

# SHOW3D: Capturing Scenes of 3D Hands and Objects in the Wild

## Supplementary Material

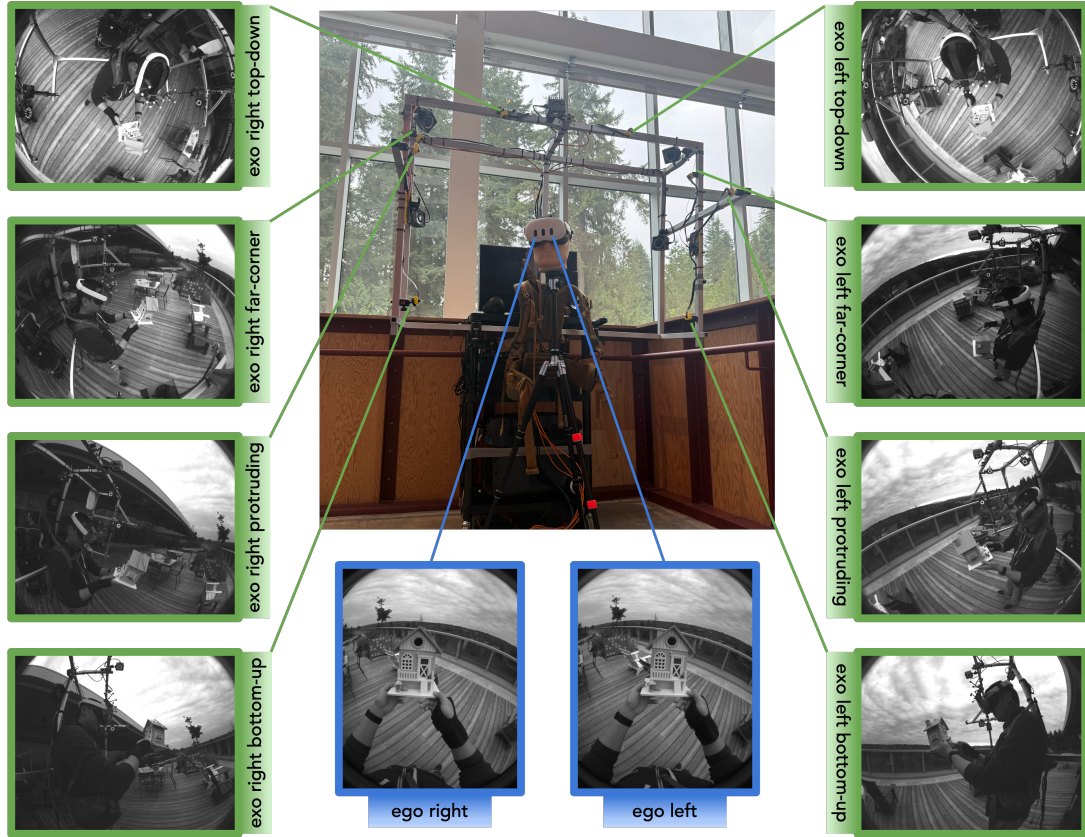


Figure 1. **SHOW3D mobile capture setup.** The rig includes eight exocentric fisheye cameras (green) providing wide-baseline views from diverse angles, and two egocentric fisheye cameras (blue) mounted on the headset. This configuration enables full-surround imaging of hand-object interactions. We use monochrome cameras instead of RGB as they are more robust to illumination variability.

### A. Mobile Capture Rig Details

Figure 1 depicts our mobile capture rig and example images captured by each camera. All camera and MoCap data are streamed, via USB and Ethernet, to a high-performance desktop workstation placed on a rolling cart. The specs of the workstation are: AMD Ryzen Threadripper PRO 3995WX 64-core processor, 128GB RAM, and 30TB SSD storage. We power the workstation with a 2048Wh portable power station, which supports up to three hours of continuous capture.

We employ five OptiTrack Prime 13W MoCap cameras for tracking the IR-reflective markers fixed to the top of the user-worn headset; the set of markers are tracked altogether as a rigid body. We use a custom solution to calibrate the MoCap system, the rig cameras, and the headset cameras all into a shared 3D space. Then, the per-frame extrinsics of the headset cameras can be inferred via the 6DoF transform of the tracked rigid body, such that the 3D ground truth

computed from rig cameras can be precisely projected into the headset cameras for every frame.

