GROUNDHAND: RETHINKING EGOCENTRIC MANIPULATION DATA COLLECTION WITH HIGH-PRECISION HAND ANNOTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric dexterous manipulation data are increasingly used for robot manipulation learning, yet existing data collection pipelines often cannot provide high-precision hand state annotations with consistent absolute physical scale. Common approaches first estimate hand poses from egocentric videos, recover ego-camera poses separately, and then transform the estimated hand states into a world coordinate system. Errors can accumulate throughout this pipeline, rendering the resulting annotations unreliable for fine-grained, dexterous manipulation. In this work, we propose GroundHand, a low-cost capture-and-annotation system that captures synchronized ego-view and wrist-view manipulation videos and produces high-precision MANO hand states directly in a unified workspace coordinate system. GroundHand uses a calibrated multi-view RGB-D setup with an exo-view annotation hardware cost of approximately $3,000, substantially lower than Vicon-based MoCap systems. Built upon this system, we develop a multi-source hand annotation model to estimate fine-grained hand pose and global translation. To continuously improve automatic annotation quality, we further introduce iterative human-in-the-loop model refinement, where sparse manual corrections are fed back during training to progressively refine the model. Experiments on existing benchmarks and our manually validated data show that GroundHand improves hand state accuracy and annotation efficiency over baseline methods.

1 INTRODUCTION

Learning from human egocentric manipulation data has been actively explored in embodied intelligence, and has demonstrated strong effectiveness as a prior for robotic manipulation (Kareer et al., 2025; Yang et al., 2025b; Punamiya et al., 2025). Beyond readily available egocentric videos, robot manipulation learning also requires precise hand state annotations with consistent absolute physical scale, especially for high-DoF dexterous hands and fine-grained manipulation tasks. In such settings, both subtle finger articulation and global hand motion determine the precision and success of the action. However, obtaining precise hand states from egocentric videos remains challenging, as severe occlusion makes fine finger articulation difficult to recover (Pavlakos et al., 2024; Potamias et al., 2025), and dynamic cameras provide limited geometric constraints for consistent hand reconstruction with absolute scale (Zhang et al., 2025; Sun et al., 2026).

To collect large-scale egocentric manipulation data with hand state annotations, existing studies have explored several strategies, yet reliable annotation accuracy remains difficult to guarantee. A common low-cost strategy is to first record egocentric videos with head-mounted cameras (Grauman et al., 2022; 2024; Damen et al., 2021), and then apply hand pose estimation methods to reconstruct hand states in post-processing (Zheng et al., 2026; Luo et al., 2025; 2026). This staged pipeline is convenient and scalable, and can even be applied to existing egocentric video datasets. However, its annotation quality is bounded by the reliability of monocular hand reconstruction. Moreover, because many egocentric videos lack accurate camera calibration and stable camera localization, transforming frame-wise hand estimates into temporally consistent hand states with absolute physical scale is inherently difficult, as illustrated in the first case of Figure 1. Another line of work relies on head-mounted devices with built-in hand tracking, such as Meta Quest (Meta, 2026), Apple Vision Pro (Apple, 2026), and PiCO devices (PICO XR, 2022). These systems enable flexible data collection,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

