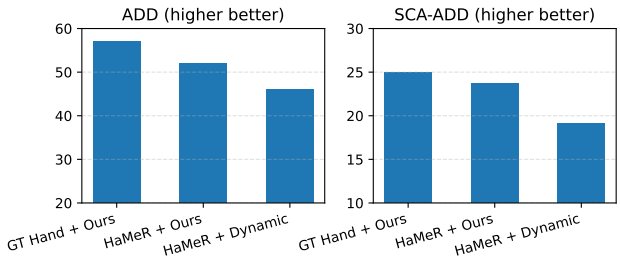

Figure 3. Robustness to noisy hand poses.

5. Sensitivity to Hand Pose Hoise

We analyse our method's sensitivity to the hand pose noise, using a random subset of 100 stable grasp segments from HOT3D. We run HaMeR [24] to obtain the finger poses and use these as input to our method.

Figure 3 compares the results. When we switch from ground-truth to estimated hand poses, our method drops reasonably for both ADD and SCA-ADD metrics. However, our method still clearly outperforms the *best* performing baseline – Dynamic. This finding is in line with the results in the paper.

6. Results on Stable Grasp in ARCTIC

In addition to HOT3D [1] and EPIC-KITCHENS [8], we also explore the ARCTIC dataset [11] with 3D ground truth for HIT reconstruction. However, due to short clips in the dataset, we only evaluate the stable grasp segments on this dataset. Similar to HOT3D, we automatically extract stable grasp sequences with threshold of $\tau = 0.5$ and identify 1303 stable grasp sequences across 9 subjects covering 11 categories. Table 3 contains per-category results for stable grasps in ARCTIC. COP outperforms the baseline [16] and alternate assumptions on all the 11 CAD-model categories. Categories like "capsule machine" see significant improvement in ADD score (+12.6). On average, COP improves ADD from 56.0 with dynamic assumption to 65.1 using the stable grasp assumption.

[1]For **object mesh**, we made per-category CAD model in Blender [7].

Table 2. **Dataset Comparison**. Here we compare various characteristics and labels provided by various datasets. We also show statistics of Stable Grasp and HIT (when available). *: object poses or segments are not provided. †: subjects in the released train/val set

Dataset	Year	Characteristics			Labels			Stable Grasps' Stats					HIT's Stats			
		In-the-wild	Funct. Intent	Ego	Pose GT	Stable Grasp	HIT	#Env	#Sub	#Cat	#Inst	#Seq	Avg. Duration	#frames	Avg. Seq. Per HIT	#Seq
FPHA [13]	2018	✗	✓	✓	3D	✗	✗	3	6	4	4	1,175	-	-	-	-
HO3D [14]	2020	✗	✗	✗	3D	✓(part)	✗	1	10	10	10	65	-	-	-	-
ContactPose [2]	2020	✗	✓	✗	3D	✓	✗	1	50	25	25	2,306	-	-	-	-
GRAB [29]	2020	✗	✓	✗	3D	✗	✗	1	10	51	51	1,334	-	-	-	-
H2O [18]	2021	✗	✓	✓	3D	✗	✗	3	4	8	8	24	-	-	-	-
DexYCB [4]	2021	✗	✗	✗	3D	✗	✗	1	10	20	20	1,000	-	-	-	-
HOI4D [19]	2022	✗	✓	✓	3D	✗	✗	610	9	20	800	5,000	-	-	-	-
Assembly101 [26]	2022	✗	✓	✗	3D Hand*	✗	✗	1	53	15	15	4,321	-	-	-	-
OakInk [32]	2022	✗	✓	✗	3D	✗	✗	1	12	32	100	1,356	-	-	-	-
SHOWMe [28]	2023	✗	✗	✗	3D	✓	✗	1	15	42	42	96	-	-	-	-
ARCTIC [11]	2023	✗	✓	✓	3D	✗	✗	1	9†	11	11	339	-	-	-	-
ARCTIC w/ Stable Grasp	2025	✗	✓	✓	3D	✓	✗	1	9	11	11	1,303	-	-	-	-
HOGraspNet [6]	2024	✗	✗	✓	3D	✓	✗	1	99	30	30	~3861	-	-	-	-
HO-Cap [31]	2024	✗	✗	✓	3D	✗	✗	1	9	64	64	64	-	-	-	-
HOT3D [1]	2024	✗	✓	✓	3D	✗	✗	4	19	33	33	295	-	-	-	-
HOT3D-HIT (ours)	2025	✗	✓	✓	3D	✓	✓	4	9	22	22	1,239	121.1s	410,650	29.1	113
Core50 [20]	2017	✓	✗	✗	2D Mask	✗	✗	11	-	10	50	550	-	-	-	-
MOW [3, 23]	2021	✓	✓	✗	✗	✗	✗	500	500	121	500	500	-	-	-	-
EPIC-HIT (ours)	2025	✓	✓	✓	2D Mask	✓	✓	141	31	9	~390	2,431	13.8s	79,736	2.8	96

Table 3. **Results on ARCTIC.** Green shows the best performing method per metric and yellow shows the second best. COP† is COP without propagation.