All of the monochrome fisheye cameras run an autoexposure policy in real-time, to adapt to variable lighting conditions. Headset cameras are controlled by Quest 3’s on-device autoexposure algorithm, while for the rig cameras, our capture software enforces a target average intensity of 60 over the entire image.

#### A.1. Inclusive Design via Stationary Configuration

To ensure that our dataset captures a diverse range of participants, we designed our system to be adaptable to different physical needs. While the primary configuration is a wearable backpack, the rig can also be fixed to a rolling cart, with a chair placed inside so the participant assumes the same relative position as if wearing it (see Figure 2).

This configuration allows participants to perform hand-



Figure 2. **Stationary configuration for inclusive capture.** The rig can be mounted statically to allow for seated data collection. This setup eliminates the weight burden of the backpack, accommodating participants with limited mobility or physical endurance.

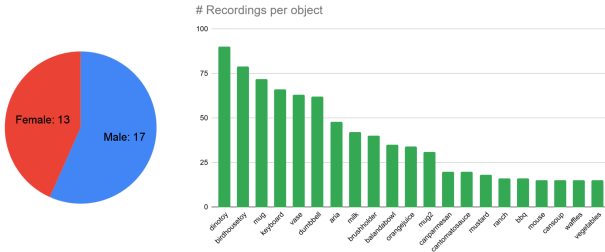


Figure 3. **SHOW3D dataset statistics.** Left: participant distribution. Right: number of recordings per each of the 21 objects.

object interactions while seated, completely removing the physical burden of carrying the backpack rig on the body. This adaptation is critical for inclusivity, as it enables participation from individuals with limited physical endurance, mobility constraints, or those who would otherwise be unable to support the weight of the rig for extended periods. This way, we ensure that our data collection is accessible to, and representative of, a broader demographic.

## B. SHOW3D Dataset Details

**Dataset Construction and Statistics.** We collected around 20 hours of synchronized ego-exo video data at 60Hz, resulting in 4.3M unique frames, or 43M unique images from all cameras. The dataset captures a large variety of indoor/outdoor environments, hand motions, and object interactions. In a subset of around 2M frames, we captured hand-object interaction with a total of 21 objects procured from the list used in HOT3D [1]. The rest of the dataset feature bare-hand motion, hand-hand interaction, or hand-object interaction with other non-HOT3D objects that we currently do not have 3D scans for.

We recruited 38 human participants (24 male, 14 female)



Figure 4. **Examples of contact annotations.** Egocentric frames from SHOW3D with annotated contact regions (yellow). For visual clarity, contact annotations are only shown between the left hand and the manipulated object.

with a wide range of hand shapes and sizes. All participants signed a form stating their consent for their collected data to be used for research purposes. In addition, we employ a commercially available high-resolution optical scanning system to obtain a personalized hand model for each participant. The personalized model is a linear blend skinning mesh in UmeTrack [4] format, which can be fit to detected 3D keypoints via Inverse Kinematics.

**2D Segmentation Masks.** We automatically generate 2D segmentation masks for both hands and objects by projecting the annotated 3D meshes into each camera view using the calibrated intrinsic parameters. We can further utilize the projected mask as a prompt for the Segment Anything Model (SAM 2) [13] to obtain high-quality, refined 2D segmentation masks suitable for training 2D vision tasks.

**Contact Regions.** Leveraging our precise mesh reconstructions, we generate detailed contact annotations by calculating the Euclidean distance from every vertex on the hand mesh to the nearest point on the object surface. Vertices with a distance below some threshold (e.g., 5mm) are labeled as contact points. This results in continuous, vertex-level contact maps (see Figure 4) that capture the functional grasp regions during dynamic interactions.

## C. 3D Hand Annotation Quality

To validate the accuracy of the 3D hand poses generated by our mobile rig via our ego-exo pipeline (see Figure 3 of the main paper, and Figure 7), we compare our automated annotations against two distinct “gold standard” sources: a stationary multi-camera studio system and human manual annotation.

**Experimental Setup.** We define the two evaluation protocols as follows:



Figure 5. **Simultaneous rig-dome capture for ground truth validation.** To evaluate the accuracy of our mobile annotation pipeline, we captured sequences where the subject wears our mobile rig inside a high-fidelity stationary multi-camera dome. This allows for a direct comparison between our mobile-rig-derived annotations and the dome-derived “gold standard.”

