PHYSICALLY CONSISTENT MULTIMODAL REPRESENTATIONS FOR PERSISTENT TARGET TRACKING AND CAPTURE WITH LEGGED MANIPULATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent target tracking and capture in unstructured environments require physically grounded multimodal representations that preserve object identity and support interaction under uncertain observations. We present a unified framework that integrates physical consistency into multimodal representation, persistent target reasoning, and coordinated capture for legged manipulators. First, synchronized visual, thermal, and radar observations are organized into a pixel-aligned physical representation for constructing metric-semantic maps and estimating target states. Subsequently, an acquire–maintain–recover mechanism combines semantic compatibility with geometric, motion, and thermal consistency to preserve target identity under occlusion and visual degradation. The resulting representation and target confidence guide coordinated locomotion and manipulation during approach, alignment, and capture. Extensive real-robot experiments cover indoor and outdoor environments, day and night conditions, and four target categories. The framework achieves 91.3% capture success across 480 episodes per method and 84.8% post-occlusion recovery, compared with 76.7% and 61.9% for the strongest baseline. Component ablations further demonstrate complementary contributions from physical representation, semantic mapping, persistent reasoning, and coordinated execution.

1 INTRODUCTION

Legged manipulators combine terrain-adaptive mobility with physical interaction, enabling retrieval beyond fixed workspaces (Liu et al., 2025a; Pan et al., 2025; Ha et al., 2025). Capturing a language-specified target requires identifying the commanded instance, tracking its motion, navigating obstacles, and coordinating the base and arm. Viewpoint changes, occlusion, and target dynamics can disrupt perception and invalidate grasping. Therefore, successful capture requires a representation that jointly preserves target identity, physical state, and actionable information throughout the interaction process.

Self-supervised features, language–image alignment, and multimodal embeddings have advanced semantic transfer (Oquab et al., 2024; Zhai et al., 2023; Girdhar et al., 2023). Open-vocabulary detection, promptable segmentation, and language-guided 3D representations further improve robotic understanding and manipulation (Cheng et al., 2024b; Liu et al., 2024; Kirillov et al., 2023; Huang et al., 2023b; Goyal et al., 2023; Ze et al., 2024). However, existing representations mainly focus on semantic recognition or spatial understanding, leaving open how heterogeneous observations can preserve physically meaningful correspondence during persistent pursuit and capture.

RGB provides semantic detail, while thermal and radar sensing offer complementary contrast, range, and motion evidence. Existing fusion approaches exploit these modalities (Zhang et al., 2023a; Kim et al., 2023; Lin et al., 2024), but heterogeneous timestamps, resolutions, and spatial supports may introduce inconsistent associations. Although visual memory and point trajectories improve temporal continuity (Cheng et al., 2024a; Karaev et al., 2025), reliable capture additionally requires validating instance identity and maintaining accurate target states for physical interaction (Ma et al., 2025; Liang et al., 2026).

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.

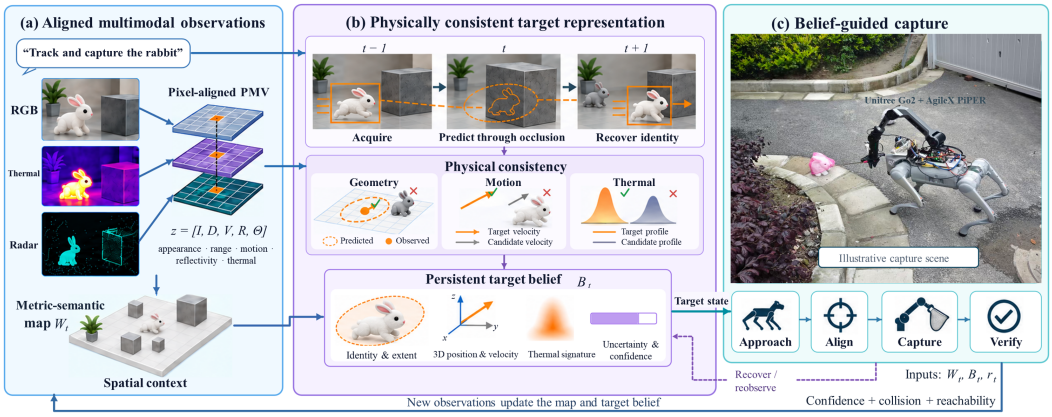


Figure 1: Framework overview. Pixel-aligned observations support metric-semantic mapping and persistent target association. Target confidence and spatial constraints guide coordinated approach, capture, and verification.

We propose *physically consistent multimodal representations*, where consistency denotes calibrated cross-sensor correspondence and temporal compatibility among geometric, radial-motion, and thermal evidence. A Pixel-Aligned Multimodal Vector (PMV) is constructed from synchronized RGB, FIR, and 4D mmWave radar observations. Rather than feature-level fusion, PMV organizes heterogeneous measurements into a physically grounded representation space, preserving spatial correspondence, measurement validity, and modality-specific physical attributes (Zhen et al., 2024; Higuera et al., 2025). It supports metric-semantic mapping and persistent target states containing instance extent, position, velocity, thermal statistics, and association confidence.

An acquire–maintain–recover mechanism preserves identity through semantic and physical validation. Acquisition integrates semantic compatibility with physical coherence, maintenance validates propagated states using geometric and sensor cues, and recovery constrains reassociation through prior target states. The resulting representation enables confidence-guided approach, alignment, and capture through coordinated base–arm execution. We evaluate the framework on a quadruped manipulator across diverse environments and conditions, including 480 capture episodes per method and a 640-trial occlusion protocol. To summarize, our contributions are threefold:

- A **physically grounded pixel-aligned multimodal representation** is introduced to preserve cross-modal correspondence, measurement validity, and physical attributes for metric-semantic mapping and persistent target-state estimation.
- We develop a **physical-consistency mechanism for persistent target reasoning** by integrating semantic association with geometric, motion, and thermal constraints for identity maintenance and recovery under degraded observations.
- A **representation-driven embodied perception-to-action framework** is established to connect physically consistent target states with confidence-guided base–arm coordination and capture execution.

2 Related Work

2.1 Multimodal Representations for Embodied Interaction

VLMaps grounds language in metric navigation maps for spatial goals and obstacle-aware navigation (Huang et al., 2023a); ConceptFusion and ConceptGraphs support multimodal querying and spatial relationships (Jatavallabhula et al., 2023; Gu et al., 2024). OpenScene, LERF, and LangSplat enable open-vocabulary 3D understanding through point, radiance-field, and Gaussian representations (Peng et al., 2023; Kerr et al., 2023; Qin et al., 2024). GO-SLAM globally optimizes poses and reconstruction (Zhang et al., 2023b), SplaTAM uses Gaussians for tracking and mapping (Keetha

et al., 2024), and PanopticNDT integrates panoptic information into spatial maps (Seichter et al., 2023). BEVFusion and UniTR construct shared bird's-eye-view features for detection and map segmentation (Liu et al., 2023; Wang et al., 2023a), while CRN and RCBEVDet fuse camera–radar observations for 3D detection (Kim et al., 2023; Lin et al., 2024). CMX studies complementary sensing, including thermal observations, for segmentation (Zhang et al., 2023a), and MetaBEV addresses corrupted or missing inputs (Ge et al., 2023). SparseDFF learns view-consistent features for manipulation transfer (Wang et al., 2024), whereas CORN uses contact-informed pretraining for nonprehensile manipulation (Cho et al., 2024). PMV exposes calibrated correspondence and channel validity as shared evidence for mapping and persistent target association.

