DEPTHUMI: ACTIVE DEPTH FOR UNIVERSAL MANIPULATION INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal Manipulation Interfaces (UMIs) scale robot learning by allowing people to collect demonstrations beyond the robot workspace. Their portability, however, shifts the bottleneck to sensing: a wrist-contained device must recover precise motion despite occlusions and rapid viewpoint changes, while preserving the local 3D geometry needed by the policy. We present DEPTHUMI, a handheld interface in which the same learned active-depth source supports both motion recovery and policy observation. Rather than treating these as separate sensing problems, DEPTHUMI combines wide-field stereo, high-rate inertial sensing, and learned active depth in one calibrated assembly. Wide-field visual–inertial reconstruction estimates the demonstrated trajectory, and a calibrated transfer maps camera motion into gripper action supervision. In parallel, learned near-field geometry constrains reconstruction and supplies metric point-cloud observations for policy learning. A simulation study guides the sensing configuration, followed by real-world evaluation of the resulting system. On an instrumented validation rig evaluated against an external Vive reference, DEPTHUMI achieves an overall mean wrist-camera translation error below 3 mm without using that reference as estimator input. Across three manipulation tasks spanning articulated pulling, peg insertion, and constrained block extraction, learned geometry matches the strong baseline on the coarse pulling task and yields its clearest gains when local geometric precision is decisive. Together, these results show that a single wrist-contained sensing system can produce accurate action supervision and geometry-aware policy inputs. Project page: `https://anonymous-submission-x.github.io/DepthUMI/`.

1 INTRODUCTION

UMIs allow people to demonstrate robot skills with handheld grippers, moving data collection beyond the robot workspace (Chi et al., 2024; Zhaxizhuoma et al., 2025; Choi et al., 2026). Their value depends on diverse, reliable data (Lin et al., 2025). Transfer also requires accurate gripper motion and observations that preserve the spatial relationships behind each action (Chi et al., 2024).

Wrist motion is hardest to recover near the interaction: the hand and object occlude the scene, nearby surfaces offer few stable features, and rapid viewpoint changes disrupt correspondence (Chi et al., 2024; Wang, 2026). Fixed base stations address this problem through high-precision tracking inside an instrumented volume, as used by TacUMI and the full-featured FastUMI Pro configuration (Zhang et al., 2026; Lumos Robotics, n.d.). Head-referenced systems instead track from a wider and more stable viewpoint: ExUMI uses headset/controller tracking, while HiFi-UMI transfers a head-mounted reference to the gripper (Xu et al., 2025; Simple AI et al., 2026). Both strategies improve observability, but require infrastructure or sensing beyond a self-contained wrist device.

Motion labels alone are insufficient because policies must also localize surfaces and objects. RGB policies leave metric distance and local shape implicit, whereas depth exposes them to voxel and point-cloud policies (Chi et al., 2023; Shridhar et al., 2023; Ze et al., 2024; Wang et al., 2024). These data products are often assembled from separate pipelines. At the wrist, however, hand–object occlusion simultaneously removes visual correspondences and weakens conventional depth. A reliable near-gripper metric signal can therefore support both demonstration reconstruction and policy observation, as well as spatial data augmentation (Huang et al., 2025).

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.

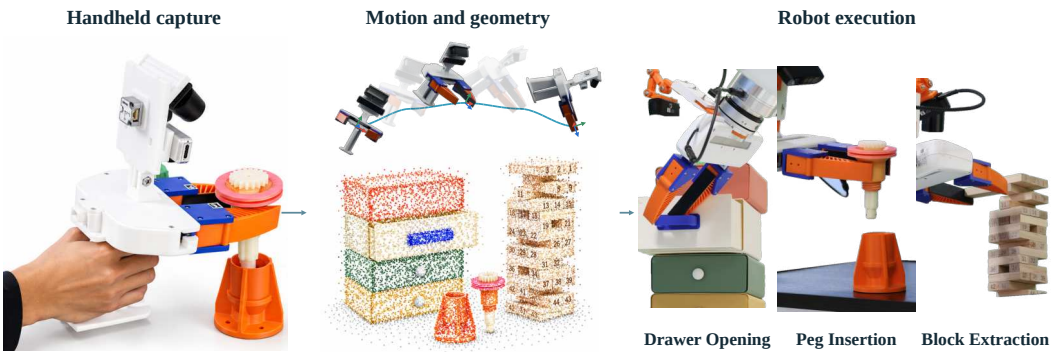

Figure 1: **DepthUMI: learned active depth for motion recovery and robot learning.** Wide-field motion cues and near-field geometry produce gripper trajectories, relative actions, and point-cloud observations for three manipulation tasks.

DEPTHUMI (Figure 1) organizes wrist-contained capture around learned metric geometry. Learned structured-light decoding recovers near-field depth (Li et al., 2025; 2026); wide-field stereo and a dedicated inertial measurement unit (IMU) preserve motion constraints through occlusion and blur. Calibration connects the recovered camera trajectory to gripper supervision and transforms the same depth into metric policy observations.

Simulation guides the sensing design across placements, tasks, and textures. Real recordings then evaluate reconstruction against an evaluation-only Vive reference, with below 3 mm overall mean camera translation absolute trajectory error (ATE). A cup study connects motion quality to action supervision, and three robot tasks compare RGB DP, DP3–Raw, and DP3–Active under common protocols. Learned geometry contributes most when local precision is decisive.

Our contributions are:

- **Wrist-contained geometric acquisition.** Active IR stereo, wide-field vision, and dedicated inertia acquire motion and geometry without base stations or a head-mounted reference.
- **Accurate learned active depth for reconstruction and learning.** Dense, reliable near-field depth strengthens reconstruction under degraded appearance and supplies metric observations for geometry-aware learning; shared calibration converts the recovered motion into gripper supervision.
- **Evidence from sensing to manipulation.** Simulation, real motion recovery, and matched DP3 comparisons trace geometry from sensing to tasks with increasing precision demands.

2 Related Work

Handheld demonstration interfaces Handheld UMI systems differ most visibly in how they recover gripper motion. The original UMI pipeline uses ORB-SLAM3 with inertial measurements, and MV-UMI and RwoR adopt the same backend for handheld pose estimation (Chi et al., 2024; Campos et al., 2021; Rayyan et al., 2025; Heng et al., 2025), making it a representative UMI-family baseline. Other systems improve observability with additional tracking hardware: FastUMI uses commercial tracking, TacUMI and the full-featured FastUMI Pro configuration use base stations, and ExUMI and HiFi-UMI introduce head-referenced sensing (Zhaxizhuoma et al., 2025; Zhang et al., 2026; Lumos Robotics, n.d.; Xu et al., 2025; Simple AI et al., 2026). UMI-3D instead develops LiDAR-centric wrist SLAM (Wang, 2026). Cloud-UMI brings point clouds into the observation stream, but obtains motion from separate tracking (Huang et al., 2025). This separation leaves motion and observation with distinct sensing assumptions. In contrast, DEPTHUMI uses one calibrated geometric system for both trajectory reconstruction and policy observation.

Active geometry and motion reconstruction Reliable depth is especially valuable when wrist imagery loses texture or stable correspondence. Visual–inertial SLAM already combines image-based

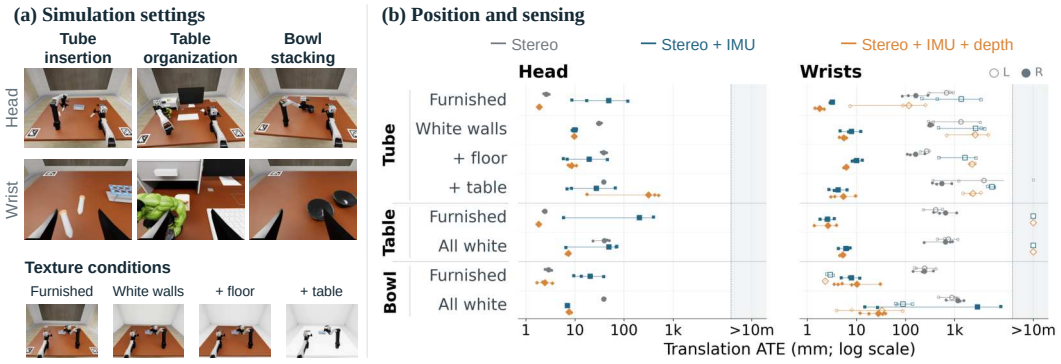

Figure 2: **Simulation-guided sensing design.** (a) RoboDojo tasks and texture regimes in Isaac Sim. (b) Camera ATE under stereo, simulator-IMU, and learned-depth combinations. Marks show all 220 solves; large marks and lines show means and min–max ($n = 3$, except the two bowl right-wrist learned-depth cells with $n = 5$).

spatial constraints with complementary high-rate inertial estimates (Leutenegger et al., 2015; Qin et al., 2018; Campos et al., 2021), and OKVIS2-X extends this formulation with dense alignment from depth or LiDAR (Boche et al., 2025). The remaining difficulty is measurement quality: active illumination creates correspondence on weak texture, yet raw factory depth often breaks down around close-range occlusion boundaries and thin structures (Zhang et al., 2018). NSL replaces hand-crafted matching with learned correspondence and geometric priors, while NSL-SLAM brings the resulting geometry into tracking (Li et al., 2025; 2026). DEPTHUMI adapts this direction to the close-range wrist regime by combining learned local depth with wide-field visual constraints and calibrated inertia.

Geometry for policy learning Policy learning has progressively moved from appearance-only inputs toward more explicit spatial reasoning. RGB methods such as Diffusion Policy infer actions directly from images (Chi et al., 2023), while auxiliary 3D objectives and depth-augmented images introduce geometry through supervision or additional channels (Zhao et al., 2026; Zhou et al., 2026; Zhaxizhuoma et al., 2025; Hentosh et al., 2026). PerAct, DP3, RISE, and 3D Diffuser Actor instead expose voxel or point-cloud geometry directly to the policy (Shridhar et al., 2023; Ze et al., 2024; Wang et al., 2024; Ke et al., 2025); iDP3 and ManiFlow extend this direction to egocentric and multimodal settings (Ze et al., 2025; Yan et al., 2025). DEPTHUMI complements these policy designs by acquiring metric observations and their action supervision together within one calibrated demonstration system.