- **Dome** (in-studio validation): We engaged a participant to wear our mobile rig *inside* a large-scale, stationary multi-camera dome system equipped with 30 calibrated cameras (see Figure 5). The Dome system provides high-precision automated 3D keypoints via massive multi-view triangulation, serving as a proxy for human-annotated ground truth in a controlled environment. To align the coordinate systems, we attached reflective markers to the mobile rig and utilized the dome’s infrared motion capture system to calibrate the rig into the dome’s 3D space. All data feeds—from both the rig and dome cameras—are synchronized with high precision.
- **Manual** (in-the-wild validation): To validate performance in realistic outdoor conditions where the Dome is unavailable, we sampled a subset of frames from our in-the-wild captures. Crowd-sourced annotators manually labeled 2D keypoints on all available camera views. We then obtained 3D ground truth via robust RANSAC-based triangulation of these manual 2D labels.

**Results and Analysis.** Table 1 reports the median and 90th percentile (P90) Mean Per-Joint Position Error (MPJPE) across three interaction complexity levels.

We observe that the **Manual** reference yields a lower median error than the **Dome** (e.g., 6.36mm vs. 7.90mm for hand-object interactions), but a higher P90 error (16.48mm vs. 12.53mm). We attribute this P90 discrepancy to the independence of the ground-truth generation methods. The

Reference	Category	MPJPE (mm)	
		Median↓	P90↓
Dome	No interactions	6.28	7.31
	Hand-hand interaction	7.46	8.95
	Hand-object interaction	7.90	12.53
Manual (in the wild)	No interactions	5.65	12.35
	Hand-hand interaction	6.03	14.63
	Hand-object interaction	6.36	16.48

Table 1. Quantitative evaluation of our ground-truth hand annotations against two references: a high-coverage 30-camera dome (**Dome**) and manual annotations (**Manual**). We report median and 90th percentile per-frame MPJPE (mm).

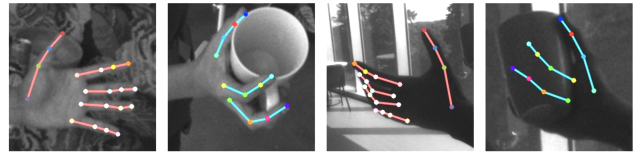


Figure 6. **Examples of manually annotated hand keypoints.** Orange: left hand, blue: right hand. Note that the annotators are instructed to annotate only the keypoints that they have high confidence for. During evaluation, we only evaluate against annotated keypoints that have valid 3D triangulation.

Dome system relies on a multi-view algorithmic pipeline (automated keypoint detection and triangulation) that is methodologically similar to our own ego-exo pipeline. Consequently, the Dome and our system likely share similar failure modes—such as occlusion patterns or challenging lighting conditions—resulting in correlated errors that artificially dampen the P90 values. In contrast, human annotators provide a truly independent ground truth. They do not suffer from algorithmic biases and can accurately label challenging views that automated detectors might miss or misinterpret. Therefore, the Manual evaluation is a more rigorous “stress test,” exposing the true outlier failure cases of our triangulation-based approach that remain hidden when comparing against another automated system.

Importantly, across both evaluation protocols and all interaction types, our pipeline consistently achieves **sub-centimeter median accuracy**. Even in the most challenging scenario—hand-object interaction—our median error remains below 8mm against the Dome and below 7mm against manual annotation. These results confirm that our mobile, markerless capture system generates 3D ground truth of sufficient quality to train downstream egocentric perception models [2–4, 11, 12, 14, 15], successfully bridging the gap between studio fidelity and in-the-wild diversity.