Category	SCA-IOU				ADD				SCA-ADD			
	HOMan	Rigid [28]	Dynamic	COP†	HOMan	Rigid [28]	Dynamic	COP†	HOMan	Rigid [28]	Dynamic	COP†
box	30.9	67.3	47.5	71.8	33.3	37.7	52.9	60.1	16.9	28.7	26.7	45.5
capsulemachine	34.4	34.9	43.8	66.6	42.1	48.4	45.3	57.9	24.5	34.1	26.1	43.7
espressomachine	36.9	49.2	52.2	73.0	44.6	48.5	64.4	72.3	28.8	36.2	38.2	55.6
ketchup	18.0	32.7	48.4	62.3	15.1	51.9	39.6	56.6	9.3	37.8	23.7	39.2
laptop	35.3	62.0	51.8	69.1	43.8	45.1	60.4	63.9	28.0	37.1	34.8	46.9
microwave	36.1	51.5	48.8	76.1	56.2	50.9	77.7	83.9	27.3	35.3	41.8	64.2
mixer	34.3	37.9	51.1	69.4	45.1	48.4	66.4	73.8	26.7	32.2	38.8	54.2
notebook	38.7	57.8	55.4	66.6	33.8	43.0	53.6	61.6	20.8	33.2	32.4	42.7
phone	39.5	36.7	52.4	62.2	28.1	39.7	34.9	46.6	19.9	29.2	21.4	30.4
scissors	5.2	0.0	16.2	23.7	7.0	47.4	43.9	70.2	5.2	36.3	25.6	47.7
waffleiron	36.3	48.3	51.0	65.0	45.0	59.5	72.5	76.3	26.2	39.2	40.1	51.5
Average	33.1	46.6	49.0	66.1	37.1	46.9	56.0	65.1	22.0	34.2	32.0	46.9

Table 4. Ablation on the Stable Grasp Loss E_{SG} variants on ARCTIC. We show improvement over the Dynamic Baseline

Obj. Vert. Selection	Frames Selection	ADD	SCA-ADD
V_o	N^2 pairs	65.1 (+9.1)	46.9
v_o^*	N^2 pairs	58.4 (+2.4)	37.2
V_o	N consecutive	59.1 (+3.1)	38.1
Dynamic Baseline		56.0 (+0.0)	32.0

Table 5. **Ablation on the weights.** We highlight our choice of λ_1 and λ_2 (blue) on ARCTIC

λ_1	λ_2	ADD	SCA-ADD
0	0.1	56.0	32.0
1	0	63.0	45.0
1	0.1	65.0	47.0
10	0.1	49.0	39.0

Figure 1 shows qualitative results on Stable Grasp from ARCTIC. In Tab. 4 similar to the ablation in the main paper for HOT3D-HIT, we ablate Stable Grasp Loss E_{SG} and show improvement over the Dynamic baseline. Furthermore, in Tab. 5, we ablate the weights on ARCTIC and draw similar conclusion as the analogous ablation on the HOT3D dataset (Table 7 in the main paper).

7. Additional Implementation Details

Physical Loss E_{push} and E_{pull}. In the main paper, we note our usage of physical repulsion and attraction losses E_{push} and E_{pull}. These are similar to the repulsion and attraction losses in [15].

The term E_{push} ensures all object vertices are located in-

side the contact surface of the hand (Fig. 2). E_{push} applies independently to each frame, hence we omit the superscript ^{n}. For each $v_o \in V_o$, we locate the nearest vertex in hand contact regions, and compute the distance along the surface normal of this hand vertex. Object vertices that penetrate into the contact surface will have negative values. We maximise those negative values, truncating the positive ones:

$$E_{push} = \sum_{v_o \in V_o} -1 * \min(d_v, 0) \quad (1)$$

$$d_v := \langle v_o - v_h^*, n_h^* \rangle \quad (2)$$

where v_h^* is the corresponding nearest vertex on the hand and n_h^* is the surface normal of v_h^*.