2.2 Persistent Object Representations and Tracking

DINOSAUR learns real-world object-centric representations from self-supervised features without object labels (Seitzer et al., 2023), and SlotFormer autoregressively predicts learned object states (Wu et al., 2023). SeqTrack predicts bounding-box sequences (Chen et al., 2023), while ViPT and OneTracker adapt foundation representations across modalities through prompting (Zhu et al., 2023; Hong et al., 2024). DEVA decouples segmentation and temporal propagation (Cheng et al., 2023), Cutie introduces object-level memory (Cheng et al., 2024a), and SAM 2 supports promptable image/video segmentation (Ravi et al., 2025). TAPIR and CoTracker estimate point trajectories (Doersch et al., 2023; Karaev et al., 2024); CoTracker3 improves robustness using pseudo-labeled real videos (Karaev et al., 2025). OmniMotion models globally consistent motion through a canonical volume (Wang et al., 2023b), while SpatialTracker and SpatialTrackerV2 exploit 3D structure and, in the latter, explicit camera motion (Xiao et al., 2024; 2025). We complement tracking with metric, radar-motion, and thermal validation of commanded-instance identity, constraining recovery through predicted state.

2.3 Coordinated and Dynamic Legged Manipulation

Visual Whole-Body Control and RoboDuet coordinate locomotion and manipulation through hierarchical or collaborative policies (Liu et al., 2025a; Pan et al., 2025). Kinematic guidance incorporates workspace information into policy learning (Hou et al., 2025). Learned force control enables compliant interaction without requiring direct force sensing (Portela et al., 2024), while guided reinforcement learning uses trajectory-optimization demonstrations for multi-contact coordination (Sleiman et al., 2025). UMI-on-Legs connects manipulation policies to whole-body trajectory tracking (Ha et al., 2025), and WholeBodyVLA integrates latent action learning with humanoid locomotion and manipulation, using action-free videos to supplement scarce robot demonstrations (Jiang et al., 2026). Coordinated badminton integrates perception, prediction, and control (Ma et al., 2025), while DQ-Net uses temporal perception for dynamic grasping (Liang et al., 2026). We emphasize persistent target-state reliability, linking identity validation and recovery to confidence-dependent approach, alignment, and capture under collision and reachability constraints.

3 Method

Problem formulation. Given a language query q, multimodal observations o_t, and robot state r_t, a legged manipulator must track and capture the specified instance despite motion, occlusion, and degraded visibility. Observations include RGB, far-infrared (FIR), and 4D mmWave radar, with auxiliary LiDAR where specified. We maintain a metric-semantic map $\mathcal{W}_t$ and target state

$$\mathcal{B}_t = (M_t, \mathbf{p}_t, \mathbf{v}_t, \Sigma_t, \bar{\Theta}_t, c_t), \tag{1}$$

comprising the instance mask, world-frame position and Cartesian velocity, spatial dispersion, a robust thermal statistic, and association confidence. During observation gaps, the state supports prediction. Dispersion and confidence are operational indicators without assumed probabilistic calibration. Coordinated base–arm actions use $(\mathcal{W}_t, \mathcal{B}_t, r_t)$ to capture the commanded instance within $T_{\max}$ and retain it for at least T_{hold}, while respecting collision, kinematic, and safety constraints.

Figure 1 presents three components: pixel-aligned multimodal representation, physical consistency for persistent tracking, and confidence-guided capture. Physical consistency combines calibrated cross-sensor correspondence with geometric, radial-motion, and thermal compatibility over time.

The resulting map and target state connect identity-preserving association to feasible approach, alignment, and capture (Jiang et al., 2026; Liu et al., 2025b).

3.1 Pixel-Aligned Multimodal Representation

Spatial and temporal alignment. A calibrated timestamp offset places sensor s on a common clock, $t = t^{(s)} + \Delta^{(s)}$. A residual offset δt introduces an approximate displacement $\|\delta \mathbf{x}\| \approx \|\mathbf{v}\|\,|\delta t|$, which can corrupt cross-modal correspondence for moving targets. We therefore synchronize measurements before association and transform radar returns into the RGB camera frame:

$$\mathbf{x}^c = R_{c\leftarrow r}\mathbf{x}^r + \mathbf{t}_{c\leftarrow r}, \qquad \mathbf{u} = \pi(K\mathbf{x}^c), \quad (2)$$

where K is the camera intrinsic matrix and π denotes homogeneous normalization. FIR observations are registered to the same pixel lattice through calibrated cross-camera mapping.

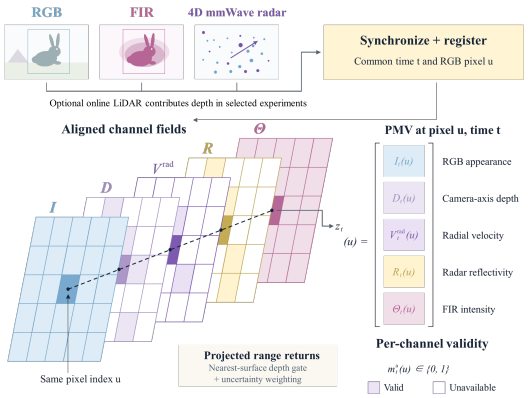


Figure 2: The structure of the proposed PMV.

Each pixel $\mathbf{u}$ then indexes a Pixel-Aligned Multimodal Vector (PMV):

$$\mathbf{z}_t(\mathbf{u}) = \left[I_t(\mathbf{u}), D_t(\mathbf{u}), V_t^{\mathrm{rad}}(\mathbf{u}), R_t(\mathbf{u}), \Theta_t(\mathbf{u})\right], \quad (3)$$

containing RGB appearance, camera-axis depth, radar radial velocity, reflectivity, and thermal evidence. Channel validity and measurement uncertainty accompany this representation, allowing downstream processing to distinguish unavailable measurements from observed values. Physical channels can remain sparse despite sharing the RGB lattice. When multiple returns project to one pixel, a nearest-surface depth gate suppresses inconsistent background contributions, followed by uncertainty-weighted aggregation of retained returns (Appendix A).