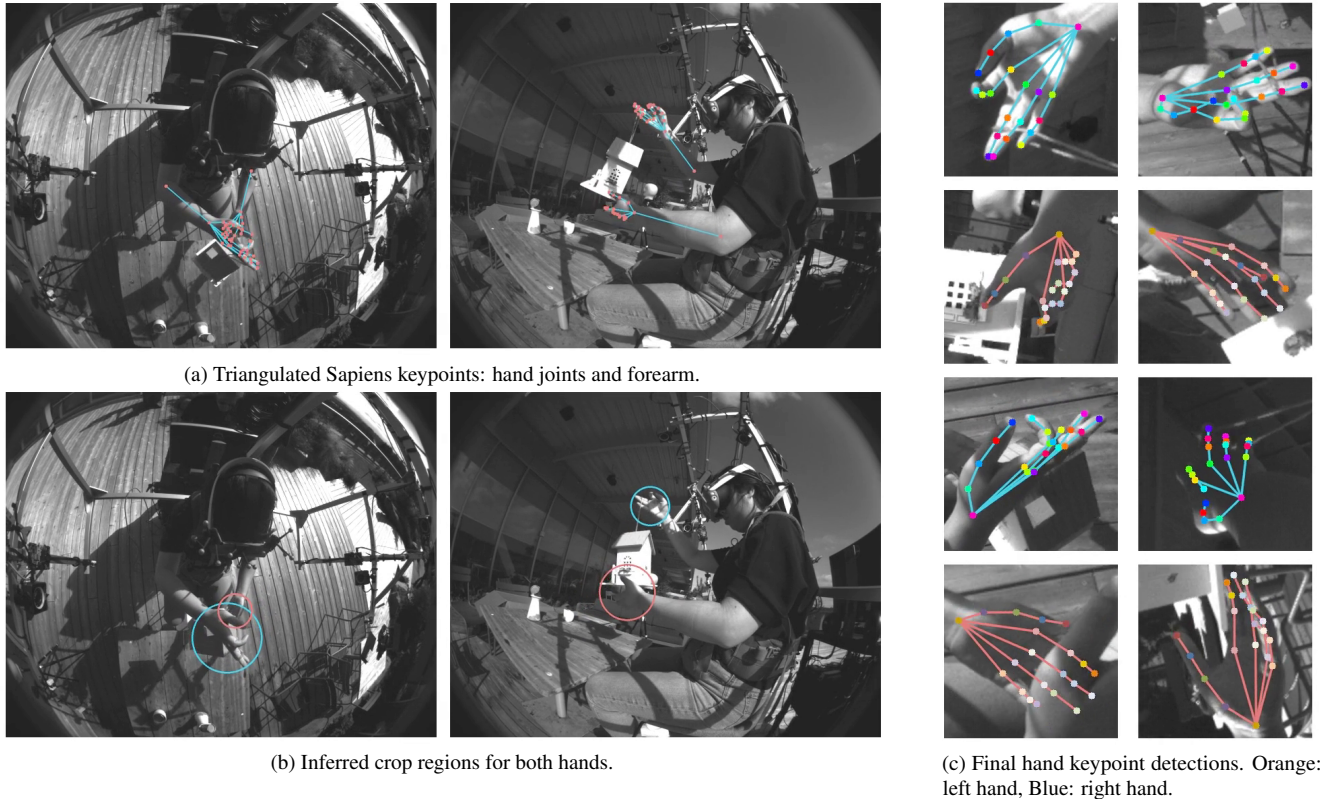


Figure 7. Our automated pipeline for 3D hand keypoint annotation. (a) Sapiens keypoints: we use the triangulated hand and forearm keypoints to define crop regions for hands. (b) Crop regions are used to generate 256x256 perspective crops [4] for both hands. (c) We apply InterNet [6] on the perspective crops, and fuse them with the initial Sapiens detections to obtain final 3D hand keypoints.

## D. 6DoF Object Annotation Quality

To quantitatively evaluate our object pose annotations, we attach optical markers to 5 selected objects and track them using a MoCap system while they are being manipulated by hands. We evaluate our GT object poses against the MoCap object poses (“gold standard”) in Table 2 by comparing the yield (% of annotated frames), translation error (TE), and rotation error (RE). The yield and accuracy of our GT object poses are generally very high, except for `bottle_bbq`, which is a smaller and thus harder to detect object. These results make SHOW3D the first HOI dataset to quantitatively evaluate both its 3D hand pose and 6DoF object pose annotations against independent references.

Object	Yield (%)	P50 TE (mm)↓	P50 RE (deg)↓
birdhouse_toy	89.5	3.51	1.73
carton_oj	92.6	3.20	1.23
dumbbell_5lb	86.1	1.53	1.63
aria_small	77.3	2.79	1.81
bottle_bbq	53.5	2.60	5.77

Table 2. Quantitative evaluation of our ground-truth object annotations against MoCap-derived poses as an independent reference.

## E. Ground Truth Pipeline Details

As mentioned in Section 3.2 of the main paper, we use a multi-stage ego-exo pipeline to obtain 3D hand annotations. We illustrate this pipeline in detail with Figure 7.