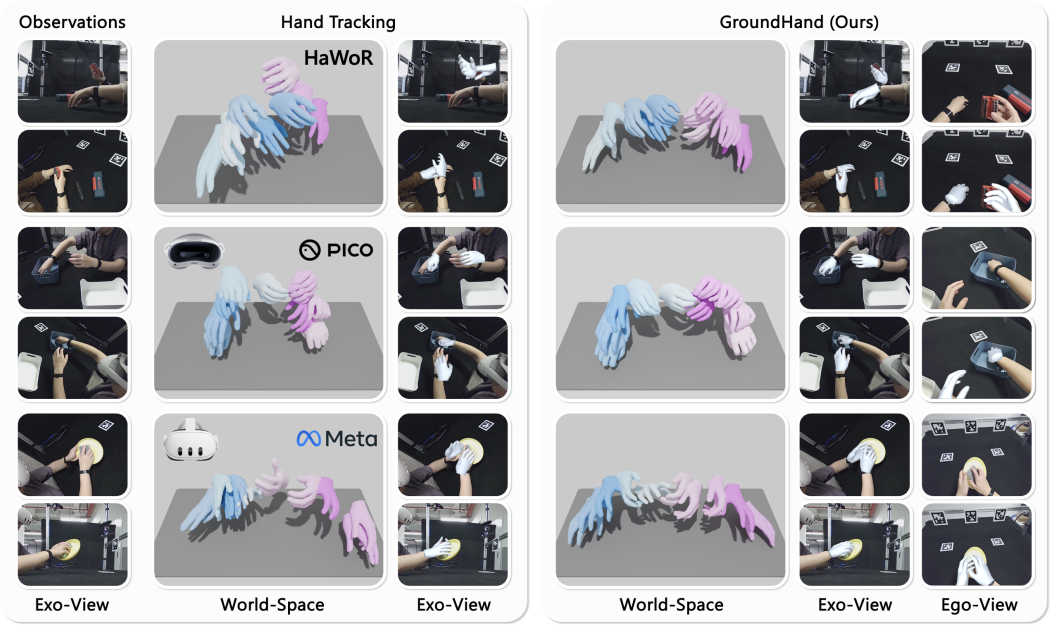

Figure 1: Existing egocentric and device-based annotation pipelines struggle to reliably reconstruct hand state. We compare representative baselines with GroundHand under challenging dexterous manipulation scenarios. For each method, we visualize the reconstructed hand motion in the unified workspace coordinate system, together with hand renderings in exo-view and ego-view images.

but the precision of hand states can still degrade under occlusion, rapid motion, and fine-grained hand-object interaction, as shown in the latter cases of Figure 1. Motion capture systems based on Vicon can provide accurate measurements, but they require constrained environments, extensive calibration, and extremely high deployment cost. For example, ARCTIC (Fan et al., 2023) relies on dense marker-based MoCap setups with 54 Vicon cameras, costing approximately 3 million dollars, which limits their scalability for egocentric manipulation data collection.

In this work, we rethink the collection of egocentric dexterous manipulation data from a systems perspective. To this end, we decouple policy observation collection from high-fidelity hand-state annotation. The ego-view and wrist-view cameras record manipulation videos for downstream policies, while calibrated exo-view RGB-D cameras provide multi-view geometric evidence for reconstructing hand states directly in a unified workspace coordinate system. Egocentric camera poses are estimated independently using AprilTag-based localization (Olson, 2011), allowing egocentric observations to be paired with accurate hand states without relying on fragile ego-view hand reconstruction. Compared with Vicon-based MoCap systems, our multi-view RGB-D setup provides a low-cost alternative for high-precision annotation, with the exo-view annotation hardware costing approximately $3,000. As a result, our system produces synchronized ego-view and wrist-view manipulation data paired with high-precision MANO hand-state annotations that preserve consistent absolute physical scale.

Built upon this capture system, we develop an annotation model for high-precision hand reconstruction to enable semi-automatic labeling. For each subject, we first optimize a subject-specific MANO shape (Romero et al., 2017), providing a hand mesh prior with accurate physical scale. Given synchronized calibrated exo-view RGB-D observations, the model fuses hand-centric image features and point-cloud features, together with the subject-specific hand mesh prior and auxiliary wrist-view observations, to estimate hand pose and global translation in the calibrated world coordinate system. RGB-D point clouds provide direct 3D geometric evidence for global hand localization with absolute physical scale, while temporal modeling helps recover occluded articulations and improves consistency across frames.

To enable continuous improvement of the automatic annotation system, inspired by SAM (Kirillov et al., 2023), we propose iterative human-in-the-loop model refinement. In each refinement stage, the current model produces initial predictions, annotators perform sparse manual corrections on their

multi-view 2D projections, and the corrected samples are further optimized and used to refine the model. Human involvement is therefore amortized into the training stage, while the final system performs fully automatic annotation after multiple refinement iterations.

We evaluate our system along multiple dimensions using both existing benchmarks and our manually validated data. We compare GroundHand with existing hand annotation methods and evaluate hand pose accuracy in the calibrated world coordinate system. We further assess geometric consistency by measuring the agreement between reconstructed MANO meshes and observed RGB-D point clouds. Our experiments show that the proposed capture-and-annotation system produces more accurate and stable hand-state annotations than existing baselines while maintaining substantially lower hardware costs than marker-based MoCap systems. Our contributions are summarized as follows:

- We build a synchronized capture system for egocentric dexterous manipulation data, pairing ego-view and wrist-view observations with high-precision MANO hand states in a unified world coordinate system.
- We develop a multi-source hand annotation model to reconstruct the hand with consistent absolute physical scale. We also utilize iterative human-in-the-loop model refinement, which progressively improves annotation accuracy, enabling automatic annotation after refinement.
- We validate the system through benchmark evaluation, manually annotated validation, physical point-cloud consistency, and efficiency analysis, demonstrating that our framework provides accurate hand-state annotations in a low-cost, scalable capture setup.