3 THE DEPTHUMI SYSTEM

DEPTHUMI is a wrist-contained interface for collecting demonstrations with both accurate motion and metric scene geometry. The device records a wide-field visual–inertial stream for motion recovery and an active-IR stream centered on the gripper workspace. A shared calibration connects the sensing rig, handheld gripper, and robot tool frames. The resulting trajectory provides action supervision, while local depth provides the spatial observation used by the policy.

3.1 SIMULATION-GUIDED SENSING DESIGN

Before fixing the hardware configuration, we use three tasks from the RoboDojo benchmark (Chen et al., 2026) in NVIDIA Isaac Sim (NVIDIA, n.d.) to probe two design choices: where sensing should be placed and which geometric and inertial cues it should retain. This conceptual study informs the sensing architecture.

The study compares head and wrist viewpoints under changing scene texture, then adds a kinematics-derived causal IMU stream and predicted active depth to stereo tracking. The outcomes vary across tasks and placements, motivating complementary rather than interchangeable sensing.

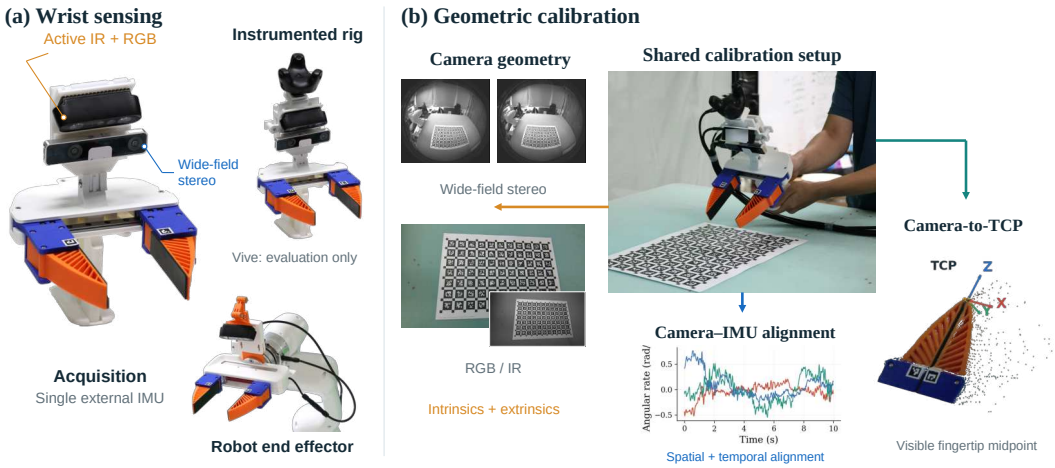

Figure 3: **Device configurations and calibration protocol.** Photographs show the compact acquisition device, instrumented validation rig, and robot-mounted sensor. Recorded calibration views support camera geometry, camera–IMU spatial and temporal alignment, and visual TCP estimation. The compact device uses one HI12 IMU; the instrumented rig records additional inertial streams and an evaluation-only Vive reference.

For the bowl right wrist, ground-truth and learned depth achieve 10.3–11.5 mm mean ATE in the furnished scene and 27.9–39.8 mm in the all-white scene, while the depth-free all-white condition contains an 8.74 m failure. Wide-field vision preserves context, inertia supports rapid local motion, and active IR supplies near-field geometry when appearance becomes unreliable. Appendix E reports the complete protocol and results.

3.2 Wrist-Contained Sensing

Guided by these observations, the sensing assembly combines modalities with complementary operating ranges. An Orbbec Gemini 335 records RGB and infrared stereo around the interaction region. An Intel RealSense T265 contributes wide-field grayscale fisheye images that retain peripheral scene structure when the hand or manipulated object occludes the near field. A dedicated IMU measures angular velocity and linear acceleration between image observations. The estimator consumes the fisheye images and calibrated inertial stream directly; the T265 proprietary pose is retained only as a system baseline.

We recover metric depth from the IR pair with a frozen learned model that adapts UFM correspondence (Zhang et al., 2025) to structured-light decoding, following NSL (Li et al., 2025). Its learned correspondence and geometric priors preserve near-field surfaces that factory decoding frequently loses around occlusion boundaries. This depth enters two parts of the system: it constrains motion reconstruction and supplies the metric XYZ observation used for policy learning. We additionally construct an offline wide-field *Fisheye depth* reference from 45 calibrated local perspective stereo pairs processed by FoundationStereo (Wen et al., 2025). This computationally intensive construction serves as a depth capability upper bound rather than a deployment pathway.

Figure 3 shows the two physical configurations. The validation rig records additional inertial streams and an evaluation-only Vive reference, which never enters the estimator. The compact acquisition device removes the reference tracker and retains a single HI12 IMU. Both configurations preserve the same camera geometry and camera-to-tool interface, while computation and power remain on a host.

3.3 Spatial and Temporal Calibration

Metric trajectories and point clouds require all observations to share consistent spatial and temporal conventions. We first calibrate the fisheye intrinsics from a wide-coverage sequence, then fix them in a common-view solve that estimates the Gemini cameras and inter-camera transforms. A motion-

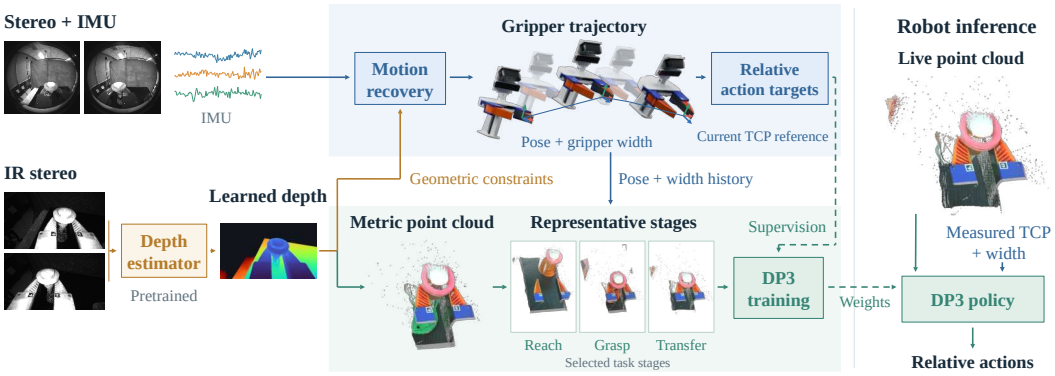

Figure 4: **Active geometry for action supervision and policy observations.** Wide-field stereo and inertia recover the demonstrated gripper trajectory, while learned active depth contributes geometric constraints and metric point-cloud observations. Calibrated poses define current-relative action targets; robot execution preserves the same observation and action conventions with measured TCP state. The cloud snapshots show representative task stages.

rich sequence subsequently identifies the camera–IMU transform and relative clock offset (Furgale et al., 2013). Device and host timestamps are preserved through processing rather than replaced by callback arrival times.

The resulting camera reprojection errors are 0.24–0.60 px RMS. An evaluation-only camera–Vive hand–eye check gives 2.990 mm and 0.477° held-out closure RMSE on 73 observations. A separate rigid transfer maps recovered camera motion to the handheld gripper, and the robot uses its calibrated end-effector-to-TCP transform during execution. These checks keep camera geometry, reference evaluation, and physical tool transfer as distinct quantities. Appendix B reports the calibration procedure and reference checks.

4 From Handheld Demonstrations to Robot Policies

The calibrated sensor record is processed along two connected paths (Figure 4). Wide-field vision, inertia, and local depth recover the demonstrated gripper trajectory, from which we derive relative action targets. The same learned depth source produces metric observations for policy training and execution. The paths share calibrated geometry, but retain separate representations suited to estimation and control.

4.1 Depth-Aided Motion Reconstruction

Our reconstruction backend builds on the tightly coupled visual–inertial and volumetric alignment formulation of OKVIS2-X (Boche et al., 2025). Let W be the estimator world frame, S the IMU frame, F_l, F_r the wide-field fisheye frames, and L the pinhole left-IR frame; T_B^A maps from B to A. Each keyframe state contains $T_{S_k}^W$, velocity, and inertial biases. Calibrated $T_{F_l}^S, T_{F_r}^S$ define stereo and temporal reprojection factors, IMU preintegration couples consecutive states, and T_L^S registers learned depth without resampling it into a fisheye image.

Let $D_b(\mathbf{u})$ be optical-axis depth at a valid pixel $\mathbf{u}$ in keyframe b, let $\tilde{\mathbf{u}}$ be its homogeneous coordinate, and let K_L be the left-IR intrinsics. Its back-projection is $\mathbf{p}_b^L = D_b(\mathbf{u}) K_L^{-1} \tilde{\mathbf{u}}$. For a submap anchored at keyframe a, this point becomes

$$\mathbf{q}_a = (T_{S_a}^W)^{-1} T_{S_b}^W T_L^S \, \mathbf{p}_b^L. \tag{1}$$

The anchored submap stores a differentiable signed log-odds field $f_a : \mathbb{R}^3 \to \mathbb{R}$ whose zero level is the local surface. We denote depth uncertainty by σ_D and local map uncertainty by $\sigma_{\text{map}} =$

$|\ell_{\min}|/(3\|\nabla f_a(\mathbf{q}_a)\|)$, where $\ell_{\min}$ is the lower log-odds parameter. The normalized residual is

$$r_{g,ab} = \frac{f_a(\mathbf{q}_a)}{\|\nabla f_a(\mathbf{q}_a)\|\sqrt{\sigma_{\text{map}}^2 + \sigma_D^2}}. \tag{2}$$

Thus, the field gradient converts log-odds error to metric surface distance, and the two uncertainties set its weight. Invalid field queries and insufficient gradients contribute no geometric constraint.

Let $\mathcal{X}$ collect all keyframe states and sparse landmarks, and let $\mathcal{V}$, $\mathcal{I}$, and $\mathcal{G}$ index visual, inertial, and depth residuals r_v, r_i, and r_g, respectively. The joint objective is

$$E(\mathcal{X}) = \underbrace{\sum_{\mathcal{V}} \rho_v(\|r_v\|^2_{\Sigma_v})}_{\text{visual reprojection}} + \underbrace{\sum_{\mathcal{I}} \|r_i\|^2_{\Sigma_i}}_{\text{IMU preintegration}} + \underbrace{\sum_{\mathcal{G}} \rho_g(r_g^2)}_{\text{depth alignment}} + E_p. \tag{3}$$

Here $\|r\|^2_\Sigma = r^\top \Sigma^{-1} r$, Σ_v and Σ_i are the factor covariances, ρ_v and ρ_g are robust losses, and E_p carries the marginalization prior. The visual and inertial terms follow standard formulations (Leutenegger et al., 2015; Forster et al., 2017). Final bundle adjustment produces the wrist-camera trajectory, which the calibrated tool transfer maps into the gripper frame. Appendix C gives the factor and depth-processing details.