While the Sapiens [5] keypoint model exhibits strong generalization on in-the-wild images, we observe that its hand keypoint quality is often suboptimal, likely because the backbone’s feature resolution is insufficient for recognizing finer details. Our insight is that applying a dedicated hand keypoint detector on cropped and resized images helps to improve performance; note that cropping can remove background variations to a large extent, resulting in an easier detection problem. On the other hand, the initial Sapiens keypoint detections are robust enough for defining rough crop regions for both hands, even under heavy occlusion during object interaction.

As described in the main paper, we fuse hand keypoint detections from Sapiens and InterNet [6] via robust triangulation. We assign a real-valued confidence to each triangulated 3D keypoint: this confidence value can be used as a weight on positional constraints in the subsequent Inverse Kinematics (IK) optimization, to adjust the contributions



Frames ahead		<i>pouring in</i>	<i>picking up</i>	<i>pouring out</i>	<i>opening</i>	<i>stirring</i>	<i>placing down</i>	<i>raising</i>	<i>shaking</i>	<i>tapping</i>	<i>washing</i>	<i>checking</i>	<i>inspecting</i>	<i>moving</i>	Mean
30	w/o text	45.8	35.7	56.1	30.4	49.9	38.2	31.0	33.8	29.3	64.5	46.8	61.9	<b>56.2</b>	44.6
	w/ text	<b>28.3</b>	<b>23.6</b>	<b>37.3</b>	<b>20.5</b>	<b>35.0</b>	<b>29.4</b>	<b>24.5</b>	<b>27.5</b>	<b>26.1</b>	<b>59.7</b>	<b>44.9</b>	<b>61.2</b>	59.1	<b>36.7</b>
	Improv.	38.2%	33.9%	33.5%	32.6%	29.9%	23.0%	21.0%	18.6%	10.9%	7.4%	4.1%	1.1%	-5.2%	17.7%
60	w/o text	53.1	41.8	60.6	34.4	48.4	40.1	37.8	40.9	<b>29.9</b>	<b>71.6</b>	55.7	<b>63.8</b>	69.7	49.8
	w/ text	<b>35.4</b>	<b>26.7</b>	<b>42.8</b>	<b>22.6</b>	<b>40.2</b>	<b>31.9</b>	<b>25.7</b>	<b>30.7</b>	32.2	73.6	<b>51.2</b>	63.9	<b>68.6</b>	<b>42.0</b>
	Improv.	33.3%	36.1%	29.4%	34.3%	16.9%	20.4%	32.0%	24.9%	-7.7%	-2.8%	8.1%	-0.2%	1.6%	15.8%

Table 3. Per-verb evaluation of text-driven 6DoF object pose forecasting, for a selected set of verbs. Reported are translation errors (mm, ↓) and the relative percentage improvement provided by text conditioning. The results are sorted by improvement at 30 frames. Our method shows robust gains on predictable actions (*pouring*, *opening*), but has limited impact on ambiguous ones (*moving*, *inspecting*).

from individual keypoints. Our confidence formulation, intuitively, rewards keypoints for having a larger inlier count and a lower reprojection error. Specifically, given RANSAC inlier count threshold  $i_T$  and 2D reprojection error threshold  $e_T$ , and a triangulated keypoint with observed inlier count  $i$  and average reprojection error  $e_{\text{rep}}$  over inlier views, we define the confidence  $C$  as follows:

$$E_{\text{rep}} = \frac{e_T - e_{\text{rep}}}{e_T} \in (0, 1), \quad (1)$$

$$C = (E_{\text{rep}})^{\gamma / \max\{1, i - i_T\}}. \quad (2)$$

First, the fraction  $E_{\text{rep}}$  is inversely related to the average reprojection error, and bounded within  $(0, 1)$ . Then, we reward higher inlier counts by raising  $E_{\text{rep}}$  to a power smaller than 1, thus increasing the resulting confidence value towards 1. The reward factor can be additionally controlled by a free parameter  $\gamma > 0$ .

Finally, we assign a per-frame confidence value to each hand after solving IK. In downstream experiments, we threshold this per-hand confidence to determine valid frames; the threshold can vary depending on the task. We use a Bayesian formulation to compute the per-hand confidence, based on the intuition that the final error in the hand pose is compounded from the errors in both keypoint detection and IK. For each hand, we simply take the product between 1) the average confidence of its associated keypoints, and 2) a term inversely proportional to the residual error from IK solving.