Shared metric-semantic context. A semantic module operates on the PMV field and produces per-pixel class probabilities $p_t(c \mid \mathbf{u}) = [\mathrm{softmax}(f_\theta(\mathbf{z}_t))]_{\mathbf{u},c}$. For pixels with valid depth, the corresponding world-frame point is

$$\mathbf{x}_t^w(\mathbf{u}) = R_{w\leftarrow c}(t) D_t(\mathbf{u}) K^{-1}\tilde{\mathbf{u}} + \mathbf{t}_{w\leftarrow c}(t), \quad (4)$$

where $\tilde{\mathbf{u}} = (u, v, 1)^\top$. Sensor range is converted to camera-axis depth before back-projection. Successive observations are integrated using the robot pose to construct $\mathcal{W}_t$. Geometry and semantic evidence are accumulated in a voxel map, a Euclidean signed distance field supports collision queries, and local surfels provide manipulation geometry. The resulting common reference links target observations to scene geometry, free space, and candidate approach configurations.

3.2 Physical Consistency for Persistent Target Tracking

Acquire. Language-conditioned proposals $(b_{k,t}, s_{k,t})$ provide candidate boxes and semantic relevance scores, which are refined into instance masks $M_{k,t}$ (Cheng et al., 2024b; Kirillov et al., 2023). To assess whether a mask groups physically coherent observations, we compute median absolute deviation (MAD) over its valid depth and radial-velocity support:

$$\begin{aligned} \ell_{k,t}^{\mathrm{coh}} &= -\frac{\mathrm{MAD}(D_t \mid M_{k,t})^2}{\sigma_D^2} - \frac{\mathrm{MAD}(V_t^{\mathrm{rad}} \mid M_{k,t})^2}{\sigma_V^2}, \\ k^* &= \arg\max_k \left[\lambda_s s_{k,t} + \lambda_p \ell_{k,t}^{\mathrm{coh}} + \lambda_\Theta \psi_\Theta(\bar{\Theta}_{k,t})\right]. \end{aligned} \quad (5)$$

The scales σ_D and σ_V normalize the physical cues, while ψ_Θ supplies complementary thermal evidence when available. Only valid measurements contribute to these terms. The highest-scoring candidate initializes a track only after passing semantic and physical acceptance checks. Its position is the coordinate-wise median of valid world-frame points, Σ_t summarizes their dispersion, and $\bar{\Theta}_t$ records a robust thermal statistic. Cartesian velocity is subsequently estimated from smoothed inter-frame displacement. This velocity describes target motion in the world frame and is distinct from the radar's radial measurement.

Maintain. A visual tracker propagates the previous mask, $\widetilde{M}_t = \mathcal{T}(I_t, I_{t-1}, M_{t-1})$ (Cheng et al., 2024a). We validate this hypothesis against the predicted target state, $\widehat{\mathbf{p}}_t = \mathbf{p}_{t-1} + \mathbf{v}_{t-1}\Delta t$ and $\widehat{\Sigma}_t = \Sigma_{t-1} + Q_t$, where Q_t accounts for uncertainty growth between updates. For hypotheses with sufficient valid physical support, compatibility is measured by $\mathrm{L}_t = \exp\left[-\frac{\|\widetilde{\mathbf{p}}_t - \widehat{\mathbf{p}}_t\|^2}{\sigma_{p,t}^2} - \gamma_v \frac{(\Delta V_t^{\mathrm{rad}})^2}{\sigma_v^2} - \gamma_\Theta \frac{(\Delta\Theta_t)^2}{\sigma_\Theta^2}\right]$, where $\sigma_{p,t}$ is a spatial tolerance derived from the uncertainty proxy. The motion residual compares observed radial velocity with the predicted target motion expressed in the current radar line of sight, accounting for sensor ego-motion. The thermal residual measures deviation from the maintained thermal statistic. Unavailable cues are omitted, and hypotheses lacking the physical support required for localization are deferred to recovery.

We combine physical compatibility with segmentation confidence:

$$c_t = \alpha_t c_t^{\mathrm{seg}} + (1 - \alpha_t) L_t, \qquad \alpha_t = \mathrm{clip}_{[0,1]}\left(\frac{\eta_t - \eta_{\min}}{\eta_{\max} - \eta_{\min}}\right), \tag{6}$$

where η_t measures visual reliability and $\eta_{\max} > \eta_{\min}$. As visual reliability decreases, physical evidence receives greater weight. Accepted hypotheses update $\mathcal{B}_t$; rejected hypotheses retain the predicted state and trigger recovery, preventing unconditional identity transfer from a propagated mask.

Recover. When $c_t < \tau_c$, or a valid tracking update is unavailable, the original query generates new candidate masks. With $\rho_{k,t} = \lambda_p \ell_{k,t}^{\mathrm{coh}} + \lambda_\Theta \psi_\Theta(\bar{\Theta}_{k,t})$, re-acquisition ranks candidates as

$$k^* = \arg\max_k \left[\lambda_s s_{k,t} + \rho_{k,t} - \lambda_g \frac{\|\mathbf{p}(M_{k,t}) - \widehat{\mathbf{p}}_t\|^2}{\sigma_{p,t}^2}\right]. \tag{7}$$

Semantic relevance preserves compatibility with the command, physical coherence validates the candidate's support, and predicted proximity constrains inheritance of the existing identity. Re-acquisition occurs only when the selected candidate passes acceptance checks; otherwise the target remains unresolved and capture is inhibited. This prior reduces implausible reassociation, although prolonged occlusion and nearby similar instances can remain ambiguous.

3.3 Confidence-Guided Capture

Approach and alignment. The map and target state jointly determine how the robot approaches the instance. Candidate base configurations are evaluated using

$$J_{\mathrm{app}} = w_g C_g + w_s C_s + w_v C_{\mathrm{vis}} + w_m C_{\mathrm{man}}, \tag{8}$$

subject to collision and kinematic constraints. Geometry and semantic costs penalize hazardous or non-traversable regions, the visibility cost evaluates target observability, and the manipulability cost uses a precomputed reachability volume. Coordinated base–arm alignment then reduces end-effector and base-pose errors under the same feasibility constraints. This couples approach planning to the visibility and manipulation requirements of the maintained target state.

Capture and verification. For quasi-static targets, feasible grasps are ranked using local geometry. For moving targets, the short-horizon prediction $\widehat{\mathbf{p}}_t(\Delta) = \mathbf{p}_t + \mathbf{v}_t\Delta$ supports interception:

$$\Delta^* = \arg\min_{\Delta \in \mathcal{F}_t} \|\mathbf{p}^{\mathrm{ee}}(t + \Delta) - \widehat{\mathbf{p}}_t(\Delta)\|^2, \tag{9}$$

where $\mathbf{p}^{\mathrm{ee}}(t + \Delta)$ is the end-effector position under the candidate coordinated trajectory, and $\mathcal{F}_t$ contains interception times satisfying the planning horizon, collision, and kinematic constraints. Receding-horizon feedback revises the plan as new observations update $\mathcal{B}_t$. Primitive selection additionally considers target coverage and environmental costs; coverage is an operational geometric score (Appendix C).

Confidence regulates execution with $\tau_a > \tau_c$. The controller invokes recovery when $c_t < \tau_c$, continues approach or alignment when $\tau_c \le c_t < \tau_a$, and permits capture at $c_t \ge \tau_a$ only after feasibility checks. Identity and retention verification determine whether the capture satisfies the task criterion. An unsuccessful attempt triggers re-alignment or recovery, completing the feedback loop from physically validated observations to action and renewed observation.

