arXiv:2603.28760v1 [cs.CV] 30 Mar 2026

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

Patrick Rim[1,2*] Kevin Harris[1] Braden Copple[1] Shangchen Han[1] Xu Xie[1] Ivan Shugurov[1]
Sizhe An[1] He Wen[1] Alex Wong[2] Tomas Hodan[1] Kun He[1]

[1]Meta Reality Labs [2]Yale University

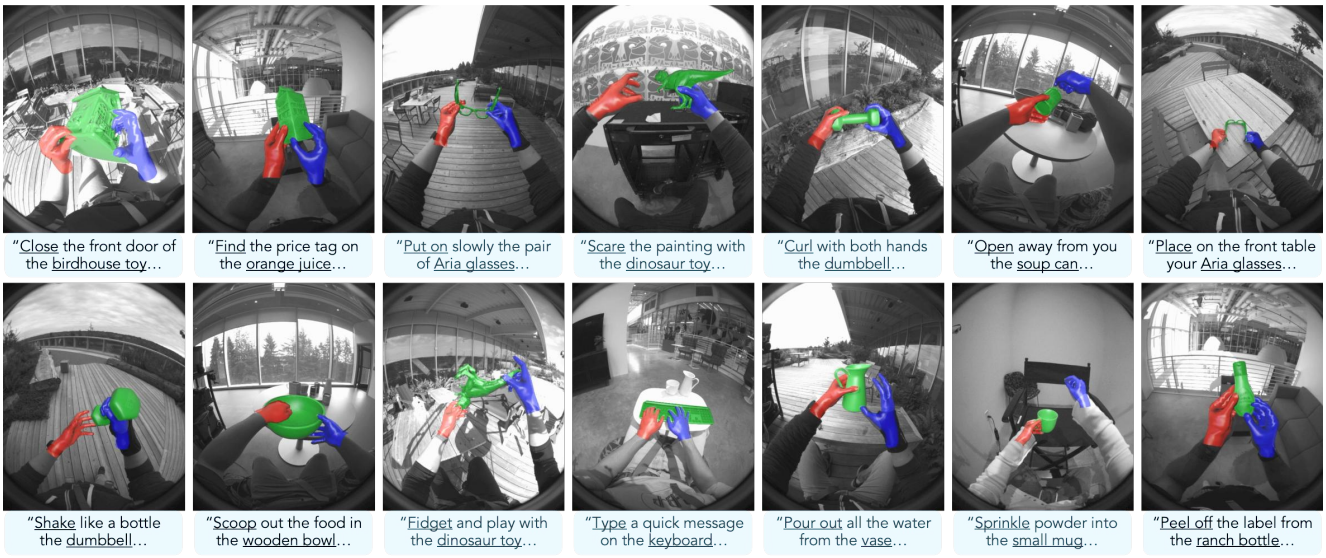


Figure 1. **SHOW3D** is the first dataset of in-the-wild hand–object interactions with accurate 3D annotations as well as text descriptions. The dataset was captured with our novel mobile multi-camera rig in diverse indoor and outdoor scenes, and annotated with 3D shapes and poses with our multi-view pipeline. Overlays show 3D annotations projected to egocentric images (hands in red and blue, object in green).

Abstract

Accurate 3D understanding of human hands and objects during manipulation remains a significant challenge for egocentric computer vision. Existing hand–object interaction datasets are predominantly captured in controlled studio settings, which limits both environmental diversity and the ability of models trained on such data to generalize to real-world scenarios. To address this challenge, we introduce a novel marker-less multi-camera system that allows for nearly unconstrained mobility in genuinely in-the-wild conditions, while still having the ability to generate precise 3D annotations of hands and objects. The capture system consists of a lightweight, back-mounted, multi-camera rig that is synchronized and calibrated with a user-worn VR headset. For 3D ground-truth annotation of hands and objects, we develop an ego-exo tracking pipeline and rigorously evaluate its quality. Finally, we present SHOW3D, the first large-scale dataset with 3D annotations that show hands interacting with objects in diverse real-world environments, including outdoor settings. Our approach significantly reduces the fundamental trade-off between environmental realism and accuracy of 3D annotations, which we validate with experiments on several downstream tasks. `show3d-dataset.github.io`

1. Introduction

Accurate 3D tracking of human hands and their interactions with objects in everyday environments is essential for egocentric perception systems. In augmented and virtual reality (AR/VR), understanding of hands and manipulated objects enables natural and expressive user interactions without reliance on external input devices or constrained input spaces. In robotics and teleoperation, this ability is critical for advising and transferring dexterous manipulation skills.

Despite recent advances, existing hand–object perception models [10, 11, 29, 30, 38, 40] often struggle when deployed outside controlled and indoor environments. There is therefore a pressing need for data collection systems that can produce realistic, unconstrained, and annotated captures of hand–object interactions *in the wild*. However, building such

*Work done during an internship at Meta.

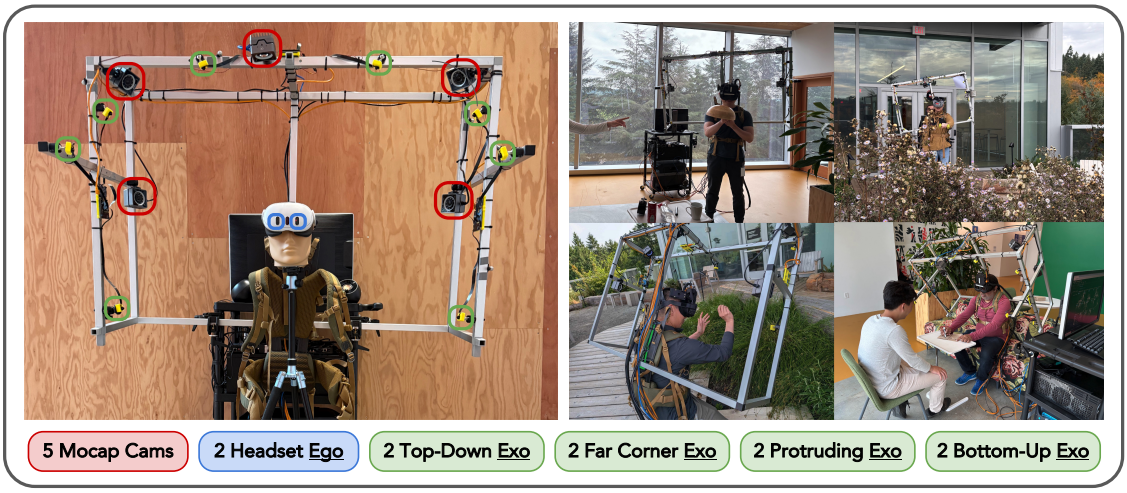


Figure 2. **Our mobile multi-camera capture rig.** Left: Annotated hardware layout showing five MoCap cameras (red), and eight exocentric monochrome cameras mounted in a half-dome configuration (green), and two egocentric monochrome cameras on the Meta Quest 3 headset (blue). The MoCap cameras are only used for headset pose tracking, by tracking optical markers rigidly attached to the headset. The ten total exocentric and egocentric monochrome cameras are used for marker-free annotation of 3D hand and object poses. Right: The rig in use during in-the-wild capture sessions, demonstrating its lightweight (about eight kilograms) and wearable design that allows natural interaction while maintaining stable, synchronized multi-view coverage under mostly unconstrained motion.

systems is highly challenging due to a fundamental trade-off. On one hand, the need for realistic hand/object appearance, environmental diversity, and user mobility limits the application of external sensors, *e.g.*, cameras, IMUs, ToF, and motion capture systems [1, 15, 32, 34, 37]. On the other hand, the availability and quality of 3D ground truth are bounded by sensor fidelity and coverage.