2 Related Work

Egocentric Manipulation Data Collection. Existing large-scale egocentric datasets such as EgoVerse (Punamiya et al., 2025), Ego4D (Grauman et al., 2022), EPIC-KITCHENS (Damen et al., 2021), EgoDex (Hoque et al., 2026), and HOI4D (Liu et al., 2022b) are often collected with a single wearable camera, such as a head-mounted, chest-mounted, or AR/VR headset camera. Motion-capture-based (Mocap) systems provide reliable annotations by directly measuring hand, object, and camera trajectories with high precision, which have been widely used for high-quality hand-object reconstruction benchmarks (Fan et al., 2023; Zhan et al., 2024; Banerjee et al., 2025; Garcia-Hernando et al., 2018). Close-range single egocentric view frequently suffers from occlusion and motion blur, while the dependence on expensive tracking infrastructure limits the scalability of Mocap systems. Given that situation, a increasing number of data capture systems choose to perform a trade-off between accuracy and affordance by introducing multi-view setting and depth (Grauman et al., 2024; Zhan et al., 2024; Kwon et al., 2021; Wang et al., 2024; Yang et al., 2022) as geometric priors. Inspired by these works, in our system, we use calibrated multi-view RGB-D cameras as the geometric backbone, and further combine them with an egocentric RGB-D camera and wrist-mounted fisheye cameras, yielding rich spatial context and improved visibility of contact-rich interactions.

Hand Reconstruction and Annotation. Commercial egocentric devices can provide hand annotations directly through built-in hand tracking (Meta, 2026; Apple, 2026), offering an integrated solution for device-centric data collection. Recent monocular reconstruction methods (Zhang et al., 2025; Yu et al., 2025) estimate hand motion and camera trajectory from videos, enabling automatic pseudo-label generation at scale. Mocap datasets like ARCTIC (Fan et al., 2023) and HOT3D (Banerjee et al., 2025) use marker-based infrared motion-capture system to obtain precise spatial positions. Egocentric hand reconstruction datasets like AssemblyHands (Ohkawa et al., 2023), rely on multi-view exocentric 2D annotations and optimization to recover hand trajectories. HO-Cap (Wang et al., 2024) further emphasizes a more automated annotation pipeline, while H2O (Kwon et al., 2021) augments optimization-based annotation with additional human quality inspection to filter out low-quality results. Following this line of vision-based annotation, we adopt a human-in-the-loop annotation model built upon synchronized multimodal observations. The optimized annotations can be fed back into model training, which demonstrates strong potential for scalable data annotation.

Learning from Egocentric Manipulation Videos. Egocentric human videos provide a scalable source of physically grounded knowledge for robot learning, as they capture how humans interact with objects, decompose long-horizon tasks, and execute dexterous manipulation in natural environments. Prior work (Nair et al., 2023; Ma et al., 2023; Wang et al., 2023) has shown that such videos can

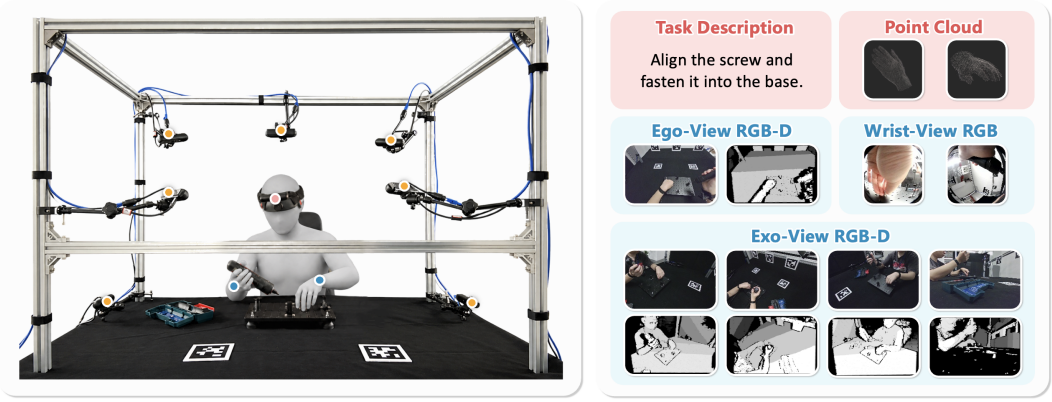

Figure 2: Overview of our data capture system setup (left) and modalities (right).

be transformed into reusable manipulation representations, visual rewards, and action priors for imitation learning and policy learning. Their value is further strengthened when paired with accurate global motion annotations and embodiment-aware supervision, which help bridge the gap between human demonstrations and robot control (Kareer et al., 2025; Yang et al., 2025b). Motivated by these findings, our capture system augments egocentric manipulation videos with synchronized multi-view RGB-D observations, bilateral wrist-mounted fisheye views, and structured language annotations, supporting geometrically grounded hand-object reconstruction and provide richer supervision for robot manipulation learning.

3 Capture System

3.1 Hardware Setup

Egocentric manipulation data collection is often affected by object occlusions and hand self-occlusions. In contrast, surrounding multi-view observations and close-range contact observations can effectively mitigate the impact of occlusions on pose estimation, providing stronger geometric priors and spatial awareness for egocentric perception.