In addition to E_{push}, which pushes the object out of the penetrating region against the hand, we use a balancing loss E_{pull} which pulls the object to touch the fingers. E_{pull} also applies independently to each frame and we omit the superscript ^{n}. We here focus on the contact regions showcased in Fig 2. For each finger tip contact region with hand vertices $\{v_h\}_C$, the region-to-object distance is defined as the minimum distance of all (v_h, v_o) pairs. We use 5 finger tip regions and minimise the average of these region-to-object distances.

$$E_{pull} = \frac{1}{5} \sum_C d(\{v_h\}_C, V_o) \quad (3)$$

$$d(\{v_h\}_C, V_o) := \min_{v_h \in \{v_h\}_C, v_o \in V_o} \langle v_h - v_o, n_o \rangle \quad (4)$$

where n_o is the surface normal of v_o.

Pose initialisation for Static segments. As the object is typically supported by a surface when static, we use 10 initialisations all with an *upright* orientation. The initialisations differ in the object's rotation around the axis of support.

Pose initialisation for Stable Grasp segments. When using datasets with 3D ground truth, the initial rotations are generated by clustering the ground-truth rotations, where clustering is performed via the axis-angle representation of the rotation matrix. The initial translation is generated by averaging the ground-truth translations. We initialise 10 rotations and 1 global translation for each (object, left/right hand) pair. For EPIC-HIT, we manually set initial object relative poses to the common poses of each category. Each (category, left/right hand) pair has on average 4.1, minimum 1 and maximum 8 initialisation poses. Importantly, all compared methods (HOMan [16], Rigid, Dynamic, COP) are initialised with these same set of initial poses, ensuring fair comparison in Table-2 and Table-3 in the main paper and Table-3 in the supp.

Pose initialisation for Unstable Contact segments. We use random initialisation.

Computational cost analysis. The main computation is due to mesh projection in E_{mask}; E_{SG} is lightweight for meshes with $\approx$ 500 vertices. The Hand-Interaction Timeline (HIT) propagation is inherently sequential, potential speed-up can be gains through engineering the per-segment optimisation, e.g. multiple initialisation in the same segment can be optimised in parallel.

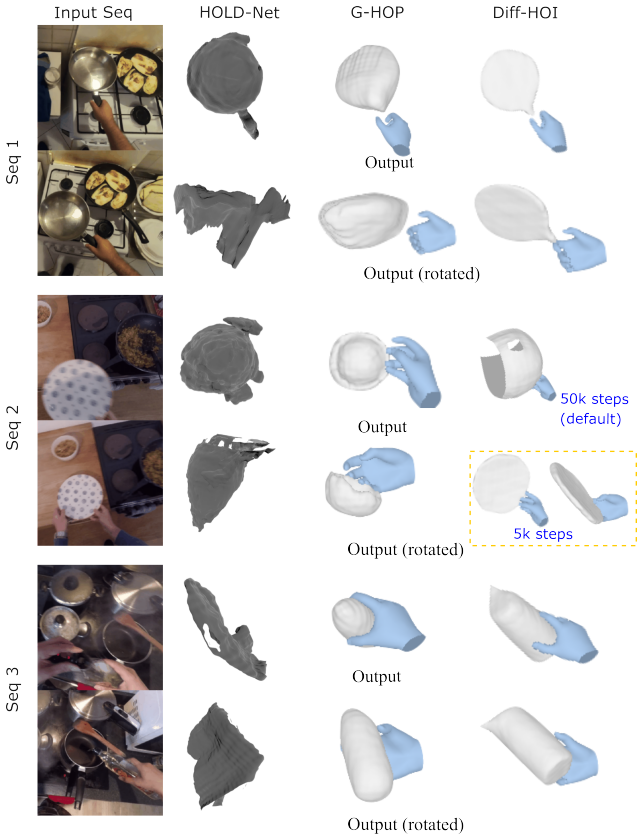

Figure 4. **In-the-wild qualitative evaluation of [12, 33, 34].** Owing to high occlusion due to fingers, the CAD-agnostic methods struggle to reconstruct the object shapes.

8. In-the-wild evaluation of CAD-Agnostic methods

In our method, we assume knowledge of the CAD model. We explore works that attempt reconstruction without CAD model's knowledge. In this section, we showcase these models to be unusable for in-the-wild hand-object reconstruction.

We evaluate CAD-agnostic methods HOLD-Net [12], G-HOP [34] and Diff-HOI [33] on the Stable Grasp from EPIC-HIT dataset. HOLD-Net is a neural rendering based multiple-view method, while G-HOP and Diff-HOI are data-driven methods that learn implicit shape priors from in-the-lab datasets.