Existing hand–object interaction datasets often fall at one of two extremes in this trade-off. Recent datasets, including GigaHands [8], HOT3D [3], and ARCTIC [7], are captured under controlled conditions, typically within indoor studio spaces equipped with motion capture systems or stationary multi-camera arrays. While these setups can offer high 3D accuracy, they inherently limit environmental diversity and, in the case of marker-based motion capture, hinder visual realism by altering natural hand and object appearance with reflective markers. At the other extreme, datasets with significant environmental variation, such as Ego-Exo4D [9] and Nymeria [20], lack dense or accurate 3D ground truth for hands and objects due to their limited sensor setup.

In this work, we break the trade-off between realism and ground-truth quality by designing a novel multi-camera capture system inspired by prior work [10] but optimized for mobility. Our lightweight backpack-style rig acquires accurate 3D ground truth while enabling users to move freely, allowing for capture of realistic interactions in diverse environments, including the outdoors (Figure 1). The system provides ten fisheye monochrome cameras synchronized at 60Hz, including eight exocentric ones attached on the rig and two egocentric ones from a user-worn Meta Quest 3 headset (Figure 2). To generate accurate ground-truth 3D poses of hands and objects, we develop a novel ego-exo annotation pipeline that leverages all ten camera views. For hand pose annotation, we use state-of-the-art models for keypoint detection, and fuse them via robust triangulation with RANSAC. For object pose annotation, we develop a multi-view extension of recent DINOv2-based methods for 2D detection [24], pose estimation [28], and refinement [25]. Both components produce their own confidence estimates, which allow for automated filtering when building training and evaluation datasets. We assess the quality of both our hand and object pose annotations via evaluation against independent references, demonstrating sub-centimeter accuracy.

We use our novel capture system and annotation pipeline to collect **SHOW3D**, a large-scale dataset of hand–object interactions with dense and accurate 3D annotations, and the first such dataset that is captured across diverse real-world environments. Besides images from ten egocentric and exocentric cameras and 3D pose annotations of hands and objects, the dataset includes fitted 3D meshes, 2D segmentation masks, contact regions, and text descriptions.

Cross-dataset experiments show that models for 3D hand pose estimation and hand–object interaction field estimation trained on existing studio datasets generalize poorly to in-the-wild data, while SHOW3D-trained models maintain strong performance on studio benchmarks, confirming that our dataset captures broader environmental diversity. Furthermore, we demonstrate the utility of our text annotations through object trajectory forecasting experiments, where text-conditioned models significantly outperform baselines relying only on past trajectory, highlighting the value of semantic context for understanding hand–object interactions.

In summary, we make the following contributions:

1. **A mobile data capture system:** The first capture setup for recording and automatically generating 3D annotations of in-the-wild hand–object interactions. Our setup consists of a mobile multi-camera capture rig that enables the acquisition of accurate ground-truth hand and object pose annotations from multi-view imagery.
2. **Ego-exo 3D ground truth:** We build a pipeline to annotate 3D hand and object poses using ego-exo fusion, and automatically filter the resulting annotations using reliable confidence estimates. We also rigorously evaluate the quality of our 3D ground truth for hands and objects.
3. **Large and diverse dataset:** Our dataset SHOW3D comprises 4.3M frames of synchronized multi-view images with comprehensive annotations including 3D hand poses and MANO meshes, 6DoF object poses with 3D models, and descriptive action-level text captions.
4. **Demonstrated dataset utility:** Empirical validation via cross-dataset generalization experiments shows that models trained on SHOW3D transfer better to other datasets than vice versa, and text-conditioned forecasting experiments confirm the value of semantic annotations.

2. Related Work

In this section, we review the most recent efforts in collecting large-scale hand–object datasets. In particular, we emphasize the capture and annotation methodologies for *real data*, as opposed to synthetic data [12, 17, 23]: while synthetic engines could alleviate the scarcity of high-quality data in certain scenarios, there generally remain significant challenges in simulating rich, realistic hand–object dynamics and environmental factors, often resulting in sim-to-real domain gaps that are difficult to overcome.

HOI4D [18] records RGB-D sequences using a Kinect v2 and an Intel RealSense camera mounted on a bicycle helmet worn by participants in indoor environments. Ground truth labels for hands and objects are obtained using manual annotation and frame-wise propagation. Despite its pioneering scale, relying solely on a single egocentric view severely constrains the fidelity of ground truth, as it is challenging to resolve occlusions and scale ambiguity.

Marker-based motion capture (MoCap) systems have been used extensively for ground truth capture purposes. ARCTIC [7] uses an off-the-shelf solution with 54 infrared MoCap cameras to track hands and objects by attaching IR-reflective markers to them. However, the stationary dome-style setup limits environmental diversity, and the presence of markers detracts from visual realism and impedes natural interactions. OakInk2 [39] employs a hybrid capture solution with 12 MoCap cameras and four RGB cameras (three exocentric, one egocentric). Similar to ARCTIC, both the hands and the objects are embedded with markers in order to generate 3D pose ground truth. TACO [19] uses a setup consisting of 12 static exocentric RGB cameras, one egocentric GoPro camera, and a MoCap system. While object poses are still marker-based, hand pose annotations are derived from a marker-less pipeline, using 2D keypoint detection and triangulation. A recent dataset, HOT3D [3], works towards realism by utilizing the Meta Quest 3 headset and Aria research glasses [6]. However, the hands and objects are still tracked with a MoCap system relying on markers.

On the other hand, marker-less solutions typically rely on a synchronized multi-camera setup. Assembly101 [33] and AssemblyHands [26] use 8 RGB cameras synchronized with a custom-built headset to capture hand–object interactions and annotate 3D hand poses, but lack ground-truth object pose. HO-Cap [36] uses multi-view RGB-D sensors to reconstruct 3D hand and object poses, and uses them as source motion in simulation. More recently, GigaHands [8] utilizes an array of 51 RGB cameras to build a 2D-to-3D pipeline for hands and objects. The dataset notably offers a large number of frames compared to previous datasets, but the stationary and restricted capture volume still results in limited environmental and contextual diversity. Furthermore, GigaHands does not contain egocentric views.

Compared to static studios, mobile data capture systems are more difficult to build and operate. A notable large-scale effort is Ego-Exo4D [9], featuring one egocentric view and four exocentric views. However, since its primary focus is on capturing full-body actions in real-world scenarios, the cameras are not optimized for observing hands, and hand annotations are available only in a sparse subset of frames. Moreover, Ego-Exo4D does not provide 3D object annotations. Nymeria [20], a similar real-world dataset captured with Aria research glasses and wearable MoCap suits, also focuses on full-body motion instead of detailed hand–object interaction. EgoDex [13] is an egocentric dataset of dexterous manipulations collected using Apple Vision Pro API for hand pose estimation, but lacks object pose annotations, and is captured in simple studio-like environments.

Across all these existing efforts, an evident trade-off emerges: employing extensive multi-sensor capture systems can yield high-quality annotations, but sacrifices environmental diversity and visual realism. To the best of our knowledge, we introduce the first truly mobile multi-camera capture rig capable of acquiring precise 3D hand–object ground truth for egocentric perception in genuinely unconstrained environments. The closest work to ours is the back-mounted multi-camera system from MEgATrack [10], but it is not designed to capture egocentric data.