Our data collection system, as shown in Figure 2, consists of eight Orbbec 336L RGB-D cameras capturing RGB and depth images at 1280×800 at 30 FPS. One RGB-D camera is mounted on the head to capture the egocentric view and active visual perception during manipulation, while the remaining seven cameras are arranged around the workspace to provide exocentric observations for accurate global hand annotation. This layout reduces blind spots and mitigates occlusion in the annotation pipeline. In addition, two fisheye cameras are mounted beneath the wrists, capturing 180-degree field-of-view RGB images at 1280×720 at 60 FPS. They provide close-range contact observations, further complementing the exocentric views in occlusion-heavy interactions.

For geometric calibration, we pairwise calibrate the exocentric cameras with a checkerboard, and express all extrinsics in the first camera's coordinate frame for unified coordinate transformation and point cloud composition. We further define a tabletop-aligned world frame whose xy-plane is parallel to the table, providing an intuitive reference for motion description. We estimate the egocentric camera pose using five AprilTags distributed above and below the workspace to improve visibility under head motion and occlusion, followed by outlier removal and temporal smoothing in post-processing.

3.2 Modalities

Depth information is non-trivial for metric-scale motion perception. With accurately annotated camera extrinsics and hand geometry, we can synthesize high-fidelity hand point clouds that directly reflect the hand's true pose and motion. As shown in Figure 2, our system records synchronized modalities from seven exocentric RGB-D cameras, one egocentric RGB-D camera, and two wrist-

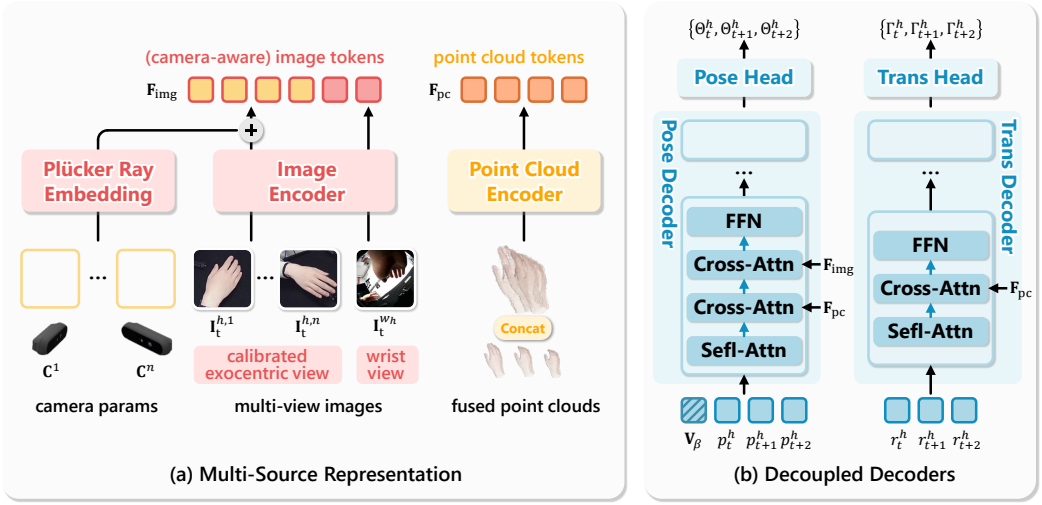

Figure 3: **Architecture of GroundHand**. The model first extracts multi-source representations, including camera-aware image representations, point-cloud tokens, and subject-specific mesh-prior tokens. It then reconstructs hand pose and root translation through decoupled decoders.

mounted fisheye RGB cameras, covering global geometry, first-person perception, and close-range hand-object interaction. The exocentric views provide dense 3D evidence for hand motion and interaction geometry, the egocentric view reflects the demonstrator's active perception and task intent, and the wrist views capture fine-grained contact cues around the hands. In addition to raw sensor streams, we provide 2D hand keypoints and bounding boxes in exocentric views, SAM3-based hand masks, and hand point clouds computed from aligned RGB-D observations and segmentation masks, offering hand-centric geometric cues for downstream annotation and optimization. Each episode is further paired with a task name and detailed language description, linking observed motions with explicit manipulation goals. Together, these modalities represent both the physical execution and semantic intent of human manipulation.

4 Multi-Source Hand Reconstruction Model

4.1 Overview

A scalable egocentric dexterous manipulation collection system requires a reconstruction model that can automatically produce accurate hand states with consistent absolute physical scale. We introduce GroundHand, a temporal reconstruction model that predicts high-precision hand states from synchronized multi-source evidence. We parameterize the hand state using the MANO model (Romero et al., 2017). Specifically, each hand state x_t^h of hand side $h \in \{L, R\}$ at frame t is represented as $x_t^h = \{\Theta_t^h, \Gamma_t^h, \beta\}$, where $\Theta_t^h \in \mathbb{R}^{16\times 6}$ denotes the hand pose in 6D rotation representation, including the global wrist orientation and articulated joint rotations, $\Gamma_t^h \in \mathbb{R}^3$ denotes the root translation, and $\beta \in \mathbb{R}^{10}$ denotes the shape parameter. The shape parameter β is subject-specific and optimized once for each subject from a dedicated calibration sequence. We define a canonical workspace coordinate system relative to the operator position and the tabletop. The camera extrinsics, point clouds, and MANO parameters are unified in this coordinate space, promising the consistent absolute physical scale.