4.2 Action Supervision and Geometric Observations

Let C denote the tracked camera and E the tool center point (TCP). The calibrated transform chain gives $T_E^W(t) = T_S^W(t) T_C^S T_E^C$. At observation index t, the current pose $R_t = T_E^W(t)$ defines the local reference. History and future action targets are

$$H_{t,j} = \underbrace{R_t^{-1}}_{\text{reference}} \underbrace{T_E^W(t-j)}_{\text{history}}, \qquad A_{t,k} = R_t^{-1} \underbrace{T_E^W(t+k)}_{\text{future target}}, \quad k \geq 1. \tag{4}$$

Each target contains translation, a six-dimensional rotation representation (Zhou et al., 2019), and gripper width. All targets from one inference share the current measured TCP reference, following UMI's relative-trajectory principle (Chi et al., 2024).

RGB DP (Chi et al., 2023) receives wrist images. The two DP3 (Ze et al., 2024) variants instead receive 4,096 calibrated XYZ points constructed from factory depth (DP3–Raw) or learned active depth (DP3–Active). The DP3 variants share the point processing, state and action representations, architecture, and training budget; only the depth source changes.

At deployment, a robot-mounted Gemini provides the visual observation and robot kinematics supplies the measured TCP state. History advances once per completed policy action, preserving the discrete training semantics, and one TCP reference anchors the full predicted action chunk. Appendix F specifies the coordinate frame, training settings, and execution timing.

5 Real-World Experiments

We evaluate DepthUMI at three connected levels: wrist-camera trajectory accuracy, downstream action supervision and execution, and robot manipulation under increasing precision demands.

5.1 Wrist Motion Recovery

To test wrist motion across changes in scene appearance and texture, we record five normal-background and two low-texture white-cloth sequences. We compare the T265 SDK trajectory, ORB-SLAM3 Stereo–Inertial, and our estimator with either Fisheye depth or learned IR depth. ORB-SLAM3 is the visual–inertial backend widely used by UMI-family systems (Chi et al., 2024; Rayyan et al., 2025; Heng et al., 2025). Fisheye depth is an offline upper bound built from 45 local perspective stereo pairs, each processed by FoundationStereo (Wen et al., 2025). Each offline system is run five times per recording, and fixed-scale SE(3) alignment to the evaluation-only Vive reference uses common timestamps. Appendix D gives initialization and scoring details.

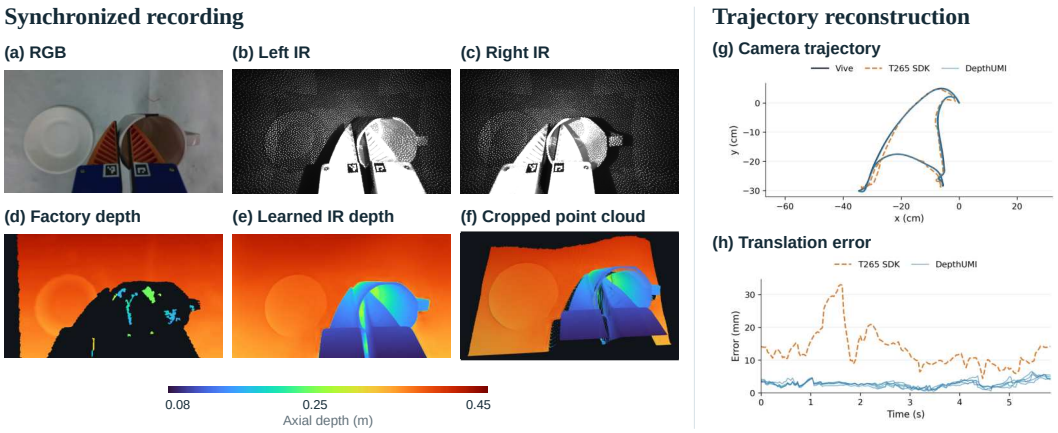

Figure 5: **Motion recovery on a complete cup pick-and-place demonstration.** (a–c) Synchronized RGB and left/right IR observations; (d–e) factory and learned IR depth inferred from the same IR pair; (f) the learned-depth point cloud. (g–h) A camera-frame example selected without cherry-picking compares the T265 SDK trajectory used by FastUMI with five DepthUMI reconstructions and the Vive reference.

Table 1: Real wrist-camera translation ATE (mm, lower is better). Offline entries are mean ± sample SD over five runs; Avg equally weights all seven recordings. Fisheye and learned-IR columns are tinted by depth source.

Recording	T265 SDK	ORB-SLAM3 Stereo–Inertial	Fisheye depth	Learned IR depth
Normal background				
N1	2.900	1.783±0.188	1.549±0.089	1.453±0.065
N2	6.405	2.874±0.438	2.067±0.082	1.993±0.023
N3	9.826	8.828±0.035	8.820±0.022	8.831±0.032
N4	4.817	1.071±0.047	1.092±0.075	1.183±0.077
N5	4.209	1.155±0.070	0.996±0.017	1.023±0.025
White-cloth background				
W1	684.111	3.068±0.071	2.617±0.061	2.636±0.069
W2	122.294	3.036±0.226	2.972±0.011	3.004±0.067
Avg	119.223	3.117	**2.873**	**2.875**

In Table 1, Fisheye and learned IR depth reach 2.873 and 2.875 mm overall mean translation ATE, compared with 3.117 mm for ORB-SLAM3 and 119.223 mm for the T265 SDK. Learned IR improves on ORB-SLAM3 by 0.242 mm (7.8%) and matches the Fisheye upper bound within 0.002 mm at roughly one-fiftieth of its depth-inference time. It therefore recovers the accuracy of the 45-view offline reference at deployable cost; ORB-SLAM3's 35/35 completion shows that the margin does not come from excluding failed runs. Figure 5 visualizes synchronized sensing and motion recovery for a complete cup pick-and-place demonstration; Appendix Figures 9 and 10 provide per-recording trajectory and error plots.

5.2 Motion Recovery on Complete Demonstrations

Camera-level ATE establishes tracking accuracy, but a UMI needs complete demonstrations that remain useful as robot actions. We therefore collect 30 cup pick-and-place demonstrations in a 100 × 100 cm tabletop workspace. Table 2 compares five motion sources under a fixed downstream DP3 observation and control recipe. All methods process the same recordings; ORB-SLAM3 contributes the 28 with complete support, while the other sources contribute all 30. Appendix F.2 defines the matched episodes, aggregation, and downstream evaluation protocol.

Table 2: **From global reconstruction to action supervision and execution.** Global columns report median cup-trajectory ATE; action cells report median translation RMSE (mm) / rotation RMSE (°) on the common episode set. Replay and policy report successes / five trials.

		Global reconstruction		Action-relative RMSE				Execution	
Motion source	**Global support**	**Trans.↓** (mm)	**Rot.↓** (°)	$k=1$	$k=3$	$k=6$	$k=16$	**Replay↑**	**Policy↑**
Vive reference	30/30	ref.	ref.	ref.	ref.	ref.	ref.	**5/5**	**5/5**
T265 SDK	30/30	18.468	2.914	7.82/.24	14.00/.40	19.77/.56	29.75/.80	1/5	3/5
ORB-SLAM3	28/30	5.177	1.362	2.83/.18	4.34/.34	6.17/.50	9.30/.68	**5/5**	4/5
Fisheye depth	30/30	2.782	**1.191**	1.42/.15	2.91/.32	4.49/.49	7.04/.68	**5/5**	**5/5**
Learned IR depth	30/30	**2.723**	1.214	1.51/.15	2.91/.32	4.53/.49	7.27/.68	**5/5**	**5/5**

Accurate global reconstruction also improves controller-facing motion. On each method's usable support in Table 2, median translation ATE is 18.468 mm for T265, 5.177 mm for ORB-SLAM3, 2.782 mm for Fisheye depth, and 2.723 mm for learned IR. On the matched common set, learned IR reduces action-relative translation RMSE from 2.83 to 1.51 mm at one step and from 9.30 to 7.27 mm at 16 steps relative to ORB-SLAM3, closely tracking the Fisheye upper bound. Accurate reconstruction is therefore preserved in the action labels.

T265 nevertheless reaches 3/5 current-relative policy rollouts despite only 1/5 direct replays and far higher ATE because local targets attenuate slowly varying global drift (Chi et al., 2024). This explains how T265-based systems can work, but also why rollout success does not certify faithful global reconstruction. Both depth-aided sources pair 5/5 replay and policy success with precise global trajectories, combining local-control robustness with accurate action supervision.

5.3 Robot Manipulation

We finally ask whether the same geometry improves closed-loop control. Figure 6 shows the progression: Drawer Opening requires acquiring and pulling a broad handle, Peg Insertion adds precise alignment, and Block Extraction requires grasping and removing a tightly packed block. We compare RGB DP, DP3–Raw, and DP3–Active on a Franka Research 3, varying task poses by up to ± 5 cm and $\pm 15°$ as well as the robot initial state.

All policies share demonstrations, action supervision, data splits, and the current-relative control convention. DP3 also shares architecture, crop, 4,096-point deterministic FPS, and training budget while changing the geometry source. Appendix G gives ranges, criteria, and intervals.

Table 3: **Manipulation performance.** FR3 success rate over 10 trials per task and policy.