3. Data Collection and Annotation

3.1. Mobile Capture Rig

We design a portable, self-contained, backpack-style multi-camera capture rig weighing roughly eight kilograms, which

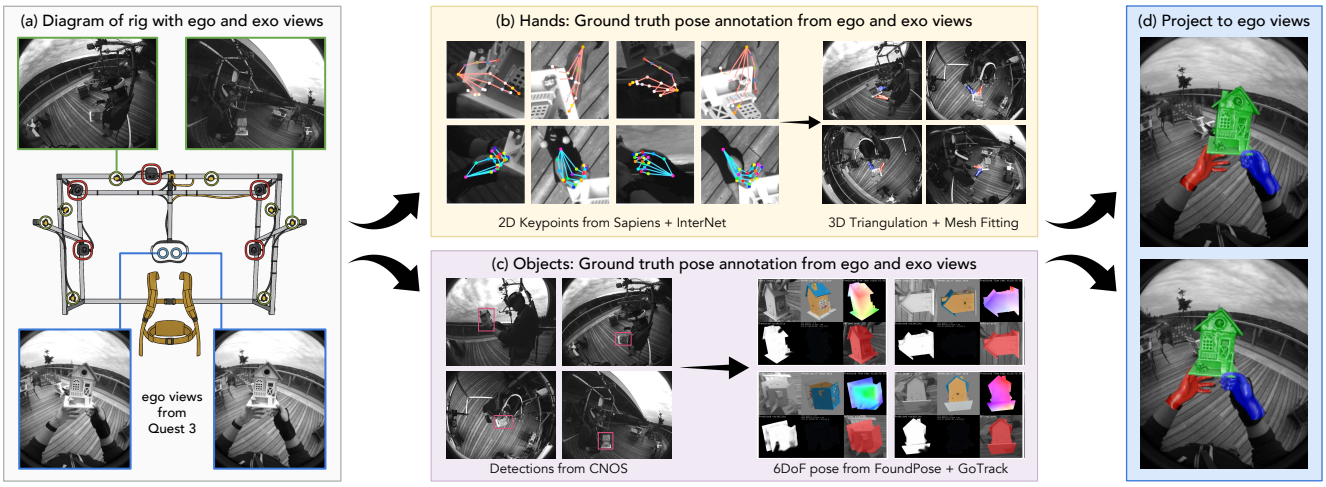


Figure 3. **Our ego-exo pipeline for 3D hand and object pose annotation.** (a) Multi-view fisheye images from our ego and exo cameras. (b) We detect 3D hand keypoints by fusing predictions from Sapiens [14] and InterNet [22], and fit personalized hand mesh via Inverse Kinematics. (c) CAD-based 3D object pose estimation using CNOS [24], FoundPose [28], and GoTrack [25]. (d) The resulting 3D ground-truth annotations are projected back into ego cameras, and can be used to train egocentric vision models.

we found can be worn by human participants without significantly restricting their natural range of motion during a capture session. Eight monochrome fisheye cameras (1024×1280 resolution, $152°$ horizontal and $116°$ vertical field of view) are rigidly mounted on the rig in a half-dome configuration to maximize visual coverage. See Figure 2 for the arrangement of cameras; exact camera placements and viewing angles were determined through empirical testing.

The participants also wear a Meta Quest 3 headset, and we record the egocentric images from the two front-facing monochrome fisheye cameras. To allow for realistic head movements, we do not fix the headset to the rig. Then, to track the headset pose relative to the capture rig, we employ MoCap the same way as in HOT3D [3]: a lightweight marker tree (rigid body) is placed atop the headset, which is tracked via the five MoCap cameras on the rig.

All ten monochrome cameras and the MoCap cameras are hardware-synchronized at 60Hz, and precisely calibrated into a shared 3D reference frame. Note that the reference frame moves with the participant, instead of being fixed to the physical world. Video and MoCap data are streamed to a high-performance desktop workstation placed on a mobile cart, which moves together with the participant guided by an operator, ensuring full mobility.

3.2. 3D Hand Pose Annotation

We design an ego-exo fusion pipeline to obtain 3D hand poses from all ten fisheye cameras in our system. First, we apply Sapiens [14], a state-of-the-art human pose estimation model, to each camera view and obtain body keypoint detections, including 21 keypoints per hand. We use a version of Sapiens finetuned on proprietary human keypoint datasets, including extensive amounts of synthetic data that, in aggregate, can approximate aspects of real-world conditions encountered during in-the-wild captures.

Given the 2D hand keypoint predictions from Sapiens, we next triangulate them into the 3D space defined by the capture rig. However, it is non-trivial to directly triangulate with fisheye distortion. To simplify the problem, we first perform perspective cropping [11] around the region of detected hand keypoints within each camera view, and warp the keypoint detections to the resulting virtual pinhole cameras.

In addition to the Sapiens keypoints, we detect another set of 2D hand keypoints from the perspective crops, using a dedicated hand pose estimation model, InterNet [22], which we also finetuned on additional studio and synthetic datasets. The reasoning is that detecting hand keypoints on cropped images is a simpler problem and can have higher accuracy, whereas directly detecting them from the full image can be suboptimal due to insufficient feature resolution for hands in the Sapiens model. We fuse the two sets of 2D hand keypoint detections by performing RANSAC-based triangulation with both sets. We find that including hand crops from the egocentric cameras often helps performance, as they capture hands from angles that are distinct from the exocentric views.

Next, the triangulated 3D hand keypoints are used to fit detailed hand meshes. We use a solution compatible with UmeTrack [11] and HOT3D [3], where a personalized linear blend skinning model of the hand is first derived from a high-resolution hand scan system; this personalization ensures high-quality ground truth meshes suitable for fine-grained tasks, such as contact estimation. Then, the personalized model is fit to the 3D keypoints via Inverse Kinematics, to obtain a temporally smooth sequence of hand poses. Finally,

Dataset	Ego / Exo	Frames	Hand Pose	Object Pose	Objects	Subjects	Environment	Annotation
SHOW3D (ours)	2 / 8	4.3M	Both / Mesh	✓	21	38	In-the-wild	Monochrome + optimization
GigaHands [8]	0 / 51	3.7M	Both / Mesh	✓	417	51	Studio	RGB + optimization
ARCTIC [7]	1 / 8	2.1M	Both / Mesh	✓	11	10	Studio	MoCap
HOT3D [3]	3 / 0	1.7M	Both / Mesh	✓	33	19	Studio	MoCap
TACO [19]	1 / 12	363K	Both / Mesh	✓	196	14	Studio	MoCap
OakInk2 [39]	1 / 3	993K	Both / Mesh	✓	75	9	Studio	MoCap
HOI4D [18]	2 / 0	2.4M	Both / Mesh	✓	800	4	Studio	RGB-D + manual
HO-Cap [36]	1 / 9	699K	Both / Mesh	✓	64	9	Studio	RGB-D + optimization
HOGraspNet [4]	0 / 4	1.5M	Single / Mesh	✓	30	99	Studio	RGB-D + optimization
Ego-Exo4D [9]	2 / 4	>100M	Both / Skel.	✗	Many	740	In-the-wild	RGB + manual
AssemblyHands [26]	4 / 8	203K	Both / Skel.	✗	Many	20	Studio	RGB + manual
EgoDex [13]	1 / 0	90M	Both / Skel.	✗	Many	–	Studio-like	Apple Vision Pro API

Table 1. **Comparison of SHOW3D against recent hand–object interaction datasets.**

we estimate hand confidence using a Bayesian formulation comprising two terms: one based on the errors in keypoint detection and triangulation, and the other based on the residual error in Inverse Kinematics.

3.3. 3D Object Pose Annotation

To generate accurate 6DoF object pose annotations, we develop a CAD-based pipeline consisting of three stages: (1) 2D object detection by CNOS [24], (2) coarse 6DoF object pose estimation by FoundPose [28], and (3) 6DoF object pose refinement by the GoTrack refiner [25].

We started with open-source implementations of these methods and extended them to support multi-view input images, which was motivated by the accuracy gains achieved with a multi-view extension of FoundPose presented in [3]. The 2D object detector is run on all available views, and the detection with the highest confidence in each view is assumed to correspond to the target object (at most one instance of each object class is assumed present in the scene). The support of calibrated multi-view images in FoundPose and GoTrack is realized by replacing standard PnP with multi-view gPnP from PoseLib [16]. Besides boosting the pose accuracy, we observed that the multi-view input makes final confidence scores produced by the GoTrack refiner more reliable, which is an important property for filtering the generated pose annotations.