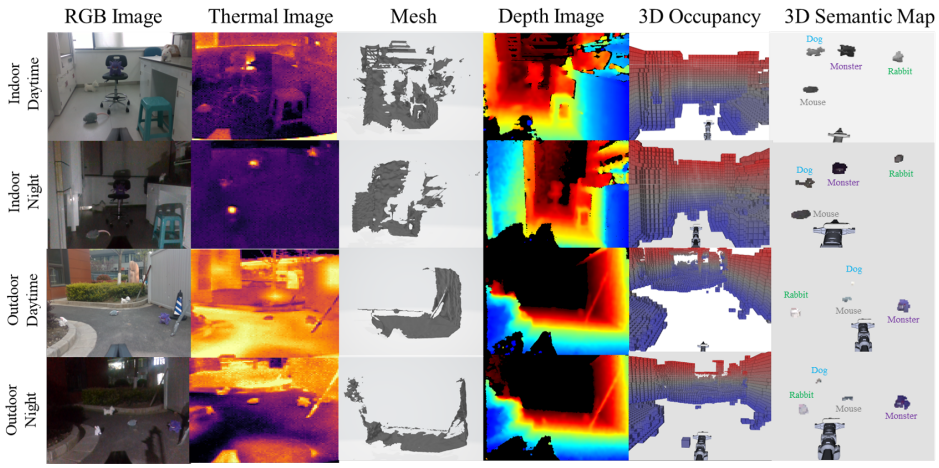


Figure 3: Multimodal observations and reconstructed spatial representations across indoor/outdoor and normal/low illumination. Aligned thermal and range observations provide complementary support for target extent, scene geometry, and semantic structure when RGB evidence weakens.

Table 1: Mapping performance under indoor and outdoor low-light conditions (mean ± standard deviation). Normal-illumination results appear in Appendix E.

Condition	Method	3D mIoU (%) ↑	Chamfer-L1 (cm) ↓	Target recall (%) ↑	Free-space F1 (%) ↑	Latency (ms) ↓
Indoor low-light	RGB-only	50.4 ± 4.2	12.1 ± 1.3	52.7 ± 4.7	65.9 ± 3.8	31.4 ± 1.9
	RGB-D	62.5 ± 3.6	8.4 ± 0.9	65.2 ± 4.1	74.8 ± 3.2	38.7 ± 2.1
	Late fusion	72.9 ± 3.1	6.2 ± 0.7	78.7 ± 3.5	82.8 ± 2.7	57.2 ± 3.0
	Ours	$\mathbf{87.3 \pm 2.1}$	$\mathbf{3.4 \pm 0.4}$	$\mathbf{92.9 \pm 2.0}$	$\mathbf{91.9 \pm 1.8}$	$\mathbf{30.6 \pm 1.8}$
Outdoor low-light	RGB-only	42.3 ± 4.7	16.0 ± 1.7	43.8 ± 5.0	59.1 ± 4.1	32.0 ± 2.0
	RGB-D	55.1 ± 4.0	11.8 ± 1.2	59.1 ± 4.3	68.8 ± 3.5	39.3 ± 2.3
	Late fusion	68.2 ± 3.5	8.6 ± 0.9	73.3 ± 3.8	77.9 ± 3.1	58.8 ± 3.4
	Ours	$\mathbf{83.8 \pm 2.5}$	$\mathbf{4.7 \pm 0.5}$	$\mathbf{89.4 \pm 2.4}$	$\mathbf{88.7 \pm 2.1}$	$\mathbf{32.1 \pm 2.0}$

4 Experiments

4.1 Experimental Setup

Experiments use a Unitree Go2 quadruped with a PiPER arm, RGB and FIR cameras, and 4D mmWave radar. PMV perception runs at least 30 Hz, planning at 10 Hz, and low-level control at 200 Hz. LiDAR provides reference geometry for mapping and is used as auxiliary ranging in selected experiments. Four inanimate targets (mouse, rabbit, dog, and monster) are evaluated across indoor/outdoor and day/night conditions, with clutter, obstacles, and uneven terrain. Rabbit and dog targets additionally move along prescribed trajectories at 0.2–1.0 m/s. Eight randomized layouts vary starting poses, obstacles, and motion directions. Such diversity complements recent efforts emphasizing broad scene, task, and embodiment coverage in robot evaluation (O'Neill et al., 2024).

Mapping is compared with RGB-only, RGB-D, and late multimodal fusion. Tracking and capture use task-adapted KineMan-RL (Hou et al., 2025), BirdieRL (Ma et al., 2025), and DQ-Net (Liang et al., 2026) on the same platform. These names refer to the evaluated adaptations. Capture uses 30 episodes per target–condition pair, totaling 480 per method; tracking uses a separate 640-trial controlled-occlusion protocol. All failed trials remain in the rate denominators, and timing conditional on success is identified explicitly Khazatsky et al. (2024); Nasiriany et al. (2024).

4.2 Quality of the Multimodal Representation

We first evaluate 3D semantic mIoU, Chamfer-L1 distance to reference geometry, target recall, free-space F1, and latency, consistent with evaluating whether multimodal representations preserve task-relevant physical information beyond a single modality (Mejia et al., 2024). Figure 3 links the sens-

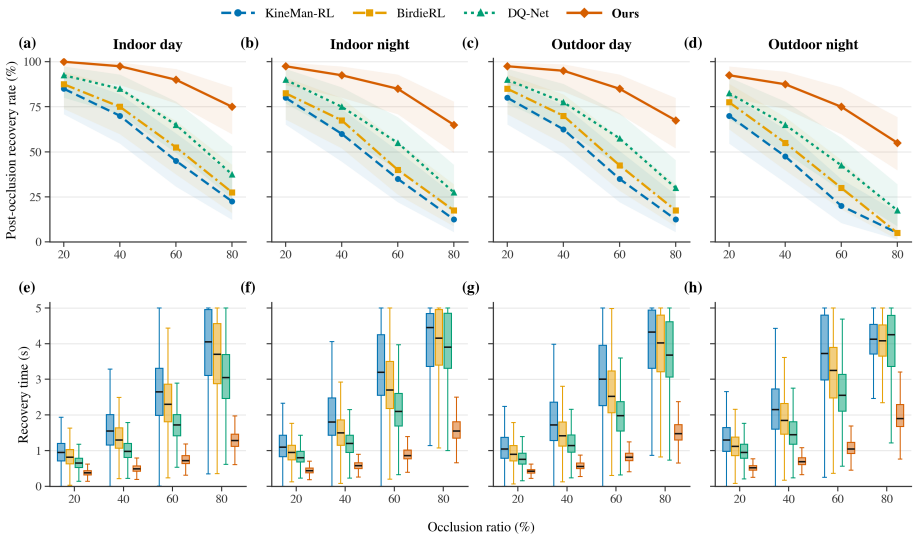


Figure 4: Occlusion recovery across four conditions. (a–d) Recovery rates with 95% Wilson intervals ($n = 40$ per point). (e–h) Recovery times for successful trials; the dashed line marks 5 s.