## F. Additional Results on Text-driven 6DoF Object Pose Forecasting

In Table 3 of the main paper, we demonstrate that text conditioning improves forecasting accuracy by 29% and 25% on average across all object categories for predicting 30 and 60 frames ahead, respectively. However, aggregating results by object potentially ignores the underlying mechanism of *how* language aids forecasting. We hypothesize that the “kinematic utility” of a text caption is less dependent on the object itself (e.g., *dumbbell* vs. *keyboard*) and more on the specific interaction protocol (e.g., *pouring* vs. *inspecting*).

To investigate this, we reorganize our evaluation based on the interaction verb (identified and extracted from the full text description) rather than the object class. Table 3 presents the translation errors grouped by a subset of 13 distinct verbs present in SHOW3D text descriptions, sorted by the relative improvement provided by text conditioning in forecasting 6DoF object pose at the 30-frame horizon.

**Impact of Action Predictability.** The results reveal a strong correlation between the *predictability* of an action and the performance gain from text conditioning.

We observe the largest improvements in complex actions that involve specific, structured 3D state changes, such as *pouring in* (38.2% improvement), *picking up* (33.9%), and *opening* (32.6%). In these scenarios, the visual history alone may be ambiguous; for example, a hand approaching a bottle could precede lifting, opening, or sliding. The text caption serves as a strong prior that disambiguates the intent, collapsing the multimodal future distribution into a specific trajectory (e.g., the typical arc that the hand and object usually take to pour).

Conversely, actions characterized by high ambiguity or stochasticity benefit least from having text description as an additional input. In the task of predicting 6DoF object pose 30 frames ahead, *inspecting* (1.1%) and *moving* (-5.2%) show negligible or even negative differences. We attribute this to the wide variability in how users execute these actions; knowing that a user is “inspecting” an object predicts that the object will likely be rotated, but it does not specify the axis or magnitude of that rotation. In these high-entropy cases, the text describes the *state* of the interaction but provides little kinematic signal regarding the future trajectory of the hand and object in 3D space.

This breakdown shows that the value of text-conditioning in SHOW3D is not uniform across hand-object interactions. Language features act as a high-level control signal that is most effective when the semantic intent dictates a specific kinematic protocol, effectively narrowing the search space for the forecasting model.