		Success rate ↑			
Policy	**Observation**	**Drawer Opening**	**Peg Insertion**	**Block Extraction**	**Total↑**
DP	RGB	70%	10%	10%	30%
DP3	Factory depth	50%	0%	10%	20%
DP3	Learned active depth	**80%**	**70%**	**50%**	**67%**

Geometry gains grow with task precision. Drawer Opening tolerates coarse handle localization, so all three settings achieve a solid 50–80% success rate in Table 3. The picture changes once contact demands precise alignment. DP3–Active improves over RGB from 10% to 70% on Peg Insertion and from 10% to 50% on Block Extraction, whereas DP3–Raw reaches only 0% and 10%. Depth alone is therefore not enough. As we can see in Figure 7, factory depth becomes sparse or invalid during the near-field approach, precisely when the policy needs geometric guidance; learned active depth continues to expose dense, accurate local surfaces for closed-loop correction. Block Extraction makes the demand especially clear: the 80 mm gripper opening exceeds the 75 mm-wide block by

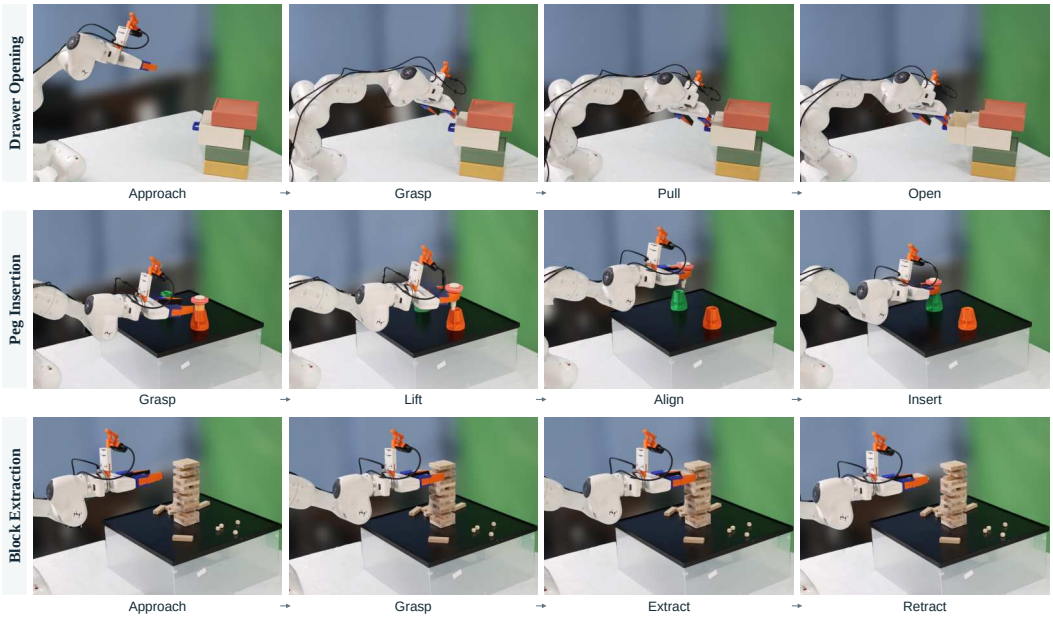

Figure 6: **Third-person real-robot executions.** Four stages of Drawer Opening, Peg Insertion, and Block Extraction.

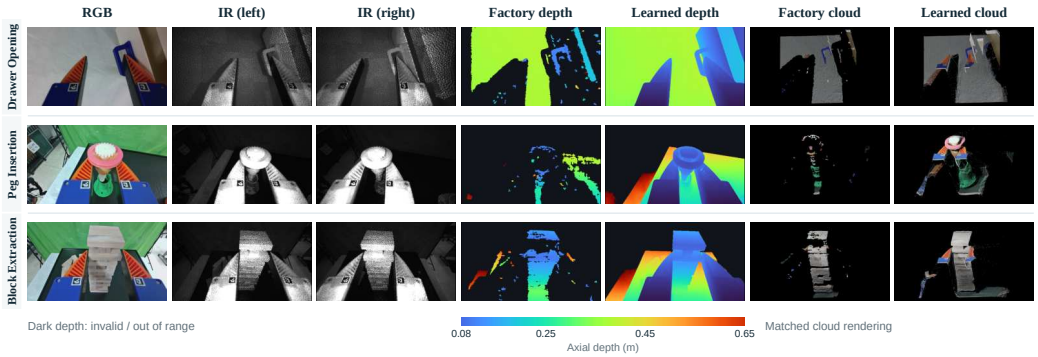

Figure 7: **Onboard deployment geometry.** RGB, IR, depth, and matched point-cloud observations.

only 5 mm in total. The substantially wider gap on Peg Insertion and Block Extraction shows that reliable near-field geometry is what turns depth into useful guidance for fine manipulation.

6 Conclusion

We introduce DepthUMI, the first UMI system to use learned active depth for both demonstration motion recovery and policy observation. Its wrist-contained interface turns wide-field motion cues and dense near-field geometry into gripper-centric actions and metric policy inputs. Across simulation and real-world experiments, two insights emerge: current-relative control can mask poor global reconstruction, while depth helps only when it remains reliable near contact. DepthUMI addresses both within one system.

While these results establish active wrist geometry as a practical foundation for UMI, several limitations remain. Fabrication and 3D-printed mounting still introduce calibration and record-to-deploy alignment error, while N3 motivates wrist-adapted local SLAM for rapid motion. As a next step, we will develop a wider-field RGB–IR stereo device shared by recording and deployment, reducing calibration burden and sensor redundancy to enable a lower-cost, more efficient UMI platform.

AI USE STATEMENT

Generative AI tools assisted with language polishing, structural revision, LaTeX formatting, scientific-figure composition and modification, experiment-design feedback, code development, result interpretation, and critical review of methodological claims and evaluation procedures. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, figures, and artifacts.

REFERENCES

Simon Boche, Jaehyung Jung, Sebastián Barbas Laina, and Stefan Leutenegger. OKVIS2-X: Open keyframe-based visual–inertial SLAM configurable with dense depth or LiDAR, and GNSS. *IEEE Transactions on Robotics*, 41:6064–6083, 2025. doi: 10.1109/TRO.2025.3619051.

Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. doi: 10.1109/TRO.2021.3075644.

Tianxing Chen et al. RoboDojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026. URL https://arxiv.org/abs/2607.04434.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems*, 2024.

Hojung Choi, Yifan Hou, Chuer Pan, Seongheon Hong, Austin Patel, Xiaomeng Xu, Mark R. Cutkosky, and Shuran Song. In-the-wild compliant manipulation with UMI-FT. In *IEEE International Conference on Robotics and Automation*, 2026.

Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2017. doi: 10.1109/TRO.2016.2597321.

Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

Liang Heng, Xiaoqi Li, Shangqing Mao, Jiaming Liu, Ruolin Liu, Jingli Wei, Yu-Kai Wang, Yueru Jia, Chenyang Gu, Rui Zhao, Shanghang Zhang, and Hao Dong. RwoR: Generating robot demonstrations from human hand collection for policy learning without robot. *arXiv preprint arXiv:2507.03930*, 2025. URL https://arxiv.org/abs/2507.03930.

Lesia Hentosh, Yaroslav Hladun, Ivan Tsmots, and Oleh Bisikalo. Depth-augmented diffusion policy with pseudo-depth for robust robotic manipulation. *Scientific Reports*, 16:28233, 2026. doi: 10.1038/s41598-026-62831-6.

Yan Huang, Shoujie Li, Xingting Li, and Wenbo Ding. UMIGen: A unified framework for egocentric point cloud generation and cross-embodiment robotic imitation learning. *arXiv preprint arXiv:2511.09302*, 2025.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3D diffuser actor: Policy diffusion with 3D scene representations. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1949–1974, 2025. URL https://proceedings.mlr.press/v270/ke25a.html.

Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015. doi: 10.1177/0278364914554813.

Jiaheng Li, Qiyu Dai, Lihan Li, Praneeth Chakravarthula, He Sun, Baoquan Chen, and Wenzheng Chen. Robust single-shot structured light 3D imaging via neural feature decoding. *arXiv preprint arXiv:2512.14028*, 2025.

Jiaheng Li, Binsheng Zhang, Xinhai Chang, and Wenzheng Chen. NSL-SLAM: High-fidelity neural structured-light depth for practical SLAM and reconstruction. *arXiv preprint arXiv:2607.24495*, 2026.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88b7b2c896506daabc8d3fd587055167-Abstract-Conference.html.

Lumos Robotics. Lumos FastUMI Pro. Product documentation, n.d. URL https://www.lumosbot.tech/products/fastumi-pro/. Accessed September 8, 2026.

NVIDIA. NVIDIA Isaac Sim. Robotics simulation framework, n.d. URL https://developer.nvidia.com/isaac/sim. Accessed September 10, 2026.

Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual–inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. doi: 10.1109/TRO.2018.2853729.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Omar Rayyan, John Abanes, Mahmoud Hafez, Anthony Tzes, and Fares Abu-Dakka. MV-UMI: A scalable multi-view interface for cross-embodiment learning. *arXiv preprint arXiv:2509.18757*, 2025.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 785–799, 2023. URL https://proceedings.mlr.press/v205/shridhar23a.html.

Simple AI, Yuteng Wei, Jinming Ma, Jiawei Wang, Weitao Zhou, Yushen Zuo, Ke Rui, Minglei Li, Jinhao Zhang, Zhikang Pan, Xiang Wang, Haoran Jia, Huan Du, Zicheng Zeng, Jun Ma, Guiyu Qin, Di Zhang, and Xiaofei Li. HiFi-UMI: Learning deployable manipulation policies from high-fidelity UMI data alone. *arXiv preprint arXiv:2607.25895*, 2026.

Chenxi Wang, Hongjie Fang, Hao-Shu Fang, and Cewu Lu. RISE: 3D perception makes real-world robot imitation simple and effective. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2870–2877, 2024. doi: 10.1109/IROS58592.2024.10801678.

Ziming Wang. UMI-3D: Extending universal manipulation interface from vision-limited to 3D spatial perception. *arXiv preprint arXiv:2604.14089*, 2026.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5249–5260, 2025.

Yue Xu, Litao Wei, Pengyu An, Qingyu Zhang, and Yong-Lu Li. exUMI: Extensible robot teaching system with action-aware task-agnostic tactile representation. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2536–2554, 2025. URL https://proceedings.mlr.press/v305/xu25e.html.

Ge Yan, Jiyue Zhu, Yuquan Deng, Shiqi Yang, Ri-Zhao Qiu, Xuxin Cheng, Marius Memmel, Ranjay Krishna, Ankit Goyal, Xiaolong Wang, and Dieter Fox. ManiFlow: A general robot manipulation policy via consistency flow training. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2268–2293, 2025. URL https://proceedings.mlr.press/v305/yan25a.html.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations. In *Proceedings of Robotics: Science and Systems*, 2024.

Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3D diffusion policies. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2873–2880, 2025.

Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. ActiveStereoNet: End-to-end self-supervised learning for active stereo systems. In *Proceedings of the European Conference on Computer Vision*, 2018.

Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. UFM: A simple path towards unified dense correspondence with flow. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html.

Zhemeng Zhang, Jiahua Ma, Xincheng Yang, Xin Wen, Yuzhi Zhang, Boyan Li, Yiran Qin, Jin Liu, Can Zhao, Li Kang, Haoqin Hong, Zhenfei Yin, Philip Torr, Hao Su, Ruimao Zhang, and Daolin Ma. TouchGuide: Inference-time steering of visuomotor policies via touch guidance. In *Proceedings of Robotics: Science and Systems*, 2026. doi: 10.15607/RSS.2026.XXII.078. URL https://www.roboticsproceedings.org/rss22/p078.html.

Haoyu Zhao, Xingyue Zhao, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. RynnWorld-4D: 4D embodied world models for robotic manipulation. *arXiv preprint arXiv:2607.06559*, 2026.

Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. FastUMI: A scalable and hardware-independent universal manipulation interface with dataset. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3069–3093, 2025. URL https://proceedings.mlr.press/v305/zhaxizhuoma25a.html.

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, and Mengyu Wang. GEM-4D: Geometry-enhanced video world models for robot manipulation. *arXiv preprint arXiv:2605.22882*, 2026. URL https://arxiv.org/abs/2605.22882.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.



Figure 8: **Photographed handheld and robot-mounted hardware.** Front and rear views show the validation and compact acquisition assemblies. The validation rig adds a Vive reference and multiple inertial streams for controlled measurement, whereas the acquisition device retains one IMU and no external pose reference. The robot-mounted camera supplies observations during execution, with TCP state obtained from robot kinematics.

A Hardware configurations and sensing roles

The evaluation apparatus and the compact acquisition device serve different purposes. The instrumented rig carries the two cameras, multiple inertial streams, and a Vive tracker so that sensing choices and an external reference can be recorded during the same motion. The compact acquisition device removes the reference tracker and retains a single HI12 IMU. Each physical assembly uses its own camera–IMU transform, time offset, and inertial noise model.

Gemini RGB observes appearance, while its active IR pair supplies learned metric depth. T265 contributes wide-field fisheye images for motion estimation. Its proprietary pose output is a separate system comparison, rather than an input to the depth-aided estimator. The external IMU measures motion between images. All localization sensors move with the gripper; the host supplies computation and power. Vive supports reference evaluation and does not enter the estimator.

The robot-side Gemini is a separate unit mounted on the FR3 gripper. Its calibration is not interchangeable with the handheld camera's. Robot kinematics supplies the measured tool state during deployment, so execution does not require demonstration SLAM to localize the robot. Figure 8 shows photographs of both handheld configurations and the robot-mounted sensor.

B Calibration procedure and reference checks

Figure 3 summarizes camera geometry, camera–IMU alignment, and visual TCP estimation alongside the hardware configurations.

B.1 Coordinate conventions and calibration order

We write T_B^A for the rigid transformation from frame B to frame A. The principal frames are world W, selected IMU S, Gemini RGB camera C, left IR camera L, and tool center point E. Distances are in meters and inertial measurements in rad/s and m/s^2. Tool-motion supervision is

$$T_E^W(t) = T_S^W(t) T_C^S T_E^C. \tag{5}$$

Changing the selected IMU changes S and its calibrated camera relationship, not the physical tool frame.

Calibration separates wide-field coverage, inter-camera alignment, inertial noise characterization, and camera–IMU alignment. This ordering prevents the restricted common view from determining

Table 4: Five-camera calibration for the instrumented validation rig. Reprojection RMS measures the target fit, not tool-motion accuracy.

Camera	Distortion model	Reprojection RMS (px)
Gemini RGB	Radial–tangential	0.5960
Gemini IR-left	Radial–tangential	0.2410
Gemini IR-right	Radial–tangential	0.2485
T265 fisheye-left	Equidistant	0.3152
T265 fisheye-right	Equidistant	0.3310

the fisheye intrinsics. The target is an A3 AprilGrid with 7×11 tags, 27.5 mm tag size, and 8.25 mm gaps. Factory parameters initialize camera models; inter-device rigid transformations are solved from recorded target views.

B.2 Wide-field and common-view camera geometry

A 60 s sequence covers fisheye image centers, edges, corners, and oblique views. A separate 120 s sequence keeps the target visible to all five cameras while exciting translation and rotation about all three axes. Accepted fisheye intrinsics are fixed in the second solve, which estimates Gemini RGB/IR intrinsics and inter-camera transformations. Both acquisitions preserve all configured streams, although their solvers select different inputs.

The fisheyes use pinhole models with equidistant distortion; Gemini cameras use radial–tangential distortion. Solved image sizes are 1280×720 for RGB, 848×480 for IR, and 848×800 for fisheyes. The IR projector is disabled during common-view calibration to preserve printed-target contrast. Demonstration processing retains the illumination and exposure settings recorded for each capture cohort.

Table 4 reports the validation-rig geometry fit. The independent fisheye stage retains 219 of 224 selected observations and estimates a 64.3059 mm baseline. The common-view stage estimates 64.4309 mm, a 0.1250 mm change. This is cross-stage consistency, not independent physical baseline metrology.

B.3 Spatial alignment and acquisition time

Fixed camera geometry, the common-view motion, and measured noise parameters enter camera–IMU spatial and temporal calibration (Furgale et al., 2013). The solve estimates rigid sensor transforms and relative time shifts. The Kalibr convention is $t_{\mathrm{imu}} = t_{\mathrm{camera}} + \Delta t$; adapters explicitly convert this convention rather than reuse an offset with an unspecified sign.

Run-local clock association and calibrated relative offsets serve different purposes. Sensors retain device and host timestamps. External IMU packets may contain several samples sharing one host receipt, so their sample clocks are reconstructed instead of assigning that receipt time to every measurement. Acceleration is interpolated at bracketed gyroscope timestamps with a bounded gap. Acquisition is software-associated, without a shared hardware trigger.

B.4 Tool transfer and evaluation-reference closure

The handheld camera-to-tool transform T_E^C, robot end-effector-to-TCP transform, and physical gripper opening limits are separate quantities. Camera reprojection cannot validate them. Physical checks establish the contact-center position and orientation and the gripper width at its end stops. We therefore report tracking error in the camera frame and do not reinterpret it as an absolute physical TCP-accuracy bound.

For reference evaluation, camera–Vive hand–eye calibration uses left-IR observations of the fixed grid and the calibrated IR-to-RGB relationship. With tracker frame V and grid frame G, closure is

$$T_V^W(t) T_C^V T_G^C(t) = T_G^W . \tag{6}$$

Of 362 synchronized observations, 289 enter fitting and 73 are held out from fitting and time-offset selection. Held-out closure RMSE is 2.990 mm in translation and 0.477° in rotation; the corre-

sponding 95th percentiles are 5.807 mm and $0.885°$. These characterize the camera–grid–tracker chain within that recording. They are neither absolute TCP accuracy nor an error bar to subtract from trajectory ATE.

C Depth and motion reconstruction details

C.1 Depth coordinates and validity

The learned IR model predicts optical-axis distance directly on the calibrated pinhole raster of the left-IR camera. Valid pixels are back-projected with their recorded intrinsics and transformed into the required camera or submap frame. Real-world reconstruction retains the native 848×480 IR raster without an additional view decomposition.

The fisheye baseline calibrates the T265 stereo pair, decomposes each frame into 45 overlapping local perspective pairs, and applies FoundationStereo (Wen et al., 2025) to each rectified pair. The local predictions are transformed and stitched into the wide-angle camera frame; Euclidean ray range is converted to optical-axis depth before geometric fusion. We use this multi-view construction as an offline capability reference for Fisheye depth.

The final tracking comparison uses the frozen UFM-based learned structured-light model without confidence filtering; policy learning uses the same learned-depth output contract fixed for each evaluation cohort. Factory and learned depth remain separate provenance-tagged modalities. Asynchronous depth and fisheye frames are associated using their recorded timestamps, which remain available throughout processing.

C.2 Geometric residual and factor configuration

The estimator adapts the volumetric alignment mechanism of OKVIS2-X (Boche et al., 2025). For a point $\mathbf{p}_b$ associated with pose T_b^W, submap coordinates are $\mathbf{p}_a = (T_a^W)^{-1} T_b^W \mathbf{p}_b$. Let $\ell_a(\mathbf{p}_a)$ be the interpolated signed log-odds occupancy field (denoted f_a in the main text), with zero level at the surface, and $\mathbf{g}_a = \nabla \ell_a(\mathbf{p}_a)$ its gradient. The geometric residual is

$$r_g = \frac{\ell_a(\mathbf{p}_a)}{\|\mathbf{g}_a\| \sqrt{\sigma_{\text{map}}^2 + \sigma_{\text{meas}}^2}}, \qquad \sigma_{\text{map}} = \frac{|\ell_{\min}|}{3\|\mathbf{g}_a\|}, \tag{7}$$

where $\ell_{\min}$ is the map's lower log-odds parameter and σ_{meas} is configured measurement uncertainty (σ_D in the main text). A missing field value or gradient, or gradient norm below 10^{-3}, gives an inactive residual. Geometry uses a Tukey loss; visual reprojection uses a Cauchy loss. Submap self-alignment constraints are excluded. Wide-field features and inertial preintegration remain necessary when local geometry is ambiguous.

Frame-to-submap and submap-to-submap factors can be enabled separately. The near plane affects occupancy integration and candidate selection, so it is not equivalent to a gripper mask. Increasing σ_{meas} changes residual weighting and the physical residual range admitted by the robust loss. These settings are not measurements of depth error. Appendix D specifies the shared-recipe evaluation protocol.

C.3 Trajectory error and relative action supervision

For current-pose-relative supervision, the target at horizon k is $A_{t,k} = T_t^{-1} T_{t+k}$. A constant world-frame change cancels. If recovered poses satisfy $\widehat{T}_t = T_t \exp(\xi_t^\wedge)$, then

$$\widehat{A}_{t,k} = \exp(-\xi_t^\wedge) A_{t,k} \exp(\xi_{t+k}^\wedge). \tag{8}$$

Action-label fidelity therefore depends on relative errors over the action horizon, not trajectory ATE alone. We distinguish trajectory agreement from relative translation and rotation error at the horizons consumed by the policy. This identity follows from the representation; it introduces no additional estimator or learned correction.