The first two stages are executed only in the first frame of a sequence or if the confidence score of the pose estimate from the previous frame is below a threshold. If the confidence score is sufficiently high, we execute only the pose refiner, using the previous pose estimate as the initial pose. This increases efficiency and also robustness in cases where object occlusion is gradually increasing.

All three stages are based on DINOv2 features [27] that allow for effective generalization to various image and camera types. In addition, none of the stages require any object-specific training, thanks to which the pipeline can be promptly applied to any object with an available CAD model. We use a subset of the objects from HOT3D [3] for which high-quality CAD models are available.

3.4. Instructions-to-Text Captions

Inspired by GigaHands [8], each recording session in our dataset is guided by a structured, goal-oriented protocol designed to elicit diverse, natural hand–object interactions from participants. Participants are instructed to perform common object manipulations, such as "spraying with a bottle", "sipping from a mug", or "cleaning a keyboard". Our instructions are intentionally broad and goal-oriented, promoting natural diversity and spontaneity.

To create rich natural language annotations from these instructions, we employ an LLM-based [2] augmentation pipeline that generates multiple semantically diverse paraphrases for each interaction. This process yields a varied collection of text captions for every captured clip, providing valuable paired natural language descriptions that enhance the dataset's utility for language-conditioned models and multimodal training.

4. The SHOW3D Dataset

Using our mobile capture system, we collect SHOW3D, a large-scale dataset comprising over 4 million frames captured from 30 subjects performing natural manipulations with 21 everyday objects. Each frame is simultaneously captured by all ten synchronized cameras, providing comprehensive multi-view coverage. The dataset spans genuinely diverse environments, including a wide range of indoor and outdoor settings that are largely absent from existing hand–object interaction datasets.

Indoor and outdoor scenes. Our outdoor captures include garden areas with natural vegetation, outdoor seating areas with furniture and architectural elements, and even recordings from building ledges where the background is predominantly sky. These outdoor environments exhibit naturally varying illumination conditions due to weather changes,

ranging from bright sunny days to overcast skies, along with dynamic shadows cast by overhead structures, trees, and varying sun angles. Indoor captures encompass kitchens, meeting rooms, hallways, cafeterias, desk areas, and recreational spaces with games and equipment. Indoor settings feature substantial illumination variability due to different artificial lighting configurations—from dim ambient lighting to bright overhead fluorescents—as well as shadow patterns from furniture and architectural features. This environmental and lighting diversity presents challenges rarely seen in studio datasets, where controlled conditions typically maintain consistent illumination and backgrounds.

Annotations. Our annotation pipeline produces rich ground-truth information: 3D hand poses for both hands (21 keypoints per hand with fitted personalized meshes), 6DoF object poses with texture-mapped 3D CAD models, 2D segmentation masks following [31], and hand–object contact annotations following [35]. Multiple natural language captions generated through our LLM-based [2] augmentation pipeline accompany each interaction sequence, offering semantic descriptions at varied levels of specificity.

Annotation quality. SHOW3D is the *first* hand–object interaction dataset to quantitatively evaluate both its 3D hand pose and 6DoF object pose annotations against independent references. To assess the accuracy of our GT *hand* pose annotations, we compare the 3D hand keypoints from our system against gold-standard annotations obtained from two sources: automatic annotation from a large multi-camera studio, and manual annotation on data captured in-the-wild. Across all conditions, our 3D keypoints achieve sub-centimeter median accuracy, approaching that of larger multi-camera systems. To assess the accuracy of our GT *object* pose annotations, we attach optical markers to a set of selected objects and track them using a MoCap system while they are being manipulated by hands. We evaluate our GT object poses against these gold-standard MoCap object poses and find sub-centimeter P50 translation error across all objects, validating the quality of our annotations. We present full experimental results in the supplementary material.

Diversity. As illustrated in Figure 4, SHOW3D occupies a broad manifold in the feature space of hand–object interactions extracted from a pretrained backbone, bridging the compact clusters formed by studio datasets such as HOT3D, GigaHands, and ARCTIC. Frames that appear near these clusters visually correspond to similar conditions (*e.g.*, one SHOW3D scene near the HOT3D cluster was captured in a comparable indoor lab setting, while another near the GigaHands cluster features trusses resembling those that host cameras in GigaHands). This shows that SHOW3D spans both in-the-wild and controlled domains, reflecting genuine environmental overlap while maintaining broader coverage.

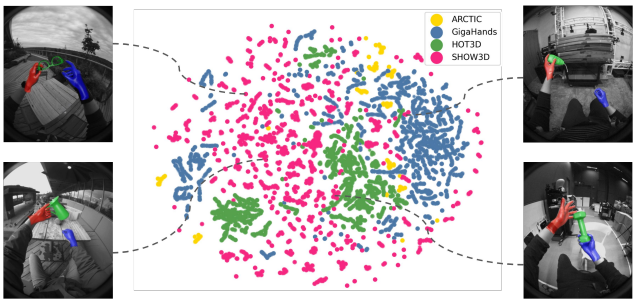


Figure 4. **Cross-dataset feature embedding.** We plot UMAP [21] embeddings of DINOv2 [27] features extracted from the raw images across different hand–object interaction datasets. SHOW3D (pink) spans diverse visual domains between datasets collected in controlled environments: GigaHands (blue), HOT3D (green), ARCTIC (yellow). Best viewed in color.

Comparison with existing datasets. Table 1 positions SHOW3D relative to existing datasets. While studio-based datasets achieve high annotation quality through controlled setups, they sacrifice environmental realism and diversity. Conversely, large-scale in-the-wild datasets like Ego-Exo4D capture authentic interactions but provide only sparse or manual 3D annotations without object pose. SHOW3D uniquely combines both: our mobile rig and RGB-based optimization pipeline enable dense, accurate 3D annotations of hands and objects across truly unconstrained real-world conditions, filling a critical gap in the hand–object interaction literature.

5. Experiments

5.1. Hand–Object Interaction Field Estimation

Existing contact detection methods primarily focus on binary contact estimation, determining whether hands and objects are in contact. However, in dexterous bimanual interactions, hands are often near objects without making contact. To capture richer spatial relationships, we adopt the *interaction field estimation* task introduced in [7] that models the continuous spatial proximity between hands and objects throughout the course of an interaction.

Problem formulation. For each vertex of each hand we aim to infer the closest distance to the object and vice versa. Specifically, an interaction field $\mathbf{F}^{a\to b} \in \mathbb{R}^{V_a}$ is defined as

Train set	Test set	ADE (mm) ↓	ACC (m/s^2) ↓
SHOW3D	HOT3D	14.70	4.05
HOT3D	HOT3D	11.29	3.21
HOT3D + SHOW3D	HOT3D	**8.80**	**2.16**
HOT3D	SHOW3D	22.57	5.61
SHOW3D	SHOW3D	13.82	**3.79**
SHOW3D + HOT3D	SHOW3D	**13.50**	3.84

Table 2. **Evaluation of interaction field estimation.**

Frames ahead		aria	bowl	cansoup	dinotoy	dumbbell	juice	keyboard	milk	mouse	mug	mustard	vase	waffles	Mean
30	w/o text	58.5	43.1	28.3	45.7	89.7	30.3	51.0	31.9	26.6	32.6	57.5	24.2	36.1	42.7
	w/ text	**49.0**	**37.5**	**25.0**	**34.0**	**73.8**	**19.3**	**37.3**	**19.5**	**14.5**	**21.6**	**16.3**	**20.3**	**27.0**	**30.4**
60	w/o text	63.6	40.5	29.0	46.4	95.9	33.1	66.5	**22.6**	27.5	37.4	77.7	27.3	39.9	46.7
	w/ text	**57.0**	**32.7**	**28.4**	**37.5**	**89.0**	**21.1**	**34.6**	25.2	**24.7**	**24.4**	**19.0**	**26.3**	**35.1**	**35.0**