Table 2: Pooled tracking results over 640 trials per method. Rates are percentages; recovery time is the median [IQR] over successful re-acquisitions.

Method	Persistent lock	Post-occlusion recovery	Recovery time (s)	Identity switch	Confidence AUC
KineMan-RL	414/640 (64.7 [60.9–68.3])	297/640 (46.4 [42.6–50.3])	1.60 [1.11–2.53]	121/640 (18.9 [16.1–22.1])	0.54 ± 0.13
BirdieRL	455/640 (71.1 [67.5–74.5])	333/640 (52.0 [48.2–55.9])	1.38 [0.99–2.17]	97/640 (15.2 [12.6–18.1])	0.60 ± 0.12
DQ-Net	516/640 (80.6 [77.4–83.5])	396/640 (61.9 [58.1–65.6])	1.18 [0.83–1.94]	62/640 (9.7 [7.6–12.2])	0.69 ± 0.10
Ours	**595/640 (93.0 [90.7–94.7])**	**543/640 (84.8 [81.9–87.4])**	**0.65 [0.48–1.01]**	**18/640 (2.8 [1.8–4.4])**	$\mathbf{0.86 \pm 0.06}$

ing modalities to their reconstructed spatial outputs, while Table 1 summarizes semantic accuracy. Full geometric, recall, free-space, and timing results are provided in Appendix E.

Table 1 highlights representation quality under low illumination. Our method achieves 87.3% mIoU indoors and 83.8% outdoors, exceeding late fusion by 14.4 and 15.6 percentage points, respectively. Chamfer-L1 decreases from 6.2 to 3.4 cm indoors and from 8.6 to 4.7 cm outdoors. Outdoor target recall increases from 73.3% to 89.4%, while free-space F1 reaches 91.9% indoors and 88.7% outdoors. Processing latency is 30.6–32.1 ms, compared with 57.2–58.8 ms for late fusion. These results indicate that the multimodal representation supports both semantic recognition and the spatial information needed for target association and approach planning when visual evidence weakens.

4.3 Persistent Tracking Under Occlusion

To test identity persistence independently of capture, we vary target occlusion over 20%, 40%, 60%, and 80% of projected area. Recent video-segmentation systems similarly identify occlusion and object reappearance as central challenges for long-term identity preservation (Li et al., 2024; Ding et al., 2025). Four targets, four conditions, four occlusion levels, and ten repetitions yield 640 trials per method. A lock is lost if identity is absent or confidence stays below threshold for more than 0.5 s. Recovery requires the correct identity and 3D state to be restored within 5 s and retained for ten frames, regardless of whether formal lock loss occurred.

Table 2 separates continuous lock from post-occlusion recovery. The proposed method improves these rates by 12.3 and 23.0 percentage points over DQ-Net, while reducing identity switches from 9.7% to 2.8%. Figure 4 resolves this comparison by condition and severity: pooled recovery is 96.9%, 93.1%, 83.8%, and 65.6% as occlusion increases, versus 88.8%, 75.6%, 55.0%, and 28.1% for DQ-Net. The corresponding margin reaches 37.5 percentage points at 80% occlusion. Conditional median recovery time rises from 0.43 to 1.53 s for our method, versus 0.78 to 3.60 s for DQ-Net. Jointly, the rates and timing indicate both more frequent and faster recovery; timing alone would conceal failed re-acquisitions. Additional pursuit sequences appear in Appendix F.

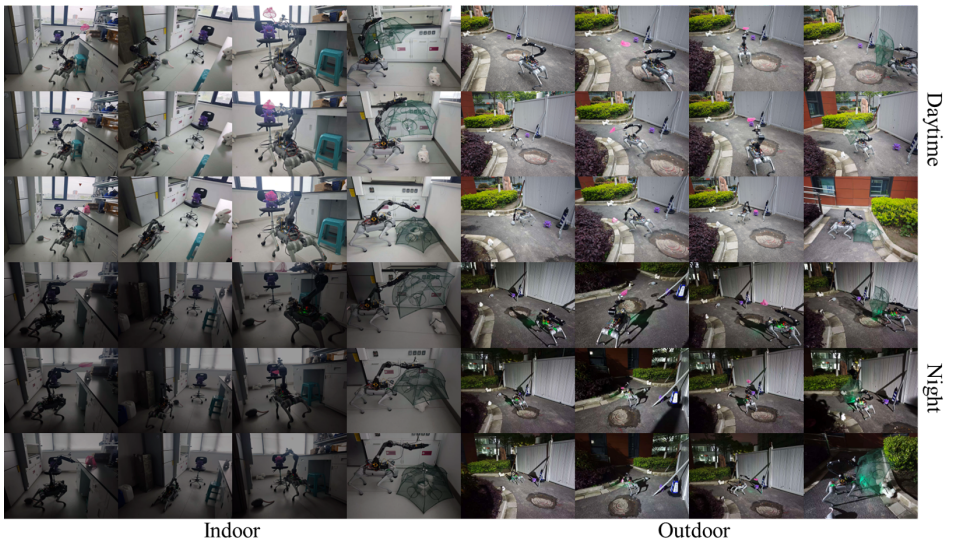


Figure 5: Capture sequences for targets of different scales. Left/right blocks show indoor/outdoor trials; upper/lower blocks show day/night conditions.

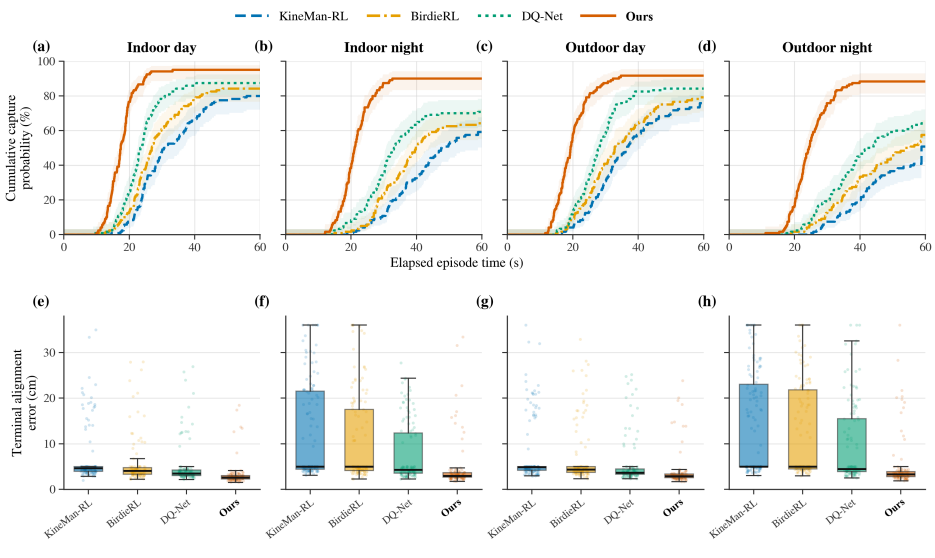
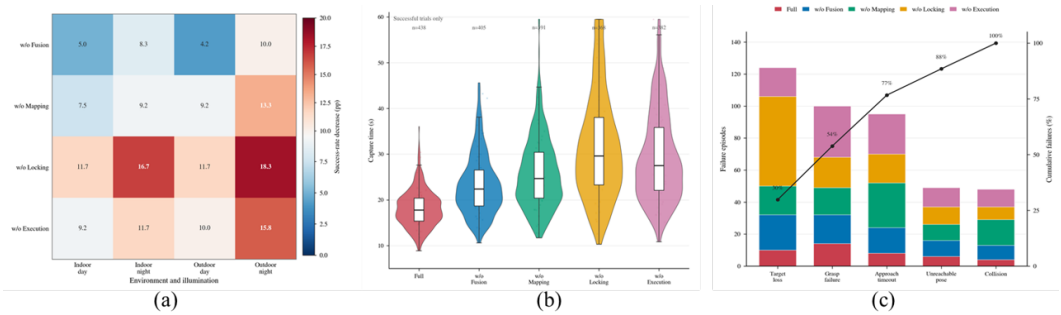