We introduce the GroundHand framework in Section 4.2. The model takes synchronized multi-source observations and cues from the capture system, and predicts the hand state sequence $\{\Theta_t^h, \Gamma_t^h\}_{t=1}^T$ in a single feed-forward pass. To further improve reconstruction accuracy, we design an iterative human-in-the-loop refinement in Section 4.3, in which sparse manual corrections are applied during training to refine the model while preserving fully automatic annotation.

4.2 FRAMEWORK

Camera-Aware Image Representation.

As illustrated in Figure 3, multi-view RGB observations provide appearance cues from calibrated viewpoints and reduce ambiguity due to single-view occlusion. For each exo-view camera n with calibrated parameters $\mathbf{C}^n$, we crop the hand region from the RGB image using the prepared hand bounding box, obtaining a hand-centric crop $\mathbf{I}_t^{h,n}$. These hand-centric inputs focus the model on the manipulation region and reduce irrelevant background variation. We use the pretrained ViT backbone from HaMeR (Pavlakos et al., 2024) as the image encoder to extract visual tokens $\mathbf{F}_{\text{img}}$. To make multi-view image features aware of camera geometry, we encode the calibrated camera parameters using Plücker rays (Plucker, 1865; Jin et al., 2024). This camera-aware representation allows GroundHand to reason over observations from different calibrated viewpoints and handle a variable number of input views without assuming a fixed camera layout, thereby supporting extension to different capture setups. The wrist-view RGB image provides close-range visual evidence for subtle finger configurations that may be difficult to observe from calibrated cameras. These wrist-view tokens are concatenated with exo-view image tokens as an auxiliary local articulation cue.

Point Cloud Representation. RGB-D observations from fixed views are fused into a hand-centric point cloud, which provides direct 3D geometric evidence for global hand localization with consistent absolute physical scale. For each calibrated RGB-D camera, we first align the RGB image and depth map, apply the hand mask to filter valid hand-region pixels, and back-project the remaining pixels into 3D using the camera intrinsics. The points are then transformed into the canonical workspace coordinate system using the camera extrinsics and fused across all fixed views. We downsample the fused point cloud and encode it with a point-cloud encoder (Wu et al., 2025), which takes 3D coordinates and RGB values as input and produces geometry-aware point tokens F_{pc}.

Subject-Specific Hand Mesh Prior. The subject-specific MANO mesh under an average pose is further used as a hand-shape prior to reduce ambiguity between shape, pose, and scale. Hand shape varies across subjects and directly affects the physical scale of the reconstructed hand. Therefore, for each subject, we estimate fixed MANO shape parameters from RGB-D point clouds before sequence reconstruction. Using these shape parameters, we generate MANO meshes under an average hand pose, denoted as $\mathbf{V}_\beta$. The mesh vertices are injected into the learnable pose tokens.

Decoupled Decoders. Given the encoded multi-source tokens, GroundHand predicts the pose and root translation of both hands for all frames in a single feed-forward pass. The reconstruction decoders use bidirectional temporal attention, allowing each frame to aggregate information from nearby frames. We adopt decoupled pose and translation decoders because fine-grained articulation and global localization rely on different sources of evidence. The pose decoder focuses on local articulation cues from calibrated multi-view image tokens, wrist-view tokens, point-cloud tokens, and subject-specific mesh priors. The translation decoder focuses more on RGB-D point-cloud geometry, which provides direct evidence for root localization with consistent absolute physical scale. This decoupled design allows GroundHand to model local hand articulation and global root translation separately, while still sharing complementary evidence throughout the reconstruction process.

4.3 ITERATIVE HUMAN-IN-THE-LOOP REFINEMENT

To enable GroundHand to perform high-precision automatic annotation, we need a sufficiently large amount of accurately annotated training data. However, manually annotating MANO parameters is prohibitively expensive. Inspired by the model-assisted data-engine strategy of SAM (Kirillov et al., 2023), we refine GroundHand through an iterative human-in-the-loop process. In each refinement stage, the current model predicts initial hand-state annotations, annotators correct sparse errors in the multi-view 2D projections, and the optimized annotations are fed back to further improve the model. Human involvement is therefore concentrated in the training stage, while the refined model performs fully automatic annotation after training.