Table 3. **Evaluation of text-driven 6DoF object pose forecasting.** Reported are per-object and mean average translation errors (mm, ↓), with and without text conditioning. Conditioning on text leads to a consistent improvement across objects.

the distance to the closest vertex on mesh $\mathbf{M}_b$ for all vertices in mesh $\mathbf{M}_a$. V_a denotes the number of vertices in mesh $\mathbf{M}_a$ (the reverse interaction field would be $\in \mathbb{R}^{V_b}$). We can formulate this problem mathematically, as follows:

$$\mathbf{F}_i^{a\to b} = \min_{1\le j\le V_b} \left\|\mathbf{v}_i^a - \mathbf{v}_j^b\right\|_2, \qquad 1 \le i \le V_a, \quad (1)$$

where $\mathbf{v}_k^{(\cdot)} \in \mathbb{R}^3$ represents the k-th vertex of mesh $\mathbf{M}_{(\cdot)}$. Our task is to estimate the interaction fields $\mathbf{F}^{l\to o}$, $\mathbf{F}^{r\to o}$, $\mathbf{F}^{o\to l}$, and $\mathbf{F}^{o\to r}$ for each frame, where l, r, and o denote the left hand, right hand, and object meshes, respectively.

Experimental setup. We train the InterField [7] model and evaluate cross-dataset generalization between SHOW3D and HOT3D. We report two metrics: Average Distance Error (ADE) in millimeters, which measures the mean absolute error in predicted distances, and Acceleration (ACC) in m/s^2, which captures temporal smoothness of predictions.

Results. Table 2 presents cross-dataset evaluation results. When testing on HOT3D, training on SHOW3D alone achieves 14.70 mm ADE compared to 22.57 mm (+54%) when training on HOT3D and testing on SHOW3D. This asymmetry demonstrates that models trained on SHOW3D generalize better to studio environments than vice versa, indicating that our in-the-wild data captures a broader distribution of interaction patterns and environmental conditions.

Furthermore, when using both datasets for training, we observe substantial improvements on HOT3D (8.80 vs. 11.29 mm, a 22% improvement) but minimal gains for SHOW3D evaluation (13.50 vs. 13.82 mm, only 2% improvement). This reveals that the domain of SHOW3D substantially encompasses the more constrained studio setting of HOT3D, while HOT3D provides limited additional coverage beyond what SHOW3D already captures. These results confirm that the environmental diversity in SHOW3D directly translates into improved model generalization, validating the value of in-the-wild hand–object interaction data.

5.2. Text-driven 6DoF Object Pose Forecasting

To validate the utility of our semantic text annotations, we evaluate whether natural language descriptions of hand-object interactions provide valuable context for predicting future object poses. Given a sequence of observed object poses and an optional text caption describing the interaction, the task is to forecast the 6DoF object pose at a future time step. Natural language captures high-level interaction intent, such as "pouring from a bottle" versus "placing a bottle down", which helps disambiguate similar pose trajectories that lead to different future outcomes.

Experimental setup. We implement a transformer-based forecasting model that takes as input a sequence of past object poses, where each pose is represented as a 7-dimensional vector (3D position and quaternion for 3D rotation). The pose sequence is embedded and augmented with sinusoidal positional encodings before being processed by a multi-layer transformer encoder. For text conditioning, we encode captions using BERT [5] and project the text features into the same latent space as the pose embeddings. Text and pose features are concatenated and processed jointly by the transformer. The model predicts future object pose using separate heads for position (3D translation) and rotation (normalized quaternion). We train models both with and without text conditioning, using sequences of 30 past poses to predict the pose 30 or 60 frames into the future. We evaluate on sequences from our test set and report average translation error in millimeters across 13 objects.

Results. Table 3 presents per-object and mean translation errors for both forecasting horizons. At 30 frames ahead, text conditioning reduces mean error from 42.7 to 30.4mm, a 29% improvement. This gain is consistent across nearly all objects, with particularly large improvements for objects like mug (34%) and mustard (72%). At the longer 60-frame horizon, text conditioning maintains its advantage, reducing mean error from 46.7 to 35.0mm (25%).

These results demonstrate that semantic descriptions provide meaningful contextual information that complements purely kinematic patterns. The consistent improvements across objects and time horizons validate the value of our text annotation pipeline and highlight the potential of language-conditioned models for understanding and predicting hand–object interactions. Ultimately, SHOW3D's multimodal annotations enable richer downstream applications beyond purely visual perception.

#	Training set	Test set	MKPE (mm) ↓
1	UmeTrack	SHOW3D	22.2 (+55%)
2	HOT3D	SHOW3D	19.6 (+37%)
3	UmeTrack + HOT3D	SHOW3D	16.4 (+15%)
4	SHOW3D	SHOW3D	15.5 (+8%)
5	UmeTrack + HOT3D + SHOW3D	SHOW3D	**14.3**
6	HOT3D	HOT3D	14.0 (+14%)
7	UmeTrack + HOT3D	HOT3D	12.7 (+3%)
8	UmeTrack + HOT3D + SHOW3D	HOT3D	**12.3**
9	UmeTrack	UmeTrack	9.7 (+2%)
10	UmeTrack + HOT3D	UmeTrack	**9.5**
11	UmeTrack + HOT3D + SHOW3D	UmeTrack	9.6 (+1%)

Table 4. **Evaluation of 3D hand pose estimation.** The reported metric is Mean Keypoint Position Error (MKPE, in mm). Models trained on existing datasets, UmeTrack [11] and HOT3D [3], generalize noticeably worse to SHOW3D, highlighting the increased difficulty of in-the-wild data.

5.3. 3D Hand Pose Estimation

Experimental setup. We evaluate the utility of our dataset as a benchmark for 3D hand pose estimation. We sample four subjects (with gender balance) from SHOW3D to build a balanced evaluation set, and apply a threshold on the estimated hand pose confidence described in Sec. 3.2 to the entire dataset to filter valid frames for training and evaluation. We benchmark generalization by training the multi-view UmeTrack hand tracking model [11] on SHOW3D, as well as evaluating cross-dataset generalization on the existing UmeTrack and HOT3D datasets.

Since we rely on marker-less ground truth generation in genuinely challenging in-the-wild conditions, the yield rate of training frames (*i.e.* frames with valid hand pose annotation) in SHOW3D—using automated filtering—is naturally less than perfect. Overall, SHOW3D retains 1.9M high-confidence training frames for hand pose estimation. Notably, this is greater than either UmeTrack or HOT3D, both of which have fewer than 1.4M training frames.

Results. As shown in Table 4, models trained on UmeTrack or HOT3D can achieve reasonable performance when tested within their respective domains (rows 6 and 9), but regress substantially when evaluated on SHOW3D (rows 1 and 2). A model trained on SHOW3D (row 4) achieves a significantly lower error on SHOW3D than a model trained on UmeTrack, HOT3D, or their combination (rows 1, 2, and 3). We also observe that combining SHOW3D with the other datasets is consistently beneficial. The model trained on the combination of UmeTrack, HOT3D, and SHOW3D outperforms other models when evaluated on SHOW3D and HOT3D (rows 5 and 8), and is a close second on UmeTrack (row 10), which was captured in a very restricted environment.