Figure 6: Capture performance across four conditions. (a–d) Cumulative success rates over time ($n = 120$ per method per condition, including failures); endpoints indicate final success rates. (e–h) Terminal alignment-error distributions.

4.4 Closed-Loop Target Capture

We evaluate whether multimodal representations and persistent tracking support capture. This end-to-end setting complements recent real-robot studies of generalist policies and whole-body manipulation under diverse execution conditions. Success requires capturing the commanded instance within 60 s, retaining it for at least 3 s, and avoiding collisions or safety-limit violations. Our method achieves 438/480 successful captures (91.3%; 95% Wilson interval: 88.4–93.5%), versus 76.7% for DQ-Net, 71.3% for BirdieRL, and 66.5% for KineMan-RL. Improvements over DQ-Net reach 14.6 percentage points overall and 24.2 outdoors at night, consistent with complementary sensing and persistent association supporting capture under degraded visual observations.



(a) (b) (c)

Figure 7: Module-level ablations. (a) Capture-success decrease relative to the full system across conditions. (b) Capture times for successful episodes, with medians and interquartile ranges. (c) Failure-mode counts across configurations and their cumulative proportion.

Table 3: Target-wise capture ablations: successes/120 (%) per target and pooled results ($n = 480$) with 95% Wilson intervals. Δ is the success-rate decrease from the full system; P_{adj} uses two-sided Fisher's exact tests against it with Holm correction over four comparisons.

Configuration	Mouse	Rabbit	Dog	Monster	Overall (95% CI)	Δ (pp)	P_{adj}
Ours (Full)	**105/120 (87.5)**	**108/120 (90.0)**	**112/120 (93.3)**	**113/120 (94.2)**	**438/480, 91.3 [88.4–93.5]**	–	–
w/o Fusion	95/120 (79.2)	99/120 (82.5)	105/120 (87.5)	106/120 (88.3)	405/480, 84.4 [80.9–87.3]	6.9	0.0015
w/o Mapping	90/120 (75.0)	95/120 (79.2)	102/120 (85.0)	104/120 (86.7)	391/480, 81.5 [77.7–84.7]	9.8	2.56×10^{-5}
w/o Locking	81/120 (67.5)	88/120 (73.3)	98/120 (81.7)	101/120 (84.2)	368/480, 76.7 [72.7–80.2]	14.6	2.94×10^{-9}
w/o Execution	87/120 (72.5)	92/120 (76.7)	100/120 (83.3)	103/120 (85.8)	382/480, 79.6 [75.7–82.9]	11.7	1.10×10^{-6}

Figure 5 illustrates execution, while Fig. 6 shows cumulative success and terminal alignment errors. Our method accumulates successful captures faster and exhibits lower median errors and narrower interquartile ranges. These results indicate that maintained target information supports timely approach and coordinated alignment. Detailed target-wise outcomes appear in Appendix E.

4.5 Component Ablations and Failure Analysis

We remove one module at a time using the same 480-episode capture protocol: late fusion replaces pixel alignment, geometry-only occupancy replaces semantic mapping, frame-wise detection replaces persistent tracking, and shortest-path navigation with independent arm control replaces coordinated execution Arm et al. (2024). Table 3 summarizes the overall effect, while Fig. 7 separates operating conditions, timing, and failure modes.

Removing persistent tracking produces the largest success decrease and longest median capture time. Removing mapping or coordinated execution also causes substantial losses, showing that identity alone is insufficient without feasible approach and alignment. Across the five configurations, target loss, grasp failure, and approach timeout account for 319 of 416 failures (76.7%). The full system has 42 failures, compared with 112 without persistent tracking. These module-level results support the coupled representation–tracking–execution design, while leaving the individual effects of position, motion, and thermal consistency to finer-grained ablations.

5 Conclusion

Physically consistent multimodal representations were developed for persistent target reasoning and capture with legged manipulators. Pixel alignment preserves cross-modal correspondence, temporal consistency supports identity maintenance and recovery, and confidence-guided control links target states to coordinated interaction. Real-robot experiments achieve 91.3% capture success and 84.8% post-occlusion recovery, while ablations support complementary roles of representation, mapping, persistent reasoning, and coordinated execution. Limitations include sensitivity to calibration and pose errors, sparse measurements, abrupt target motion, uncalibrated confidence, and incomplete isolation of individual consistency cues. Future work will improve robustness to sensing errors and evaluate generalization to unseen object categories and motion regimes.

Reproducibility Statement

The main text describes the multimodal representation, tracking scores, control interface, and evaluation protocols. The appendix provides additional equations, detailed mapping metrics, target-wise outcomes, and qualitative sequences. Supplementary materials include a reference implementation, evaluation scripts, experimental result tables, demonstration videos, and documentation of the hardware and evaluation protocols. These materials are available in the anonymous repository: https://anonymous.4open.science/r/iclr2027-pmv-capture-supplement-2CBE/.

Statement on the Use of Large Language Models

A large language model assisted with manuscript organization, language editing, notation, reference formatting, and preparation of the LaTeX source. It did not conduct experiments or generate new experimental measurements. The authors are responsible for the technical content and final verification.

References

Philip Arm, Mayank Mittal, Hendrik Kolvenbach, and Marco Hutter. Pedipulate: Enabling Manipulation Skills using a Quadruped Robot's Leg. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5717–5723. IEEE, 2024.

Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14572–14581, 2023.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking Anything with Decoupled Video Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326, 2023.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the Object Back into Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161, 2024a.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911. IEEE, 2024b.

Yoonyoung Cho, Junhyek Han, Yoontae Cho, and Beomjoon Kim. CORN: Contact-based Object Representation for Nonprehensile Manipulation of General Unseen Objects. In *The Twelfth International Conference on Learning Representations*, 2024.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. SAM2Long: Enhancing SAM 2 for Long Video Segmentation with a Training-Free Memory Tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13614–13624, 2025.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking Any Point with Per-Frame Initialization and Temporal Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10061–10072. IEEE, 2023.

Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. MetaBEV: Solving Sensor Failures for 3D Detection and Map Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8721–8731, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. RVT: Robotic View Transformer for 3D Object Manipulation. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 694–710. PMLR, 2023.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 5021–5028. IEEE, 2024.

Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. UMI-on-Legs: Making Manipulation Policies Mobile with Manipulation-Centric Whole-body Controllers. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 5254–5270. PMLR, 2025.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 885–915. PMLR, 2025.

Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, and Wenqiang Zhang. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19079–19091, 2024.

Dianyong Hou, Chengrui Zhu, Zhen Zhang, Zhibin Li, Chuang Guo, and Yong Liu. Efficient Learning of A Unified Policy For Whole-body Manipulation and Locomotion Skills. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5455–5461. IEEE, 2025.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual Language Maps for Robot Navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 10608–10615. IEEE, 2023a.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Fei-Fei Li. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 540–562. PMLR, 2023b.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua Tenenbaum, Celso de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Proceedings of Robotics: Science and Systems*. Robotics: Science and Systems Foundation, 2023.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards Unified Latent VLA for Whole-body Loco-manipulation Control. In *International Conference on Learning Representations*, 2026.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It Is Better to Track Together. In *Proceedings of the European Conference on Computer Vision*, pp. 18–35. Springer Nature Switzerland, 2024.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, 2025.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. SplaTAM: Splat Track & Map 3D Gaussians for Dense RGB-D SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21357–21366, 2024.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19729–19739, 2023.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, 2024.

Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. CRN: Camera Radar Net for Accurate, Robust, Efficient 3D Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17615–17626, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026. IEEE, 2023.

Minghan Li, Shuai Li, Xindong Zhang, and Lei Zhang. UniVS: Unified and Universal Video Segmentation with Prompts as Queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3227–3238, 2024.

Qiwei Liang, Boyang Cai, Rongyi He, Hui Li, Tao Teng, Haihan Duan, Changxin Huang, and Runhao Zeng. Whole-Body Coordination for Dynamic Object Grasping with Legged Manipulators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18434–18442. Association for the Advancement of Artificial Intelligence (AAAI), 2026.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-camera Fusion in Bird's Eye View for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14928–14937, 2024.

Minghuan Liu, Zixuan Chen, Xuxin Cheng, Yandong Ji, Ri-Zhao Qiu, Ruihan Yang, and Xiaolong Wang. Visual Whole-Body Control for Legged Loco-Manipulation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 234–257. PMLR, 2025a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55. Springer Nature Switzerland, 2024.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation. In *International Conference on Learning Representations*, 2025b.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 2774–2781. IEEE, 2023.

Yuntao Ma, Andrei Cramariuc, Farbod Farshidian, and Marco Hutter. Learning coordinated badminton skills for legged manipulators. *Science Robotics*, 10(102):eadu3922, 2025.

Jared Mejia, Victoria Dean, Tess Hellebrekers, and Abhinav Gupta. Hearing Touch: Audio-Visual Pretraining for Contact-Rich Manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6912–6919. IEEE, 2024.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-Scale Simulation of Household Tasks for Generalist Robots. In *Proceedings of Robotics: Science and Systems*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Guoping Pan, Qingwei Ben, Zhecheng Yuan, Guangqi Jiang, Yandong Ji, Shoujie Li, Jiangmiao Pang, Houde Liu, and Huazhe Xu. RoboDuet: Learning a Cooperative Policy for Whole-Body Legged Loco-Manipulation. *IEEE Robotics and Automation Letters*, 10(5):4564–4571, 2025.

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D Scene Understanding With Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, 2023.

Tifanny Portela, Gabriel B. Margolis, Yandong Ji, and Pulkit Agrawal. Learning Force Control for Legged Manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 15366–15372. IEEE, 2024.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D Language Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, 2025.

Daniel Seichter, Benedict Stephan, Söhnke Benedikt Fischedick, Steffen Müller, Leonard Rabes, and Horst-Michael Gross. PanopticNDT: Efficient and Robust Panoptic Mapping. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7233–7240. IEEE, 2023.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schoelkopf, Thomas Brox, and Francesco Locatello. Bridging the Gap to Real-World Object-Centric Learning. In *International Conference on Learning Representations*, 2023.

Jean Pierre Sleiman, Mayank Mittal, and Marco Hutter. Guided Reinforcement Learning for Robust Multi-Contact Loco-Manipulation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 531–546. PMLR, 2025.

Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, and Liwei Wang. UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6792–6802, 2023a.

Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking Everything Everywhere All at Once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19795–19806, 2023b.

Qianxu Wang, Haotong Zhang, Congyue Deng, Yang You, Hao Dong, Yixin Zhu, and Leonidas Guibas. SparseDFF: Sparse-View Feature Distillation for One-Shot Dexterous Manipulation. In *International Conference on Learning Representations*, 2024.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models. In *International Conference on Learning Representations*, 2023.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: Tracking Any 2D Pixels in 3D Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20406–20417, 2024.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: Advancing 3D Point Tracking with Explicit Camera Motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6726–6737, 2025.

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*. Robotics: Science and Systems Foundation, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023a.

Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. GO-SLAM: Global Optimization for Consistent 3D Instant Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3727–3737, 2023b.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61229–61245. PMLR, 2024.

Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual Prompt Multi-Modal Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9516–9526, 2023.

A Pixel-Aligned Representation and Map Integration

This section expands Section 3.1. Physical consistency here concerns the spatial and temporal support of measurements; it does not imply that calibration, pose estimates, or individual sensors are error-free.

A.1 Coordinate conventions and channel validity

Let $\mathbf{T}_{a\leftarrow b}$ transform coordinates from frame b to frame a. Sensor timestamps are placed on the reference clock using $t = t^{(s)} + \Delta^{(s)}$. For a range return with sensor-frame position $\mathbf{x}_k^s$, projection into the RGB camera gives

$$\mathbf{x}_k^c = R_{c\leftarrow s}\mathbf{x}_k^s + \mathbf{t}_{c\leftarrow s}, \qquad d_k = \mathbf{e}_3^\top \mathbf{x}_k^c, \qquad \mathbf{u}_k = \pi(K\mathbf{x}_k^c), \tag{A.1}$$

where $\mathbf{e}_3 = (0,0,1)^\top$. Only positive-depth returns projecting inside the image can populate a pixel. The depth channel stores d_k, the camera-axis depth, rather than the sensor's radial range. FIR observations are registered through the calibrated cross-camera mapping. These operations establish correspondence but do not make sparse physical measurements dense.

For channel $a \in \{I, D, V^{\mathrm{rad}}, R, \Theta\}$, let $m_t^a(\mathbf{u}) \in \{0,1\}$ indicate whether a valid observation or explicitly reconstructed estimate is available. Validity accompanies the PMV field. An unavailable measurement is distinct from an observed zero value; in particular, missing Doppler evidence is not evidence of a stationary target. Likewise, FIR intensity is a thermal observation, not an absolute temperature unless supported by radiometric calibration. The evaluated targets are inanimate, so thermal evidence depends on measured target–background contrast.