We first pretrain GroundHand on DexYCB (Chao et al., 2021), which provides the supervision our model requires, including synchronized RGB-D observations, camera calibration, and MANO hand parameters. This pretraining stage gives GroundHand an initial ability to reconstruct hand states from calibrated multi-source evidence. For each refinement stage, the current GroundHand first predicts

Table 1: Comparison with existing annotation methods on DexYCB.

Method	MPJPE ↓	RR-MPJPE ↓	PA-MPJPE ↓	MPVPE ↓	RR-MPVPE ↓	PA-MPVPE ↓
Single View Hand Mesh Reconstruction Method						
HandOccNet Park et al. (2022)	14.0	-	5.8	13.1	-	5.5
MobRecon Chen et al. (2022)	14.2	-	6.4	13.1	-	5.6
H2ONet Xu et al. (2023)	14.0	-	5.7	13.0	-	5.5
Zhou *et al.* Zhou et al. (2024)	12.4	-	5.5	12.1	-	5.5
MaskHand Saleem et al. (2025)	11.7	-	5.0	11.2	-	4.9
Multiple View Hand Mesh Reconstruction Method						
FTL-MeshTR Remelli et al. (2020)	8.7	9.8	5.5	8.8	9.8	5.8
PE-MeshTR Liu et al. (2022a)	7.5	8.9	4.8	7.4	8.7	4.7
POEM-small Yang et al. (2025a)	7.2	8.9	4.2	7.5	8.8	4.5
POEM-large Yang et al. (2025a)	6.3	7.5	3.7	6.5	7.5	3.8
GroundHand (pretrained)	6.3	6.0	3.5	6.2	5.6	3.1
GroundHand (ours)	6.1	5.6	3.4	6.1	5.3	3.1

Table 2: Hand annotation quality on the evaluation subset of our captured data.

Annotation Method	PC2Mesh ↓	PC2Mesh-Q90 ↓	Mesh2PC ↓	Mesh2PC-Q90 ↓	*Average* ↓
WiLoR Potamias et al. (2025)	106.357	155.885	153.415	237.660	129.886
HaWoR Zhang et al. (2025)	60.488	96.608	53.283	83.905	56.886
PiCO Tracking PICO XR (2022)	23.353	43.847	21.233	42.517	22.293
Meta Quest Tracking Meta (2026)	15.960	32.340	18.260	39.146	17.110
GroundHand (pretrained)	9.256	16.605	9.834	19.997	9.545
GroundHand (ours)	6.479	10.719	6.126	13.683	6.302
Human Correction and Optimization	6.334	10.493	5.930	13.449	6.132

initial MANO hand states for the captured sequences. These predictions are used as initialization for the manual correction process. We project the predicted 3D hand keypoints onto calibrated exo-view images, and annotators inspect the multi-view 2D projections and correct inaccurate keypoints when necessary. Given these corrected 2D keypoints, we perform multi-view optimization to update the MANO pose and root translation. The resulting optimized MANO parameters are treated as human-verified refined annotations and used to fine-tune GroundHand for the next refinement stage. In our implementation, we perform two refinement stages on the captured data. This iterative process gradually improves annotation fidelity and enables the final GroundHand model to automatically annotate new sequences.

5 Experiments

We evaluate GroundHand from multiple perspectives centered on **whether it can serve as a reliable automatic hand-state annotator for egocentric dexterous manipulation data**. In Section 5.2, we compare GroundHand with existing hand reconstruction methods to assess its ability to recover hand states with consistent absolute physical scale. In Section 5.3, we further evaluate the quality of the annotations on our captured data through geometric consistency with RGB-D point clouds. We also analyze the iterative human-in-the-loop refinement process and ablate the major components of our multi-source hand reconstruction model in Section 5.4 to understand how each design choice contributes to annotation accuracy and accelerates subsequent refinement.

5.1 Experimental Setup

Datasets. We conduct experiments on both public hand pose benchmarks and our captured dexterous manipulation data. For public benchmark evaluation, we use DexYCB (Chao et al., 2021), which provides synchronized multi-view RGB-D observations, calibrated cameras, and MANO annotations. In addition, to evaluate the annotation quality of our collection system, we capture an extra evaluation subset. During the capture of this subset, the operator also wears a Meta Quest or PiCO device, enabling comparison with device-based hand tracking under the same manipulation sequences.