Figure 5 shows qualitative examples on the SHOW3D test set. First, due to the lack of object interactions in the UmeTrack training set, the UmeTrack-trained model struggles in the face of even minor occlusions from objects. We observe that the HOT3D-trained model handles occlusions better, but still achieves suboptimal performance, especially in environments with novel background/lighting. Finally, the model trained with SHOW3D data shows significantly improved robustness to object interaction and background clutter. Once again, these results highlight the unique value of real in-the-wild data with 3D ground truth, which offers robustness benefits not captured by prior datasets collected in more controlled environments (indoor studios).

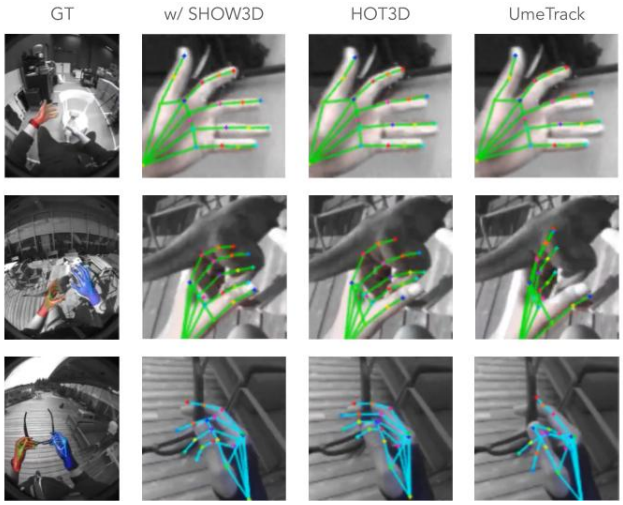


Figure 5. **Hand pose estimation on the SHOW3D test set.** From left to right: results from models trained on UmeTrack + HOT3D + SHOW3D, HOT3D, and UmeTrack, respectively. Including training data from SHOW3D significantly improves robustness against object occlusion and background clutter.

6. Conclusion

We present SHOW3D, the first in-the-wild dataset with 3D ground truth for egocentric perception of hand–object interaction. To capture real-world diversity, SHOW3D features over 1,200 video sequences collected across more than 30 distinct locations. SHOW3D is captured with a lightweight multi-camera rig optimized for high mobility, and annotated by an automated ego-exo pipeline that produces accurate 3D hand and object poses. It is demonstrated on three downstream tasks that SHOW3D enables learning robust and generalizable models, proving its value over existing datasets that are predominantly captured in restricted environments.

While this work has focused on the visual understanding of hand–object interaction, future work will explore the integration of additional modalities, such as tactile sensing and range. Furthermore, we are interested in expanding the downstream use cases of our data collection capabilities in embodied applications, such as teleoperation for robotics.

References

[1] Dimitrios S Alexiadis, Anargyros Chatzitofis, Nikolaos Zioulis, Olga Zoidi, Georgios Louizis, Dimitrios Zarpalas, and Petros Daras. An integrated platform for live 3d human reconstruction and motion capturing. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):798–813, 2016. 2

[2] Anthropic. Claude: A family of large language models. 2024. Accessed: 2025-11-14. 5, 6

[3] 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, 3, 4, 5, 8, 12

[4] Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, et al. Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In *European Conference on Computer Vision*, pages 284–303. Springer, 2024. 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019. 7

[6] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. 3

[7] 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. 2, 3, 5, 6, 7, 13

[8] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17461–17474, 2025. 2, 3, 5

[9] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 3, 5

[10] 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. 1, 2, 3, 13

[11] 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. 1, 4, 8, 12, 13, 14

[12] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024. 3

[13] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2025. 3, 5

[14] Rawal Khirodkar, 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. 4, 14

[15] Young Min Kim, Christian Theobalt, James Diebel, Jana Kosecka, Branislav Miscusik, and Sebastian Thrun. Multi-view image and tof sensor fusion for dense 3d reconstruction. In *2009 IEEE 12th international conference on computer vision workshops, ICCV workshops*, pages 1542–1549. IEEE, 2009. 2

[16] Viktor Larsson and contributors. PoseLib - Minimal Solvers for Camera Pose Estimation, 2020. 5

[17] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20395–20405, 2023. 3

[18] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 3, 5

[19] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, pages 21740–21751, 2024. 3, 5

[20] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2024. 2, 3

[21] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 6

[22] Gyeongsik Moon, Shoou-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. 4, 14

[23] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[24] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, 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. 2, 4, 5, 17

[25] 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. 2, 4, 5, 17

[26] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12999–13008, 2023. 3, 5

[27] 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. 5, 6, 17

[28] Evin Pınar 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. 2, 4, 5, 17

[29] 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. 1, 13

[30] 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. 1, 13

[31] 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. 6, 12

[32] Patrick Rim, Hyoungseob Park, Ziyao Zeng, Younjoon Chung, and Alex Wong. Protodepth: Unsupervised continual depth completion with prototypes. In *Proceedings of the computer vision and pattern recognition conference*, pages 6304–6316, 2025. 2

[33] 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

[34] Yue Sun, Zhuoming Huang, Honggang Zhang, Zhi Cao, and Deqiang Xu. 3drimr: 3d reconstruction and imaging via mmwave radar based on deep learning. In *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pages 1–8. IEEE, 2021. 2

[35] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 6

[36] 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, 5

[37] Chao Xia, Chenfeng Xu, Patrick Rim, Mingyu Ding, Nanning Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Quadric representations for lidar odometry, mapping and localization. *IEEE Robotics and Automation Letters*, 8 (8):5023–5030, 2023. 2

[38] 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. 1, 13

[39] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024. 3, 5

[40] 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. 1, 13

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

Supplementary Material

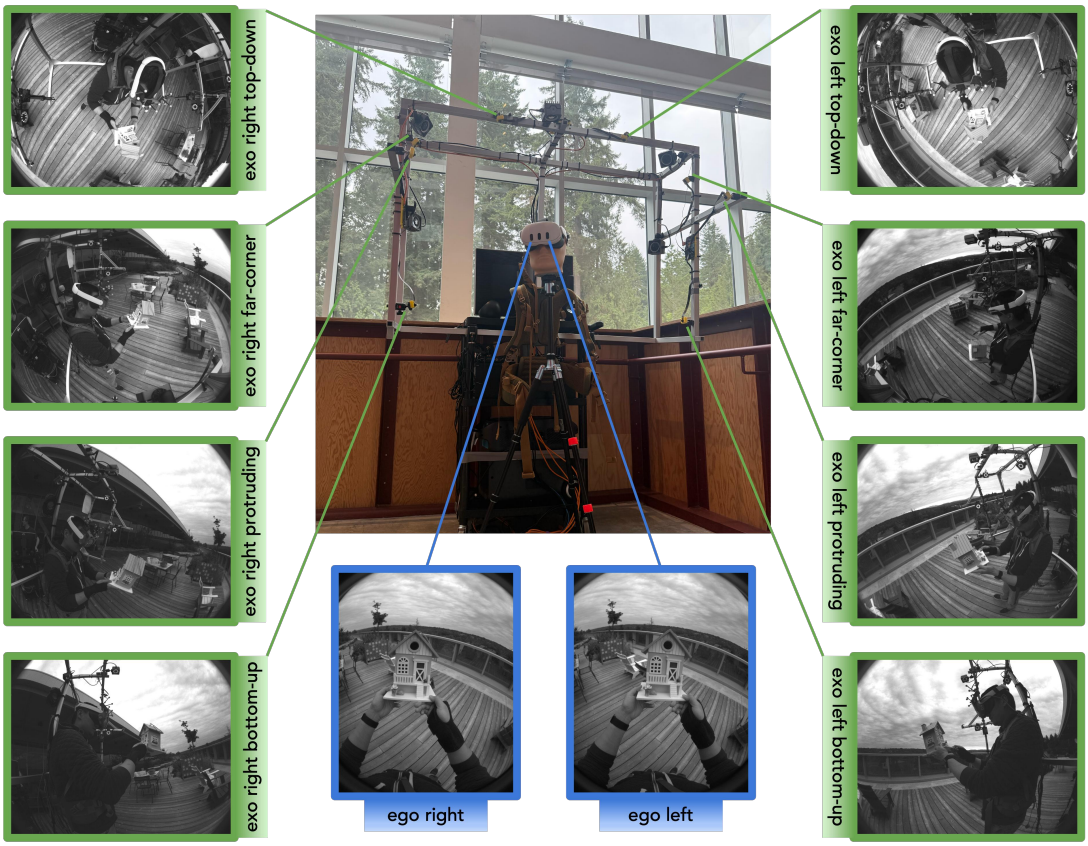


Figure 6. **SHOW3D mobile capture setup.** The rig includes eight exocentric fisheye cameras (green) providing wide-baseline views from diverse angles, and two egocentric fisheye cameras (blue) mounted on the headset. This configuration enables full-surround imaging of hand–object interactions. We use monochrome cameras instead of RGB as they are more robust to illumination variability.