Figure 4 shows HOLD-Net is able to reconstruct the object's visible surface. However, HOLD-Net is unable to

Method	ADD↑	SCA-ADD↑	err_{rot} (°) ↓	err_{trans} (cm) ↓
FoundPose [22]	5.0	1.3	78.0	33.5
FoundPose [22] w/o texture	0.0	0.0	100.1	28.9
COP (Ours)	**70.0**	**32.9**	**39.2**	**1.3**

Table 6. Comparison with data-driven methods. We show avg. rotation and translation errors.

generate the complete object surface due to finger occlusion. As input views are typically limited in egocentric videos, HOLD-Net also struggles with the unseen surfaces – the bottle's symmetry is not reconstructed, see the rotated output. In-the-wild videos are also challenging for data-driven methods[2]. In Fig. 4, G-HOP fails to produce the shape for the pan and generates a bowl shape for the plate. Diff-HOI also performs poorly. Diff-HOI can generate the plate shape at an intermediate step (see 5K steps result in yellow square), but produces a wrong shape eventually (at the default 50k steps), highlighting robustness limitations.

Overall, these methods are at an infancy stage. Our method can extend to CAD-Agnostic methods when these are more robust. Importantly, it is not obvious how to quantitatively compare these methods on the same CAD-based metrics due to the need for alignment of the predicted shapes to the ground-truth CAD-model. This alignment is not obvious and has a significant impact on the numerical evaluation.

We also compare against FoundPose [22], a CAD-known but training free method. FoundPose use DINOv2 [21] to build correspondence between the image and the CAD model. Unlike other object pose estimator, FoundPose does not require training, therefore has the potential of scaling to unseen objects. In Tab. 6, we compare results on a random subset of 40 HOT3D stable grasp sequences. FoundPose is significantly worse than COP. Note that FoundPose relies on the texture of the instance CAD model, which is not a requirement for our method.

9. Limitations and Future Direction

Whilst results in-the-wild are very promising, our pipeline relies on hand pose estimation as a first stage. Despite the robustness incorporated by the multiple-view joint optimisation, our method fails when the predicted hand poses are incorrect (see main paper Figure 8). Our method also struggles with extreme occlusions and ambiguity from limited views.

Another limitation of our approach is its reliance on the knowledge of the category's CAD model. We show in Sec. 8 that current CAD-agnostic methods [12, 33, 34] struggle in-the-wild. CAD-agnostic reconstruction and generalisation to unknown objects is the ultimate goal, however

[2]authors of these papers acknowledge their limitations in in-the-wild

Figure 5. SAM-3D results on EPIC-HIT frames from the Static segments of interaction timelines (i.e. when object is not in contact). The predicted models match the CAD models in EPIC-HIT and showcase potential extension into CAD-free assumption.

current approaches do not provide sufficiently representative shapes for hand-object reconstruction where accurate object vertices are required for predicting contact.

In addition, we note that the recently published SAM-3D model [5] could be used to obtain candidate CAD models, examples shown in Figure 5. While SAM3D is not integrated into the proposed pipeline, this is a plausible direction to address the known-CAD limitation.

Finally, we note that our definition of stable grasp is geometry-based. Exploring force closure and physical stability is left for future works.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7061–7071, 2025. 1, 2, 3

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Proceedings of the European Conference on Computer Vision*, pages 361–378, 2020. 1, 3

[3] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12417–12426, 2021. 3

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1, 3

[5] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang,

Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 5

[6] Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, and Tae-Kyun Kim. Dense hand-object(ho) graspnet with full grasping taxonomy and dynamics. In *Proceedings of the European Conference on Computer Vision*, 2024. 3

[7] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*, pages 720–736, 2018. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1, 2

[10] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems*, 2022. 2

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 1, 4, 5

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 3

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2020. 1, 3

[15] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 3

[16] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *International Conference on 3D Vision (3DV)*, pages 659–668, 2021. 2, 4

[17] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing Hand-Held Objects from Monocular Video. In *Proceedings of SIGGRAPH Asia 2022 Conference Papers*, 2022. 1

[18] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10138–10148, 2021. 1, 3

[19] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 3

[20] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26, 2017. 3

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 5

[22] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. Foundpose: Unseen object pose estimation with foundation features. In *Proceedings of the European Conference on Computer Vision*, 2024. 5

[23] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022. 3

[24] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 2

[25] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *2025 International Conference on 3D Vision (3DV)*, 2025. 2

[26] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 3

[27] Edgar Sucar, Kentaro Wada, and Andrew Davison. NodeSLAM: Neural Object Descriptors for Multi-View Shape Reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 1

[28] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. Showme: Benchmarking object-agnostic hand-object 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1935–1944, 2023. 1, 3

[29] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision*, pages 581–600, 2020. 1, 3

[30] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Advances in Neural Information Processing Systems*, 2023. 2

[31] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction, 2024. 3

[32] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 3

[33] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19717–19728, 2023. 1, 4, 5

[34] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 4, 5

[35] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural Manipulation Synthesis with a Hand-Object Spatial Representation. *ACM Transactions on Graphics*, 40(4), 2021. 1