Table 5: Background-group translation ATE for the instrumented-rig comparison (mm). Each group equally weights its recordings; overall equally weights all seven. ORB-SLAM3 is the UMI-family stereo–inertial baseline; the other two rows use our OKVIS2-X-based pipeline.

Input	Normal (5)	White-cloth (2)	Overall (7)
T265 SDK	5.631	403.203	119.223
ORB-SLAM3	3.142	3.052	3.117
Fisheye depth	2.905	2.794	2.873
Learned IR depth	2.897	2.820	2.875

D Real tracking protocols and additional results

D.1 Recording groups, selection, and repetitions

The benchmark contains five normal-background recordings, N1–N5, and two white-cloth recordings, W1–W2. Calibration-derived sensor geometry and one globally fixed estimator recipe are applied uniformly to all seven recordings and both depth inputs; there is no recording-specific tuning or best-run selection.

ORB-SLAM3 is configured as the UMI-family stereo–inertial tracking baseline. It uses the recorded Stereo–Inertial input and sets the low-motion initialization-reset threshold to zero. This adjustment prevents low-motion reset during initialization and differs from the official default configuration and from prior UMI hardware implementations. It does not alter the recorded measurements or the common scoring grid.

Five fresh invocations per offline system and recording give 105 offline comparison runs; T265 contributes seven recorded device trajectories. Every offline cell uses all five runs with sample standard deviation, and all 35 ORB-SLAM3 runs are complete. Repetitions measure solver variation on the same observations, not independent demonstration captures or matched random seeds.

D.2 Reference, alignment, and aggregation

Estimator poses are transformed from the selected IMU to the T265-left optical frame before scoring. Vive uses fixed calibrated hand–eye geometry and reference timing. Final bundle-adjusted trajectories receive one rigid SE(3) alignment without scale fitting. The previously fixed 296/297-point grids cover 99.0–99.3% of each acquisition interval. A run missing required support does not receive a shorter scoring interval. N3 retains the complete frozen grid even though its Vive reference contains isolated 40–80 mm jumps that are not corroborated by the SDK trajectory or local inertial measurements. Its ATE is therefore reported against the recorded reference rather than treated as certified absolute motion error.

Translation ATE above 100 mm defines divergence in this real benchmark. Finite divergent scores remain in averages. T265 SDK contributes one recorded online trajectory per case, with its internal IMU and processing budget.

The overall mean equally weights seven recording means, not the two background means, since group sizes differ. Table 5 reports the aggregates. The fisheye and learned-depth systems both remain below 3 mm overall; their means differ by 0.002 mm, well below the physical calibration residual discussed in Appendix B.

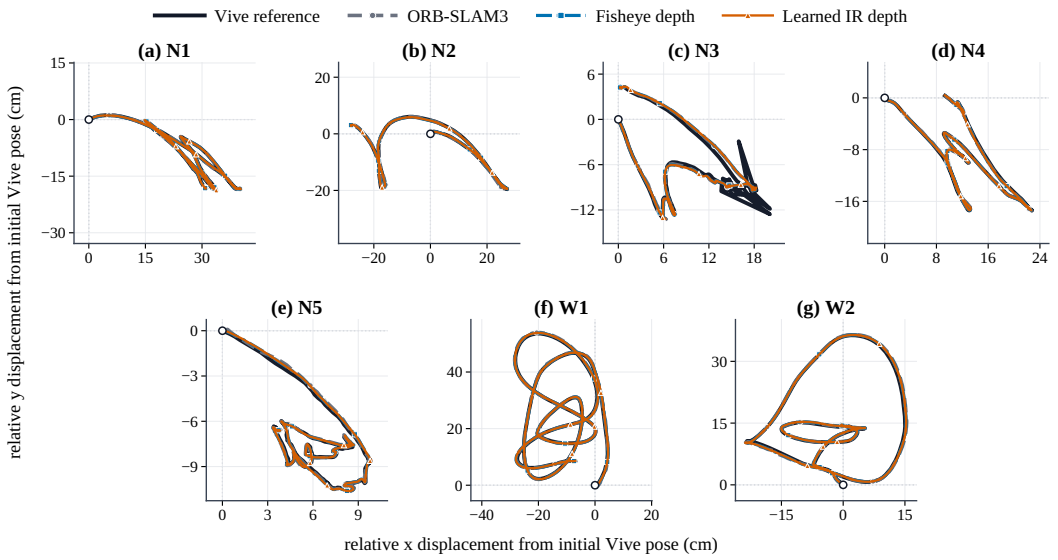

Figure 9: **Trajectory agreement across the seven-recording benchmark.** Each panel projects one evaluation-aligned camera trajectory relative to its initial Vive pose and is independently scaled to that recording's planar extent. Vive is the external reference; offline curves show all five runs faintly and their timestamp-wise mean prominently. Compare methods within each panel; cross-recording magnitude is reported by the quantitative metrics.

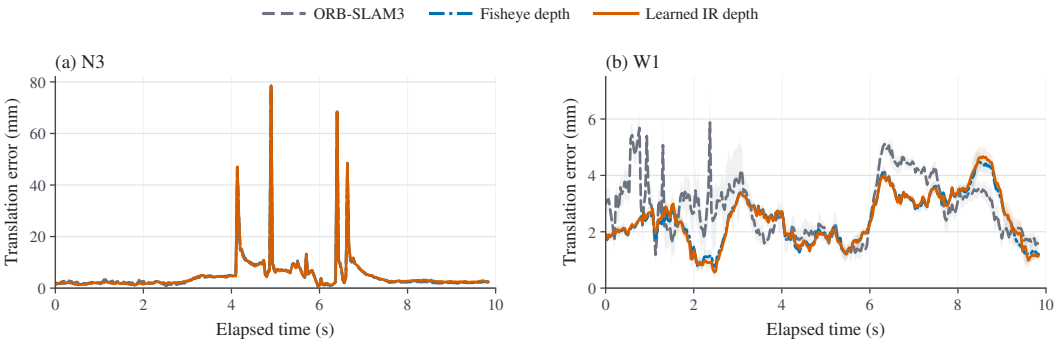

Figure 10: **Time-resolved tracking error.** Translation error to the evaluation-only Vive reference for N3 and W1; lines and bands show mean ± sample standard deviation over five runs. N3 retains the frozen reference discontinuities, while the separately labeled W1 scale preserves millimeter-level structure.

E Simulation-guided sensing design details

E.1 Task, background, and placement coverage

Tube insertion supplies four rendered backgrounds: a furnished room, white walls, white walls and floor, and white walls, floor, and tabletop. Table organization and bowl stacking each supply the furnished and all-white endpoints for quantitative comparison. The three tasks use distinct motion recordings. All white-surface conditions also remove tabletop markers; the furnished scenes retain them.

Each task/background case contains independently tracked head, left-wrist, and right-wrist observations. Every placement uses stereo, stereo plus causal simulator inertial input, and stereo plus predicted depth and that inertial input. Cells use three invocations, except the two bowl right-wrist learned-depth cells with five, giving 220 plotted solves. Ten corresponding ground-truth-depth

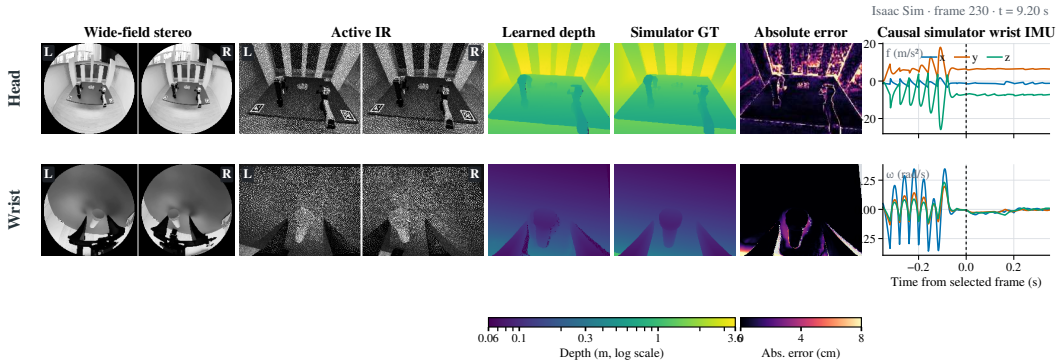

Figure 11: **Synchronized simulation observations.** Head and wrist views from one Isaac Sim timestamp pair wide-field stereo and active IR with learned and ground-truth depth; absolute error is shown in centimeters, alongside a short causal kinematics-derived wrist-IMU window. The modalities expose complementary scene context, high-rate motion, and near-field geometry that motivated the final sensing stack.

solves are reported in Table 8. Within each task, background changes preserve motion. Placements follow different motions, rather than matched tracking-difficulty trials or a head-to-hand localization system.

E.2 Sensor and evaluation contracts

Images run at 25 Hz and the causal kinematics-derived IMU at 1 kHz. Estimator gravity is 9.81 m/s^2 with nonzero inertial noise priors. The IMU is derived from simulator kinematics rather than a recorded device stream. A mapping-only pinhole camera consumes NSL predictions from rendered active IR. For this simulation study, predictions are RGB-aligned and unthresholded, with dilation during preparation; they are neither simulator GT depth nor the recorded native-IR real input.

The transport mask retains 0.1–10 m before uint16-millimeter conversion; mapping retains 0.1–3 m. A native-reader timestamp offset is applied consistently and removed before scoring. Optical poses and IMU mounting transforms are checked before export. No artificial stationary prefix or ground-truth initial pose is added.

For bowl right-wrist depth evaluation, cost-only and Jacobian evaluations use the same depth residual, and `sigma_a_c=` 0.1 accounts for numerical error in the differentiated synthetic IMU. All depth samples and timestamps, both frame-to-submap and submap-to-submap factors, and the remaining estimator settings are used. No depth warm-up or image mask is used.