A. Mobile Capture Rig Details

Figure 6 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 7).

This configuration allows participants to perform hand-



Figure 7. **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.

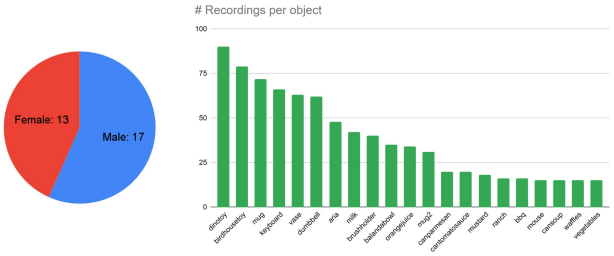


Figure 8. **SHOW3D dataset statistics.** Left: participant distribution. Right: number of recordings per each of the 21 objects.

object interactions while seated, completely removing the physical burden of carrying the backstrap 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 [3]. 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 9. **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 [11] 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) [31] 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 9) 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 12), 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 10. **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 10). 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 5 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↓	MPJPE (mm) 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 5. 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).

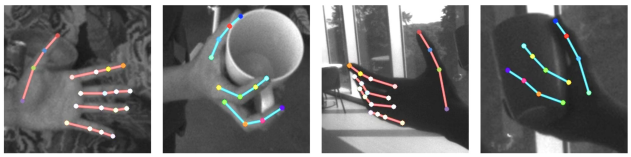
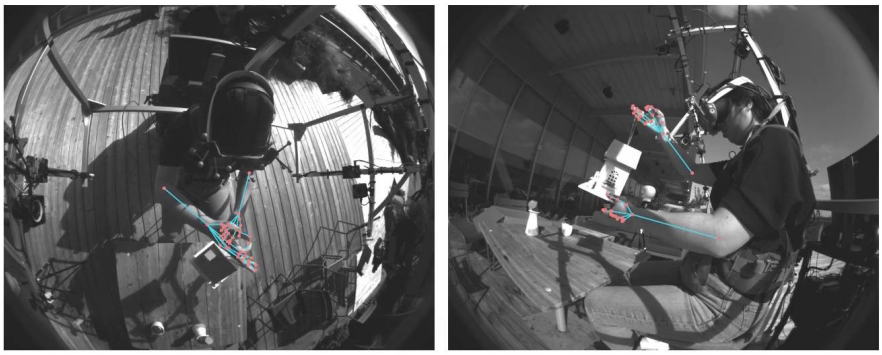


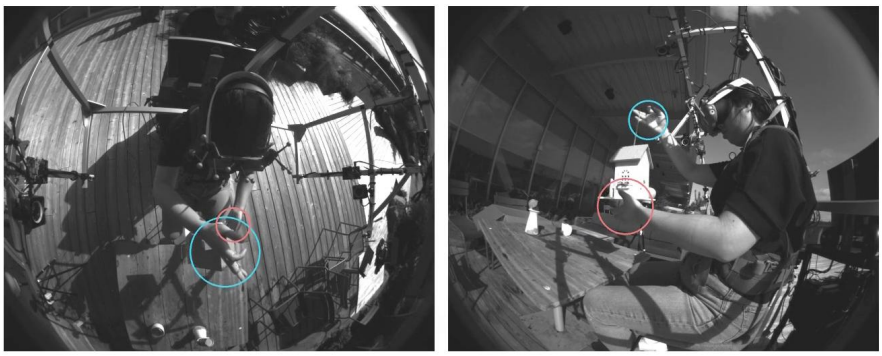
Figure 11. **Examples of manually annotated hand keypoints.** Orange: left hand, blue: right hand. Note that the annotators are instructed to annotate only the keypoints that they have high confidence for. During evaluation, we only evaluate against annotated keypoints that have valid 3D triangulation.

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.

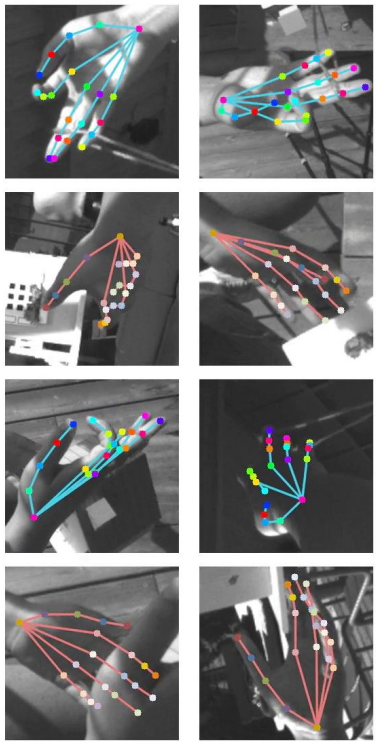
Importantly, across both evaluation protocols and all interaction types, our pipeline consistently achieves **sub-centimeter median accuracy**. Even in the most challenging scenario—hand-object interaction—our median error remains below 8mm against the Dome and below 7mm against manual annotation. These results confirm that our mobile, markerless capture system generates 3D ground truth of sufficient quality to train downstream egocentric perception models [7, 10, 11, 29, 30, 38, 40], successfully bridging the gap between studio fidelity and in-the-wild diversity.



(a) Triangulated Sapiens keypoints: hand joints and forearm.



(b) Inferred crop regions for both hands.



(c) Final hand keypoint detections. Orange: left hand, Blue: right hand.

Figure 12. 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 [11] for both hands. (c) We apply InterNet [22] 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 6 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 6. 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 pipline in detail with Figure 12.

While the Sapiens [14] 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 [22] 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		pouring in	picking up	pouring out	opening	stirring	placing down	raising	shaking	tapping	washing	checking	inspecting	moving	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	**56.2**	44.6
	w/ text	**28.3**	**23.6**	**37.3**	**20.5**	**35.0**	**29.4**	**24.5**	**27.5**	**26.1**	**59.7**	**44.9**	**61.2**	59.1	**36.7**
	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	**29.9**	**71.6**	55.7	**63.8**	69.7	49.8
	w/ text	**35.4**	**26.7**	**42.8**	**22.6**	**40.2**	**31.9**	**25.7**	**30.7**	32.2	73.6	**51.2**	63.9	**68.6**	**42.0**
	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 7. 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_{rep} over inlier views, we define the confidence C as follows:

$$E_{\text{rep}} = \frac{e_{\text{T}} - e_{\text{rep}}}{e_{\text{T}}} \in (0, 1), \tag{2}$$

$$C = (E_{\text{rep}})^{\gamma / max\{1, i - i_{\text{T}}\}}. \tag{3}$$

First, the fraction E_{rep} is inversely related to the average reprojection error, and bounded within $(0, 1)$. Then, we reward higher inlier counts by raising E_{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 7 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.