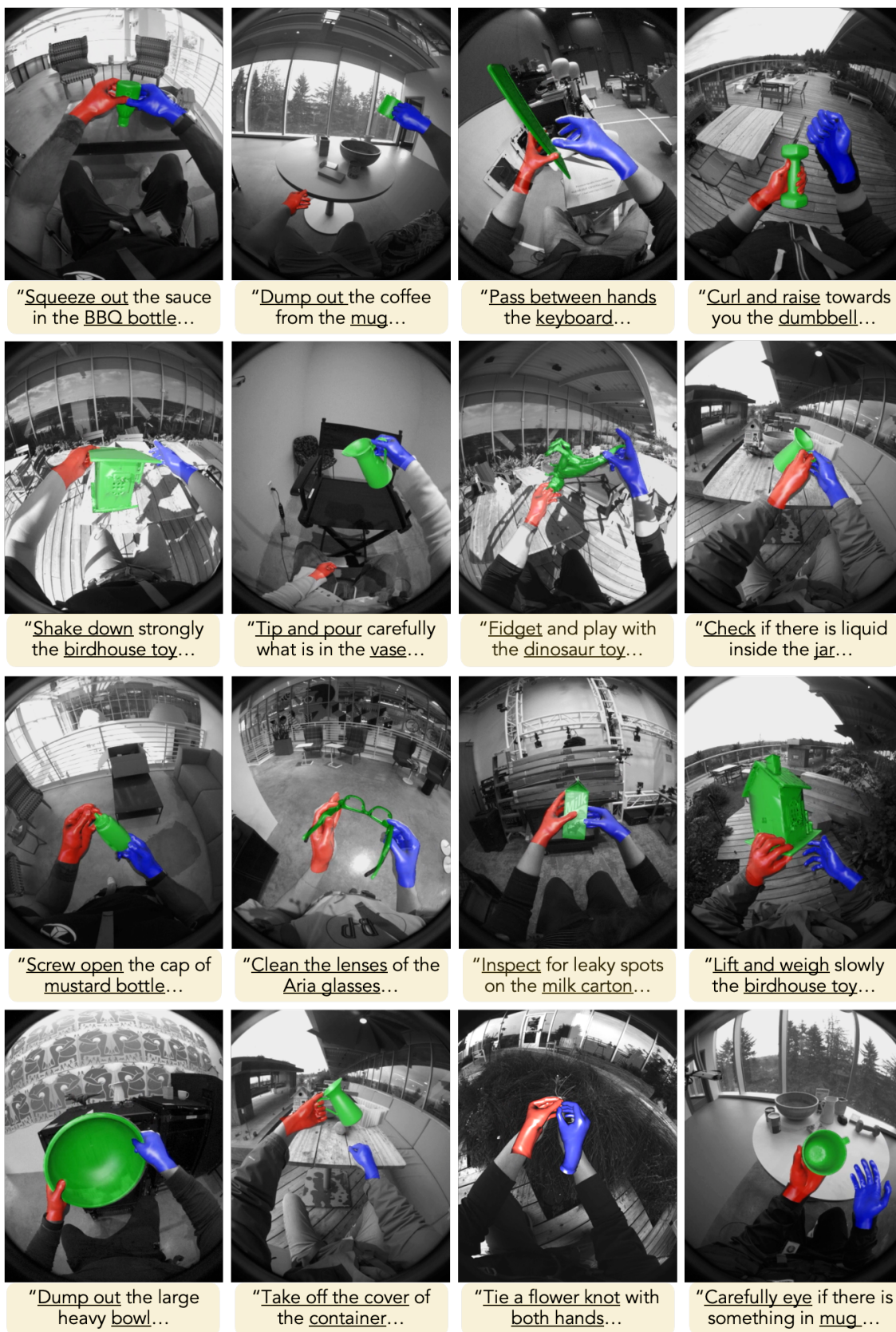


Figure 8. Additional ground truth examples from SHOW3D: hand pose (red and blue), object pose (green), and text captions.

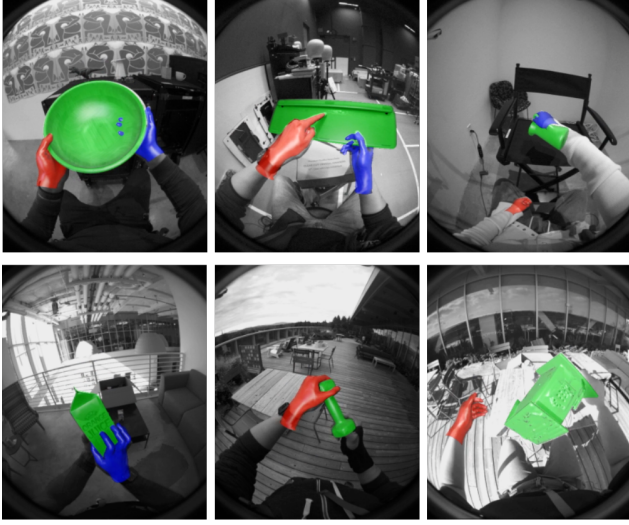


Figure 9. **Hand annotation failure cases.** We visualize two common error modes in our automated annotation pipeline. **Top row:** Examples of **inaccurate 3D pose estimation**. While all entities are successfully detected, minor errors in the estimated 6DoF pose result in physically implausible interpenetrations (clipping) between the hand and object meshes. **Bottom row:** Examples of **missed hand detections**. In scenarios with severe occlusion—where a hand is significantly obstructed by the object or the opposing hand—our pipeline occasionally fails to detect the occluded hand entirely.

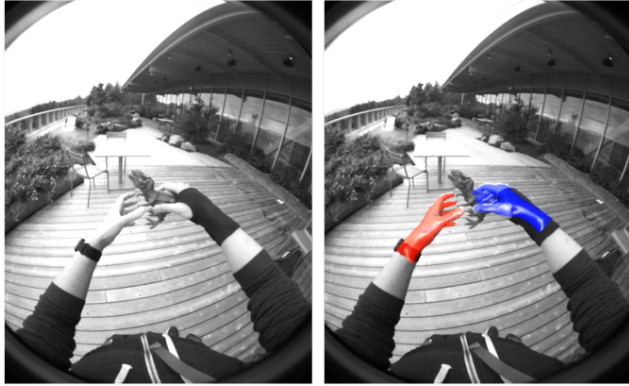


Figure 10. **Object annotation failure case:** Missed object under severe occlusion and outdoor illumination. **Left:** The raw image from egocentric view. **Right:** The projected 3D hand meshes (red/blue) overlaying the image. Although the hands are correctly tracked, the object is not detected due to occlusion from both hands and strong outdoor lighting, and thus is entirely not tracked.