All reported trajectories within each task/background use a shared scoring grid: 343 timestamps for insertion, 862 for organization, and 355 for bowl stacking, covering 99.1%, 99.7%, and 99.2% of their respective temporal spans. These intersections are computed separately for each case. Alignment is rigid SE(3) without scale fitting. Wrists use position-based rotation fitting; the less translationally excited head triggers orientation-based fallback. Tube head optical-camera path length is 0.297 m, versus 2.825 m and 1.170 m for left and right wrists. Cross-placement ATE is descriptive, not a causal placement ranking.

E.3 Complete results and initialization

Table 6 gives all insertion wrist results. Tables 7, 8, and 9 complete the head, bowl-stacking, and organization cases. Means and sample standard deviations use every reported solve. All processes produce evaluable trajectories, including retained large errors; 13 exceed 10 m, all from left-wrist trials. Plot overflow bands retain these identities and counts, and means use unclipped values.

For right-wrist insertion, depth reduces stereo–inertial mean ATE from 3.25, 7.94, and 10.17 mm to 1.78, 5.52, and 6.20 mm in the first three backgrounds. The all-white tabletop changes from 4.25 to 5.40 mm. Organization changes from 2.59 to 2.64 mm in the furnished scene and from 6.13 to

Table 6: Simulation camera translation ATE (mm, lower is better), mean ± sample SD of three unselected solves per cell. I: causal kinematics-derived IMU; D: learned depth decoded from active IR in a separate pinhole mapping camera. All 72 solves use one backend and 343 common timestamps (99.1% coverage). Large errors are retained. The two wrists are different motions, not repeated trials.

Input	Furnished room	White walls	+ White floor	+ White tabletop
Left wrist				
Stereo	681.0±336.2	1339.7±1749.3	267.2±59.6	3958.6±5505.9
Stereo + I	1364.0±1791.8	2678.0±1938.8	1617.3±1120.6	5763.9±912.6
Stereo + D + I	117.0±125.8	2621.2±2060.4	2272.1±287.5	2344.0±978.7
Right wrist				
Stereo	161.8±109.1	317.3±34.5	160.3±76.4	550.1±280.7
Stereo + I	3.2±0.3	7.9±3.9	10.2±2.6	4.3±1.9
Stereo + D + I	1.8±0.4	5.5±0.9	6.2±0.6	5.4±3.6

Table 7: Head-placement results for tube insertion and table organization: camera translation ATE in mm, mean ± sample SD over three solves. I is a causal kinematics-derived IMU; D is learned depth decoded from active IR. All six cases are retained.

Task / background	Stereo	Stereo + I	Stereo + D + I
Insertion / furnished	2.601 ± 0.347	48.553 ± 61.742	1.878 ± 0.050
Insertion / white walls	30.724 ± 3.217	9.891 ± 1.004	9.670 ± 0.498
Insertion / + white floor	38.625 ± 4.891	19.556 ± 22.793	8.544 ± 1.807
Insertion / + white table	38.478 ± 1.000	27.227 ± 33.876	313.588 ± 259.849
Organization / furnished	2.429 ± 0.158	205.314 ± 195.955	1.823 ± 0.008
Organization / all-white	39.604 ± 14.426	49.076 ± 36.873	7.348 ± 0.608

5.26 mm in the all-white scene. For insertion and organization, head depth fusion improves five of six case means, but all-white insertion changes from 27.23 to 313.59 mm.

Bowl stacking further exposes placement-dependent behavior. For the right wrist, learned depth gives 10.26 ± 11.42 mm in the furnished scene and 27.89 ± 12.21 mm in the all-white scene; ground-truth depth gives 11.46 ± 11.48 and 39.85 ± 16.99 mm. None of these 20 depth-fused runs exceeds 100 mm. The matched depth-free condition gives 7.92 ± 3.44 mm in the furnished scene but 2926.33 ± 5032.40 mm in the all-white scene, where one of three runs reaches 8.74 m. The comparison shows that both learned and ground-truth depth stabilize this texture-poor condition.

The estimator initializes attitude by aligning mean measured acceleration with gravity. The left wrist is already moving during that window for insertion and organization, where its initial mean direction differs from gravity by 7.475° and 19.910°, respectively. In bowl stacking, the right wrist also starts immediately: its first 60 ms have 13.867 m/s^2 mean specific-force norm and about 5.29° apparent tilt. Source checks find no axis swap, gravity-sign error, unit mismatch, gross timing offset, or camera/IMU motion mismatch. The reported inertial model accounts for the synthetic stream's finite differentiation accuracy instead of treating it as noise-free inertia.

Table 8: Complete bowl-stacking camera translation ATE in mm, mean ± sample SD on 355 common timestamps. I is a causal kinematics-derived IMU; D and GT are learned and ground-truth depth. Cells use three runs, except the right-wrist D and GT cells, which use five.

Background / placement	Stereo	Stereo + I	Stereo + D + I	Stereo + GT + I
Furnished / head	2.905 ± 0.437	20.504 ± 15.724	2.446 ± 0.912	—
Furnished / left wrist	244.106 ± 149.607	2.928 ± 0.429	2.313 ± 0.103	—
Furnished / right wrist	243.413 ± 114.214	7.918 ± 3.442	10.257 ± 11.423	11.460 ± 11.480
All-white / head	38.386 ± 2.374	6.997 ± 0.078	7.528 ± 0.768	—
All-white / left wrist	880.430 ± 360.714	89.049 ± 42.727	33.031 ± 46.907	—
All-white / right wrist	1164.370 ± 447.203	2926.331 ± 5032.401	27.890 ± 12.208	39.846 ± 16.992

Table 9: Complete table-organization wrist results: camera translation ATE in mm, mean ± sample SD over three solves. Large left-wrist errors are divergent trajectories, not unavailable measurements. Scientific notation preserves their scale.

Background / wrist	Stereo	Stereo + I	Stereo + D + I
Furnished / left	409.093 ± 167.781	$(6.599 \pm 3.725) \times 10^7$	$(4.688 \pm 1.993) \times 10^7$
Furnished / right	653.095 ± 389.801	2.586 ± 0.876	2.642 ± 1.265
All-white / left	743.294 ± 374.213	$(1.171 \pm 1.095) \times 10^8$	$(3.330 \pm 4.440) \times 10^6$
All-white / right	649.355 ± 359.638	6.126 ± 1.633	5.263 ± 0.615

F Demonstration data and policy learning

F.1 Data and policy settings

Tables 10 and 11 report the instantiated training and online-inference settings for DP and DP3, respectively. Drawer Opening, Peg Insertion, and Block Extraction use the bounded spatial perturbations specified in Section 5.3. Training and development episodes are disjoint, and checkpoint selection uses development rollouts rather than formal robot outcomes.

Table 10: Training and inference settings for RGB Diffusion Policy (DP). Checkpoints are selected by development rollouts; formal robot outcomes are held out from model selection.

Hyperparameter	Value
Visual observation	Wrist RGB, 224×224; no image augmentation
Temporal / action	Three observations at 10 Hz; 10-D current-TCP-relative pose and gripper command
Prediction / execution horizon	16 / 6 policy steps
Visual encoder	Pretrained CLIP ViT (Radford et al., 2021); 512-D feature
Denoiser / conditioning	U-Net $[256, 512, 1024]$, kernel 5, 8 groups; global condition; 128-D step embedding
Diffusion / inference	DDIM, 50 training steps, ϵ prediction, squared-cosine noise; DDIM / 16 inference steps
Optimizer / learning rate	AdamW; 3×10^{-4} (encoder: 3×10^{-5})
Adam / regularization	$\beta = (0.95, 0.999)$; $\epsilon = 10^{-8}$; weight decay 10^{-6}
Schedule / batch / precision	Cosine; 1,000-step warmup; 32 / bfloat16
Training / selection	1,000 epochs; seed 42; EMA power 0.75, max decay 0.9999; candidates at epochs 300, 800, and 1,000; reported checkpoint 1,000

DP consumes wrist RGB and encodes each frame with a pretrained visual backbone. Its action chunks are expressed relative to the measured TCP at the current policy step; the same reference is retained for every target within that chunk. The three-frame history advances only after a completed robot action, rather than at the camera frame rate.

Table 11: Training and inference settings for 3D Diffusion Policy (DP3). The factory-depth and learned-depth variants share these settings; only the depth source used to construct XYZ changes.

Hyperparameter	Value
Geometric observation	Metric XYZ from factory or learned active depth
Frame / workspace crop (m)	Left-IR optical frame; $x \in [-1, 1]$, $y \in [-0.8, 0.8]$, $z \in [0.03, 1.6]$
Point construction	4,096 XYZ points; deterministic farthest-point sampling
Temporal / action	Three observations at 10 Hz; 10-D current-TCP-relative pose and gripper command
Prediction / execution horizon	16 / 6 policy steps
Point encoder	PointNet; 64-D output; LayerNorm
Denoiser / conditioning	U-Net $[256, 512, 1024]$, kernel 5, 8 groups; FiLM in all blocks; 128-D step embedding
Diffusion / inference	DDIM, 100 training steps, sample prediction, squared-cosine noise; DDIM / 16 inference steps
Optimizer / learning rate	AdamW; 10^{-4}
Adam / regularization	$\beta = (0.95, 0.999)$; $\epsilon = 10^{-8}$; weight decay 10^{-6}
Schedule / batch / precision	Cosine; 250-step warmup; 32 / float32
Training / selection	1,000 epochs; seed 42; EMA power 0.75, max decay 0.9999; candidates at epochs 300, 800, and 1,000; reported checkpoint 1,000

Table 12: **Deployed policy configuration ledger.** Every task uses the same 27/3 training/development episode split across its three conditions. DP3 consumes 4,096 XYZ points from deterministic FPS; all reported policies use seed 42 and the final epoch-1,000 checkpoint. Exact run paths and manifests are bound in the formal rollout sidecars.

Task	**Condition**	**Policy input**	**Point construction**	**Split**	**Checkpoint**
Drawer Opening	RGB DP	wrist RGB	—	27/3	1,000
	DP3–Raw	factory depth XYZ	4,096-point FPS	27/3	1,000
	DP3–Active	learned IR XYZ	4,096-point FPS	27/3	1,000
Peg Insertion	RGB DP	wrist RGB	—	27/3	1,000
	DP3–Raw	factory depth XYZ	4,096-point FPS	27/3	1,000
	DP3–Active	learned IR XYZ	4,096-point FPS	27/3	1,000
Block Extraction	RGB DP	wrist RGB	—	27/3	1,000
	DP3–Raw	factory depth XYZ	4,096-point FPS	27/3	1,000
	DP3–Active	learned IR XYZ	4,096-point FPS	27/3	1,000

Learned IR depth uses the frozen UFM-based structured-light model evaluated in this paper, without confidence filtering. All policy demonstration cohorts use the Gemini 335 IR projector at energy level 1. This reduced projector energy mitigates close-range IR overexposure and is shared across policy conditions.