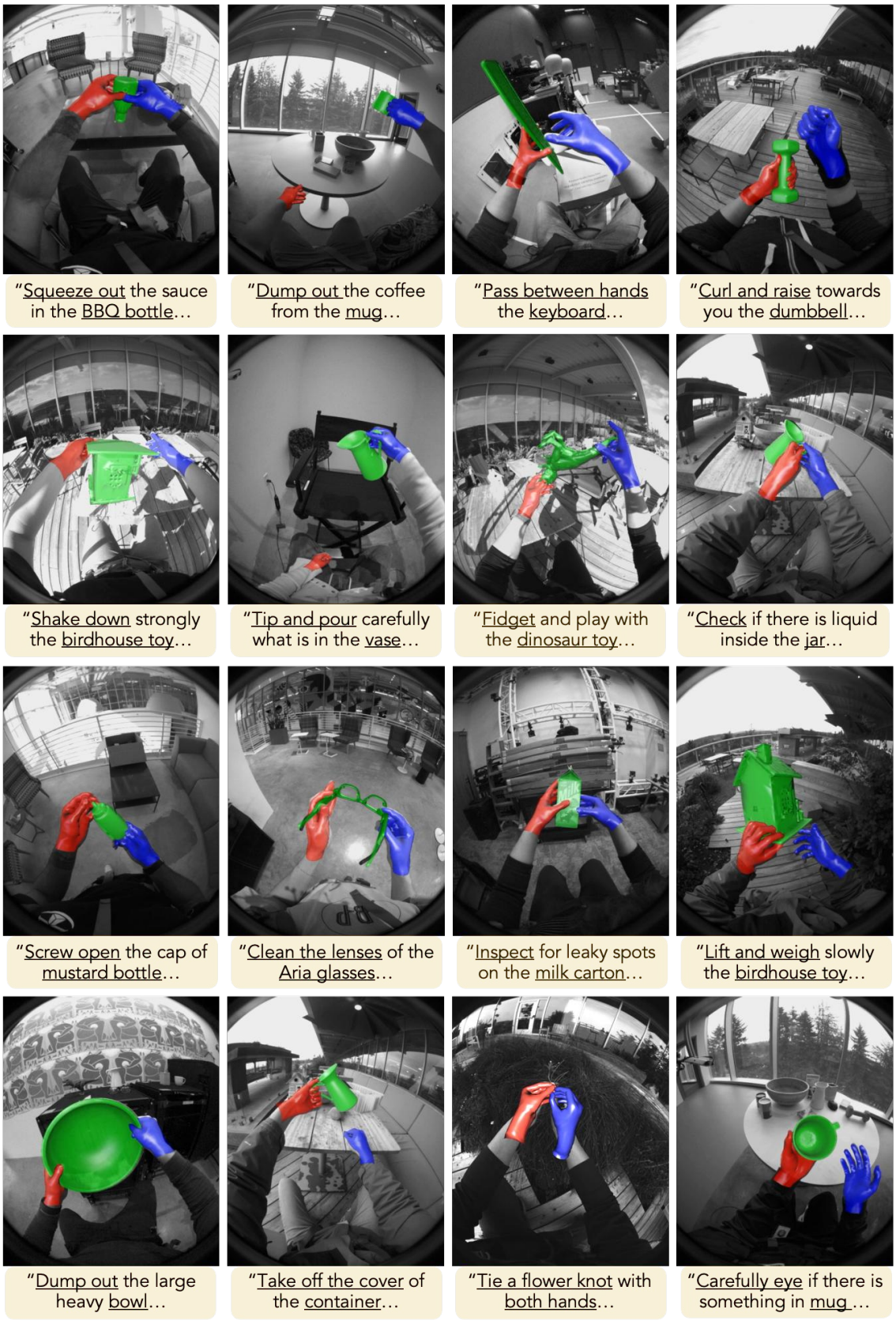


Figure 13. Additional ground truth examples from SHOW3D: hand pose (red and blue), object pose (green), and text captions.

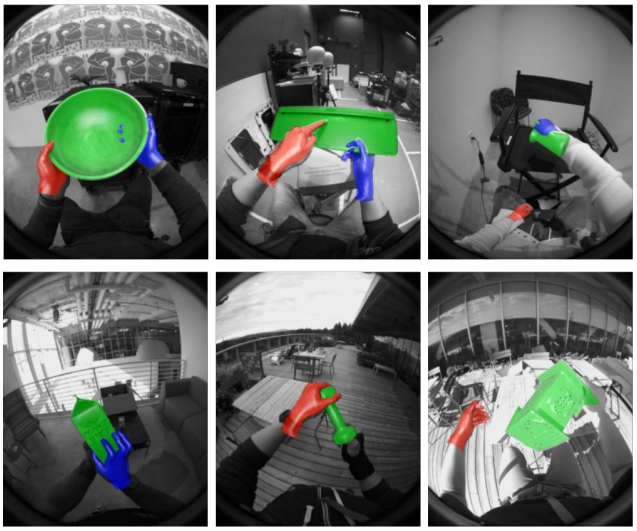


Figure 14. **Hand annotation failure cases.** We visualize two common error modes in our automated annotation pipeline. **Top row:** Examples of **inaccurate 3D pose estimation**. While all entities are successfully detected, minor errors in the estimated 6DoF pose result in physically implausible interpenetrations (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.

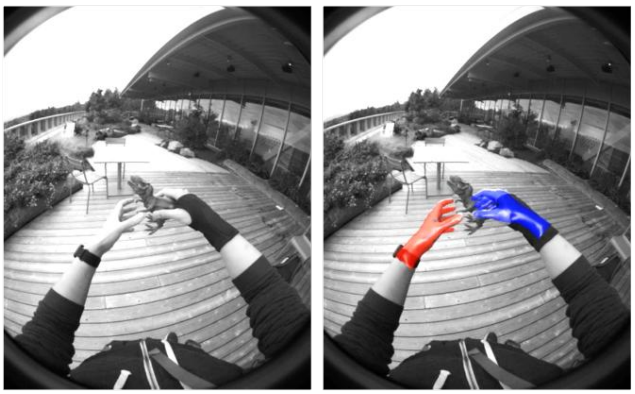


Figure 15. **Object annotation failure case:** Missed object under severe occlusion and outdoor illumination. **Left:** The raw image from egocentric view. **Right:** The projected 3D hand meshes (red/blue) overlaying the image. Although the hands are correctly tracked, the object is not detected due 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 13. 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, Figure 14 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.

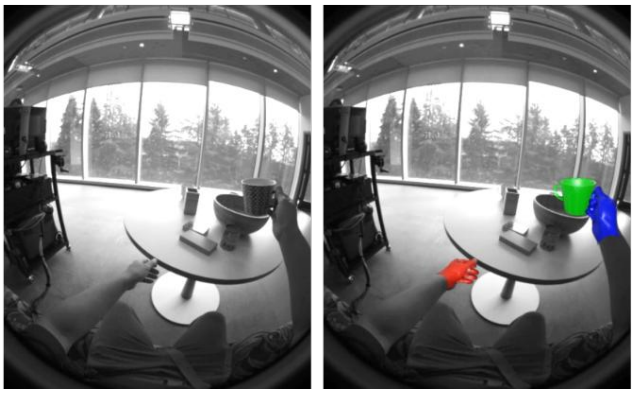


Figure 16. **Object annotation failure case:** Rotational ambiguity in symmetric objects. **Left:** The raw image from egocentric view. **Right:** The projected 3D object mesh (green) overlaying the image. Note that the pipeline correctly tracks the 3D position but fails to resolve the azimuthal rotation, incorrectly predicting the handle to be visible ("hallucinated") despite not being present in the image.

We also observe specific failure modes of our object tracking pipeline. In Figure 15, 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 [27] 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 16, 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 [24, 25, 28] 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.