## G. Additional Ground Truth Examples

We provide additional visualizations of our 3D ground truth annotations in Figure 8. These examples further illustrate the diversity of our captured environments and the high fidelity of our markerless tracking pipeline.

**Failure Modes.** To contextualize these examples, Fig-

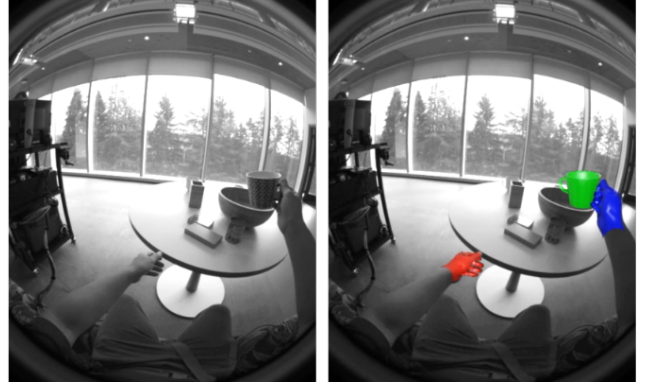


Figure 11. **Object annotation failure case:** Rotational ambiguity in symmetric objects. **Left:** The raw image from egocentric view. **Right:** The projected 3D object mesh (green) overlaying the image. Note that the pipeline correctly tracks the 3D position but fails to resolve the azimuthal rotation, incorrectly predicting the handle to be visible (“hallucinated”) despite not being present in the image.

ure 9 summarizes the two most common failure types observed in our automated hand annotation pipeline. The top row illustrates cases where all entities are successfully detected, but slight inaccuracies in the estimated 6DoF hand pose lead to physically implausible interpenetrations with object meshes. The bottom row shows the opposite failure mode, in which a hand is missed entirely due to heavy occlusion from the object or the opposing hand. While both issues occur infrequently relative to the dataset scale, they underscore the inherent challenges of annotating fine-grained hand–object interactions in egocentric views.

We also observe specific failure modes of our object tracking pipeline. In Figure 10, the hands are successfully tracked but the object is missed entirely due to severe occlusion and strong outdoor illumination. In these settings, our object tracking pipeline, which relies on DINOv2 [9] features, struggles when object features are either heavily occluded or when strong sunlight induces appearance changes, causing the object to go undetected by some or all cameras even when it is partially visible.

A second failure mode is shown in Figure 11, where the object tracking pipeline accurately localizes the 3D position of the cup but fails to estimate its azimuthal rotation, resulting in the handle being “hallucinated” in a visible position (right, green mesh) despite being fully occluded in the input image (left). This likely occurs because the DINOv2 features used in our pipeline [7, 8, 10] are highly semantic and robust to local variations. While this robustness aids in tracking the dominant cylindrical body under occlusion, the patch-based features may lack the spatial sensitivity required to infer the orientation of the thin handle, particularly when the object appears largely rotationally symmetric.



## 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 Computer Vision and Pattern Recognition Conference*, pages 7061–7071, 2025.
- [2] 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/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.
- [3] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020.
- [4] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022.
- [5] Rawal Khrodar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024.
- [6] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020.
- [7] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023.
- [8] Van Nguyen Nguyen, Christian Forster, Sindi Shkodrani, Vincent Lepetit, Bugra Tekin, Cem Keskin, and Tomas Hodan. Gotrack: Generic 6dof object pose refinement and tracking. *arXiv preprint arXiv:2506.07155*, 2025.
- [9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. In *arXiv preprint arXiv:2304.07193*, 2023.
- [10] Evin Pinar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pages 163–182. Springer, 2024.
- [11] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.
- [12] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *European Conference on Computer Vision*, pages 183–202. Springer, 2024.
- [13] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [14] Jian Yang, Jiakun Li, Guoming Li, Zhen Shen, Huai-Yu Wu, Zhaoxin Fan, and Heng Huang. Mlphand: Real time multi-view 3d hand mesh reconstruction via mlp modeling. *arXiv preprint arXiv:2406.16137*, 2024.
- [15] Yu Zhang, Chi Xu, and Li Cheng. Learning to search on manifolds for 3d pose estimation of articulated objects. *arXiv preprint arXiv:1612.00596*, 2016.