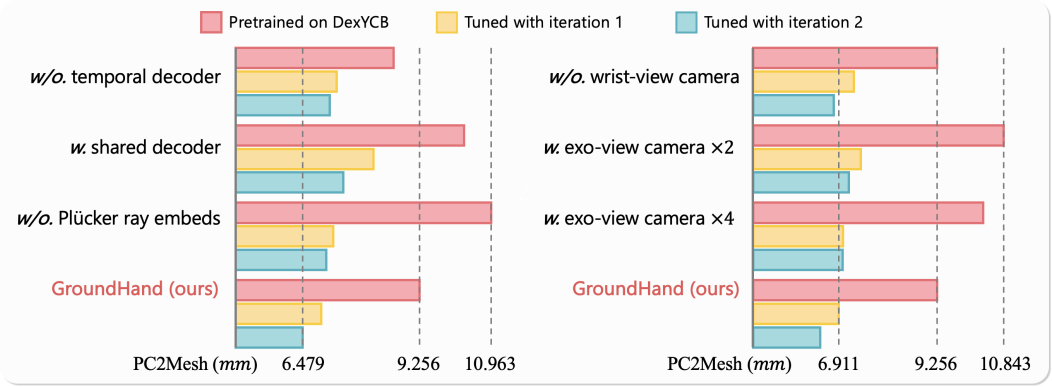

Figure 4: **Ablation of model components and iterative refinement.** We report PC2Mesh error under different model variants and camera configurations.

Metrics. On public hand pose benchmarks, we report mean per-joint position error (MPJPE) and mean per-vertex position error (MPVPE). Following common practice, we evaluate them under two reference systems: root-relative (RR), which removes the global translation and focuses on hand articulation, and Procrustes-aligned (PA), which further aligns the prediction to the ground truth with a similarity transformation. To evaluate the annotation quality of our captured data, we measure 3D geometric consistency between the reconstructed MANO mesh and the observed RGB-D hand point cloud. Since RGB-D point clouds only cover visible hand surfaces, we do not treat them as complete ground truth. Instead, we report directional point-cloud/mesh distances. PC2Mesh measures the distance from observed RGB-D points to the nearest reconstructed mesh surface, while Mesh2PC measures the distance from mesh vertices to the nearest observed point. We further report the 90th percentile errors, PC2Mesh-Q90 and Mesh2PC-Q90, to reflect tail misalignment. The Average metric is the mean of PC2Mesh and Mesh2PC. All distances are reported in millimeters.

5.2 Comparison with Hand Reconstruction Methods

We first evaluate GroundHand on DexYCB to validate its reconstruction accuracy under standard 3D supervision. As shown in Table 1, GroundHand achieves competitive hand pose and mesh reconstruction performance compared with existing single-view and multi-view hand mesh reconstruction methods. The results suggest that calibrated multi-source evidence enables accurate MANO reconstruction, even before refinement on our captured data.

We further compare GroundHand with ego-centric and device-based annotation pipelines in our capture setting. Figure 1 shows representative results by visualizing hand renderings in ego-view and exo-views and reconstructed hand motion in the workspace. Existing pipelines can suffer from inaccurate 3D localization and inconsistent absolute physical scale, while GroundHand produces hand annotations that are better aligned with RGB-D geometry and visual observations.

5.3 Annotation Quality on Captured Data

We further evaluate the annotation quality of different pipelines on our captured dexterous manipulation data. We capture a dedicated evaluation subset and process the same sequences with representative ego-view reconstruction methods, device-based tracking systems, GroundHand, and human-corrected optimization. WiLoR (Potamias et al., 2025) and HaWoR (Zhang et al., 2025) estimate hand states from ego-view images, and their predictions are transformed into the canonical workspace coordinate system using the ego-camera pose. PiCO Tracking and Meta Quest Tracking represent device-based baselines recorded during the same capture sessions and aligned to the same coordinate system. Table 2 reports geometric consistency between the reconstructed MANO mesh and the observed RGB-D hand point cloud. GroundHand substantially improves geometric consistency over ego-view reconstruction and device-based tracking pipelines, and closely approaches the quality of human-corrected optimization.

Table 3: Effect of iterative human-in-the-loop refinement.

Iteration	Auto Annotation				Human Correction	
	Model Version	PC2Mesh (mm)	Mesh2PC (mm)	Average	Correction Ratio (%)	Time Cost (s/frame)
1	pretrained	9.256	9.834	9.545	48.1	16.9
2	tuned with it.1	6.911	7.552	7.232	14.8	4.1
3	tuned with it.2	6.479	6.126	6.302	-	-

5.4 Effect of Human-in-the-Loop Refinement

We analyze the effect of iterative human-in-the-loop refinement in two ways: whether the model design yields better automatic annotations, and whether the refinement loop reduces the human correction required in later stages. Figure 4 reports the PC2Mesh error of different model variants across three training stages: DexYCB pretraining, tuning with the first refinement iteration, and tuning with the second refinement iteration. The results show that each component contributes to annotation quality. Removing temporal decoding, Plücker ray embeddings, wrist-view observations, or using a shared decoder consistently degrades geometric alignment. Varying the number of exo-view cameras further shows that additional calibrated views improve reconstruction quality, while GroundHand remains effective under different camera configurations.