Raw and learned depth come from the same recordings. Our post-tracking pipeline produces one set of motion labels that is shared by every policy in Table 3. Both DP3 conditions share camera transformations, crop, 4,096 XYZ points, deterministic farthest-point sampling, architecture, training budget, and control conventions; the depth source changes in training and live observations. Table 12 records the deployed input, split, and final checkpoint for each condition.

F.2 Motion-source bridge

The cup study uses complete pick-and-place episodes and the same DP3 observation recipe, three-observation history, current-TCP-relative targets, and training/execution settings across the five motion sources in Table 2. Every pose history and future target is regenerated from its selected trajectory. The rigid alignment used to score ATE is never applied to replay or policy execution.

Table 13: **Upper-tail action-relative error.** The 90th percentile across 24 episode-level RMSE values; each cell reports translation (mm) / rotation (°) at the indicated policy horizon.

Motion source	$k = 1$	$k = 3$	$k = 6$	$k = 16$
T265 SDK	10.323 / .271	17.812 / .447	23.500 / .670	38.451 / .893
ORB-SLAM3	3.464 / .216	5.509 / .392	7.140 / .565	10.267 / .768
Fisheye depth	1.753 / .196	3.255 / .367	5.152 / .554	8.313 / .774
Learned IR depth	1.842 / .195	3.350 / .367	5.372 / .553	8.299 / .770

The benchmark contains 30 complete episodes. Each trajectory is aligned to the calibrated Vive camera reference with one scale-fixed SE(3) transform and scored on a frozen timestamp grid. For each offline estimator, full-trajectory translation and rotation RMSE are averaged across three runs before taking the median over usable episodes; recorded reference and SDK sources each contribute one trajectory. ORB-SLAM3 contributes 28 episodes because it does not recover a complete global trajectory containing the first frozen-grid sample in two recordings; these are the same underlying 30 recordings used by every method. Vive, T265 SDK, Fisheye depth, and learned IR depth contribute all 30. Different usable subsets are therefore not visually ranked.

ORB-SLAM3 uses the recorded Stereo–Inertial input and the same initialization adjustment as the seven-recording study. Fisheye and learned IR depth use the OKVIS2-X-based reconstruction described in Appendix C. Projector operation follows each recording cohort and is not an evaluated factor.

The downstream protocol holds the policy configuration fixed and evaluates five direct replays and five policy rollouts per source. Training and development use the common set of episodes for which every required trajectory, action, and gripper stream is available. Relative errors over the policy horizon supplement ATE because a constant global rigid error can cancel under current-relative actions. For the common 24 episodes with complete trajectories from all four estimated motion sources, Table 2 evaluates steps 1, 3, 6, and 16 on the 10 Hz policy grid. Main-table values are medians across episodes of the within-episode RMSE; Table 13 reports the 90th percentiles. No global trajectory alignment is used because the current-relative action already cancels a constant world transform. Table 2 reports successful executions / trials. For five trials, the two-sided 95% Wilson intervals are 3.6–62.4% for 1/5, 23.1–88.2% for 3/5, 37.6–96.4% for 4/5, and 56.6–100.0% for 5/5.

F.3 Execution timing

All formal comparisons use a three-observation history with poses expressed relative to the measured TCP at the current policy step. That TCP pose anchors all sixteen targets from an inference, and the robot executes a six-node prefix. History advances after acknowledged action nodes using synchronized visual observations, measured TCP, and gripper state. Camera frames and servo ticks do not create extra history. The endpoint is held until acknowledgement, and the first observation is repeated until three real nodes exist.

F.4 Complementary modalities

Figure 7 shows interaction frames from dedicated robot replay recordings with the IR projector enabled. Learned depth is inferred offline from the recorded IR pair associated with each RGB frame. Factory and learned depth use the same display scale; their clouds use the recorded calibration, identical spatial crops, and a shared rendering viewpoint. These visualizations retain all finite points within the display crop rather than the 4,096 points sampled for policy input. They do not add trials to the policy success counts.

G Task protocols and outcome definitions

Every task begins before contact. Full-task success includes grasp acquisition; stage counts diagnose failures but never replace the full-attempt denominator. Table 14 gives the operational success criteria. Drawer Opening ends at the specified opening distance and has no closing phase. Block

Extraction uses one designated exposed block and the same prepared tower for all methods. Across the current 30-demonstration cohorts, initial TCP position spans are $18.73 \times 12.44 \times 12.89$ cm for Drawer Opening, $7.15 \times 5.75 \times 8.62$ cm for Peg Insertion, and $2.38 \times 5.53 \times 2.64$ cm for Block Extraction. Maximum pairwise initial orientation differences are $16.6°$, $10.1°$, and $15.1°$, respectively. These TCP spans include the varied robot initialization and are distinct from the bounded object and fixture perturbations.

Table 14: **Task protocols.** All attempts begin before grasping and use the bounded pose variation in Section 5.3. Each criterion evaluates the complete task.

Task	Full-task success criterion	Spatial demand and reset
Drawer Opening	Grasp the handle and pull until the drawer front crosses the fixed opening mark; no closing phase.	Coarse handle acquisition and directed articulated motion. The cabinet starts closed and its pose follows the common perturbation protocol.
Peg Insertion	Pick up, align, insert, and release the peg; the part remains seated after the gripper retracts.	Local mating under partial gripper occlusion. The peg and secured receiver follow the common perturbation protocol.
Block Extraction	Fully separate the designated block while retaining the grasp; every non-target layer remains standing at the episode end.	Narrow-clearance grasping and contact-sensitive withdrawal. The prepared tower and target definition are shared across policies.

Scheduling and uncertainty Policies are evaluated in interleaved order under the same bounded pose perturbations and reset protocol. Each of the five motion-source conditions uses five direct replays and five policy rollouts. No formal robot outcome is reported before its evaluation cohort is complete. Development rollouts select checkpoints; formal outcomes do not. Final reporting uses successes/actual attempts and two-sided 95% Wilson binomial confidence intervals for each condition. The exact checkpoint, training manifest, feature contract, depth source, and execution settings are stored with every formal rollout sidecar. Training uses seed 42; rollout repetitions are physical trials rather than training seeds. Aborted attempts remain failures unless a predefined apparatus-fault rule applies; exclusions are counted and explained.

Figure 12 shows synchronized examples from two formal cohorts. All reported manipulation outcomes, including Peg Insertion, are bound to timestamped rollout sidecars and their checkpoint, feature contract, and evaluation label.

Table 15: **Manipulation outcomes with uncertainty.** Cells report successes/attempts, observed success rate, and two-sided Wilson 95% confidence interval in percentage points.

Policy	Drawer Opening	Peg Insertion	Block Extraction
RGB DP	7/10, 70% [39.7, 89.2]	1/10, 10% [1.8, 40.4]	1/10, 10% [1.8, 40.4]
DP3–Raw	5/10, 50% [23.7, 76.3]	0/10, 0% [0.0, 27.8]	1/10, 10% [1.8, 40.4]
DP3–Active	8/10, 80% [49.0, 94.3]	7/10, 70% [39.7, 89.2]	5/10, 50% [23.7, 76.3]

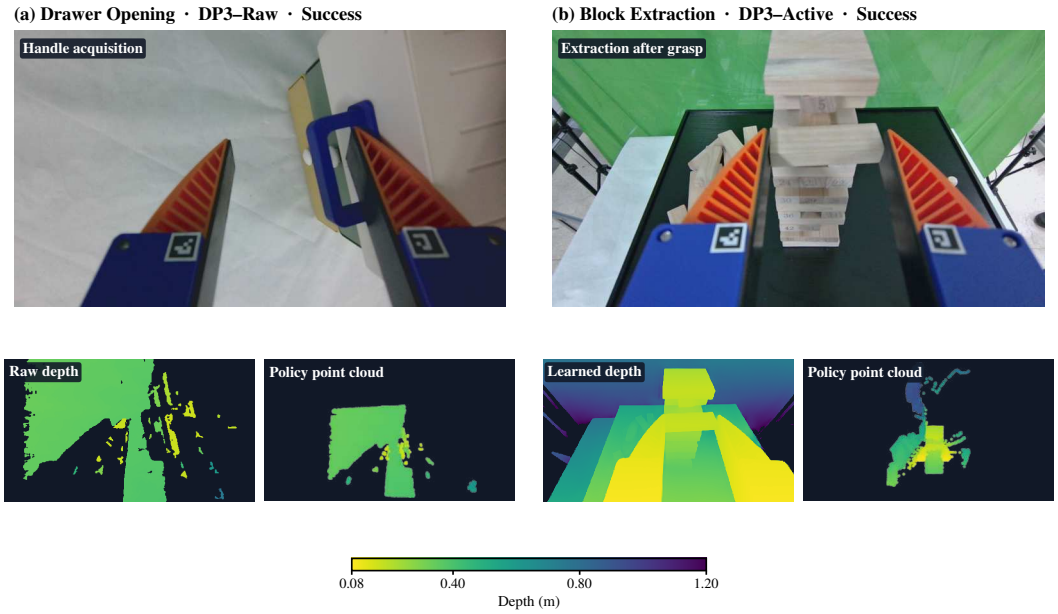

Figure 12: **Traceable policy-outcome stages.** Synchronized RGB, observed depth, and the 4,096-point policy input are shown for one formal Drawer Opening rollout and one formal Block Extraction rollout. Each panel documents a single recorded stage and does not by itself establish a general success or failure mechanism.

H Reproducibility and Artifact Organization

Each HDF5 episode retains raw camera, IMU, and gripper streams with acquisition timestamps. Derived depth and trajectories preserve source modality and calibration identity; LeRobot is an interoperability export. Experiment records retain tracking inputs, configuration snapshots, build and command histories, trajectories, and all attempted outcomes, with evaluation cohorts kept separate. Policy artifacts bind checkpoints to data splits, label/depth sources, point processing, reference poses, history, and execution settings (Appendix F).