Beyond component ablation, Table 3 evaluates how iterative refinement changes both automatic annotation quality and human correction effort. Each iteration uses the model refined from the previous stage as the pre-annotation model for the next annotation round. We report PC2Mesh, Mesh2PC, and their average distance to measure annotation quality. For human correction, we report the correction ratio and annotation time. The correction ratio denotes the percentage of projected 2D keypoints that require manual correction among all visible keypoints that annotators inspect. A lower correction ratio, therefore, indicates that the model produces more accurate initial projections and requires less human intervention. The geometric error decreases steadily across iterations, while the correction ratio and annotation time also drop substantially. This demonstrates that human corrections are progressively absorbed into the model: early manual feedback improves the pre-annotation model, and the improved model reduces the correction burden in later annotation rounds.

6 Conclusion

This work rethinks egocentric dexterous manipulation data collection from the perspective of annotation reliability. While ego-view videos are easy to collect and useful for robot learning, they are not by themselves sufficient for producing accurate hand states with consistent absolute physical scale. GroundHand addresses this gap by separating observation collection from high-fidelity annotation: ego-view and wrist-view cameras capture the visual data needed by downstream tasks, while calibrated exo-view RGB-D cameras provide the geometric evidence for accurate MANO hand-state reconstruction in a canonical coordinate system.

Building on this design, we develop a multi-source temporal reconstruction model that combines camera-aware exo-view image features, RGB-D point-cloud geometry, subject-specific hand mesh priors, and wrist-view articulation cues. To progressively improve the model's annotation capability, we introduce iterative human-in-the-loop refinement, where sparse manual corrections are fed back to refine GroundHand during training. Experiments show that GroundHand produces more reliable annotations than existing ego-centric and device-based pipelines, and significantly improves geometric consistency with RGB-D observations.

Scalability and Extensibility. GroundHand is designed as an extensible collection framework rather than a specific camera configuration. The reconstruction model only requires calibrated exo-view RGB-D observations and can therefore support different camera numbers, camera placements, and capture environments. The same design naturally allows additional sensing modalities to be synchronized with the system and paired with high-precision hand-state annotations. In ongoing collection, we are scaling GroundHand to more diverse subjects, objects, and long-horizon dexterous tasks, and plan to incorporate tactile sensing to capture contact-rich interaction signals. We hope this work provides a step toward scalable data collection systems.

STATEMENT

AI Use Statement. Generative AI tools were used to assist with language editing, manuscript organization, and improving the clarity of presentation. The authors developed, reviewed, and verified all research ideas, experimental design, data processing, analyses, and conclusions. The authors take full responsibility for the content of this paper.

REFERENCES

Apple. Handtrackingprovider. https://developer.apple.com/documentation/arkit/handtrackingprovider/, 2026. Apple Developer Documentation. A source of live data about the position of a person's hands and hand joints. Accessed: 2026-05-04. 1, 3

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

Gary Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. 15

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6, 7, 15

Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3

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*, 2023. 2, 3

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, 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*, 2024. 1, 3

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026. 3

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024. 6

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *IEEE International Conference on Robotics and Automation*, 2025. 1, 4

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, 2022a. 7

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*, 2022b. 3

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. 1

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, et al. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026. 1

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations*, 2023. 3

Meta. Hand tracking overview for meta quest in unity. https://developers.meta.com/horizon/documentation/unity/unity-handtracking-overview/, 2026. Meta Horizon OS Developer Documentation. Hand tracking enables the use of hands as an input method for the Meta Quest headsets. Accessed: 2026-05-04. 1, 3, 7

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, Proceedings of Machine Learning Research, 2023. 3

Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assembly-Hands: towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2011. 2, 15

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7

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*, 2024. 1, 6

PICO XR. PICO 4. https://www.picoxr.com/global, 2022. VR headset. Accessed: 2026-05-07. 1, 7

Julius Plucker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, 1865. 6

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 7, 8

Ryan Punamiya, Simar Kareer, et al. Egoverse: An egocentric human dataset for robot learning from around the world, 2025. https://egoverse.ai/assets/pdf/EgoVerse.pdf. 1, 3

Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 7

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 2017. 2, 5, 15

Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. Maskhand: Generative masked modeling for robust hand mesh reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 7

Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. Unihand: A unified model for diverse controlled 4d hand motion modeling. In *International Conference on Learning Representations*, 2026. 1

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 201–221. PMLR, 2023. 3

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. *arXiv preprint arXiv:2406.06843*, 2024. 3

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22193–22204, 2025. 6

Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7

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

Lixin Yang, Licheng Zhong, Pengxiang Zhu, Xinyu Zhan, Junxiao Kong, Jian Xu, and Cewu Lu. Multi-view hand reconstruction with a point-embedded transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a. 7

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Rizhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025b. 1, 4

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

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*, 2024. 3

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3, 7, 8

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026. 1

Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7