A.2 Nearest-surface aggregation

Let $\mathcal{K}_t(\mathbf{u})$ contain geometrically valid projected returns, and let $\mathcal{K}_t^a(\mathbf{u})$ contain those supplying channel a. For a nonempty candidate set, define

$$\begin{aligned} d_{\min}(\mathbf{u}) &= \min_{k\in\mathcal{K}_t(\mathbf{u})} d_k, \\ w_k &= \sigma_k^{-2}\mathbb{I}[d_k \le d_{\min}(\mathbf{u}) + \delta_d], \\ a_t(\mathbf{u}) &= \frac{\sum_{k\in\mathcal{K}_t^a(\mathbf{u})} w_k a_k}{\sum_{k\in\mathcal{K}_t^a(\mathbf{u})} w_k}. \end{aligned} \tag{A.2}$$

Here $\sigma_k > 0$ represents return uncertainty and δ_d is the depth-consistency tolerance. The gate reduces contributions from background surfaces projecting to the same pixel, while inverse-uncertainty weighting favors reliable returns. A zero denominator leaves the channel invalid. Channel-specific candidate sets prevent, for example, a LiDAR return from being treated as a radar velocity observation. The gate is a local heuristic and cannot eliminate every multipath return or calibration error.

A.3 Geometry, semantics, and spatial context

Valid depth is back-projected using Eq. (4) and integrated with the robot pose. For a voxel v, the weighted signed-distance update is

$$\phi_t(v) = \frac{w_{t-1}(v)\phi_{t-1}(v) + \widehat{w}_t(v)\widehat{\phi}_t(v)}{w_{t-1}(v) + \widehat{w}_t(v)}, \qquad w_t(v) = w_{t-1}(v) + \widehat{w}_t(v). \tag{A.3}$$

This observation-integration rule is distinct from the Euclidean signed distance field used for collision queries. Local surfels retain surface geometry for manipulation.

Let $\boldsymbol{\pi}_{t-1}(v)$ denote the accumulated semantic scores and $\boldsymbol{\ell}_t(\mathbf{u}) = p_t(\cdot \mid \mathbf{u})$ the current per-pixel output. Their normalized multiplicative update is

$$\boldsymbol{\pi}_t(v) = \frac{\boldsymbol{\pi}_{t-1}(v) \odot \boldsymbol{\ell}_t(\mathbf{u})}{\mathbf{1}^\top[\boldsymbol{\pi}_{t-1}(v) \odot \boldsymbol{\ell}_t(\mathbf{u})]}. \tag{A.4}$$

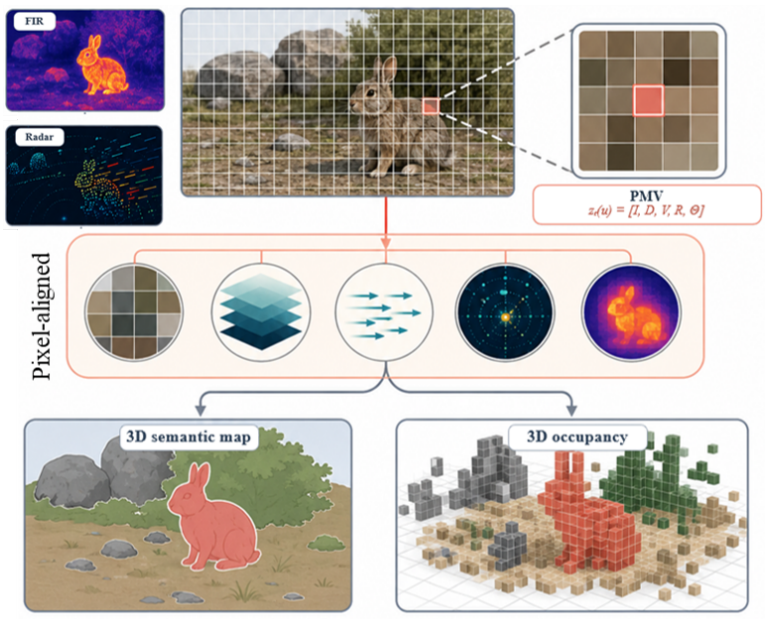


Figure 8: PMV structure. A common pixel lattice associates appearance with available depth, radial motion, reflectivity, and thermal evidence; channel validity preserves missing-data information.

The update requires a positive normalization denominator; positive scores also permit equivalent log-domain accumulation. Adjacent observations are correlated, so the accumulated scores are not interpreted as independently calibrated Bayesian posteriors.

The shared world frame makes metric relations available without repeating cross-modal association. For valid point sets $\mathcal{P}_A$ and $\mathcal{P}_B$ belonging to regions A and B, their observed separation is

$$d(A, B) = \min_{\mathbf{x}\in\mathcal{P}_A,\, \mathbf{y}\in\mathcal{P}_B} \|\mathbf{x} - \mathbf{y}\| . \tag{A.5}$$

This provides spatial context for approach and manipulation. It is a distance between observed supports, not a guarantee about unobserved surfaces. Furthermore, the integration equations alone do not remove stale geometry from moving objects; global mapping and short-horizon target prediction have different temporal roles.

B Persistent Target-State Estimation and Validation

B.1 Observed State and Short-Horizon Prediction

For an accepted mask M_t, define $\Omega_t^a = \{\mathbf{u} \in M_t : m_t^a(\mathbf{u}) = 1\}$ and $\mathcal{P}_t = \{\mathbf{x}_t^w(\mathbf{u}) : \mathbf{u} \in \Omega_t^D\}$. Position and thermal statistics are computed from their respective valid supports:

$$\mathbf{p}_t = \text{median}_{\mathbf{x}\in\mathcal{P}_t}\, \mathbf{x}, \qquad \bar{\Theta}_t = \text{median}_{\mathbf{u}\in\Omega_t^\Theta}\, \Theta_t(\mathbf{u}), \qquad \Sigma_t = \text{Cov}_{\text{rob}}(\mathcal{P}_t). \tag{B.1}$$

The position median is coordinate-wise. Thermal statistics remain unavailable when their support is empty. The robust dispersion Σ_t includes effects of object extent, sampling, and localization error; it is not a calibrated covariance of the target center.

Let $s < t$ be the preceding accepted observation time. Cartesian velocity uses the elapsed time between accepted positions:

$$\mathbf{v}_t = \beta \frac{\mathbf{p}_t - \mathbf{p}_s}{t - s} + (1 - \beta)\mathbf{v}_s, \qquad 0 \le \beta \le 1. \tag{B.2}$$

For consecutive accepted frames, this reduces to the inter-frame update. Predicted positions during an observation gap are not inserted as new measurements into this difference. Until sufficient observations are available, Cartesian velocity is an initialized estimate.

Prediction advances the last accepted state according to

$$\widehat{\mathbf{p}}_t = \mathbf{p}_s + \mathbf{v}_s(t-s), \qquad \widehat{\mathbf{v}}_t = \mathbf{v}_s, \qquad \widehat{\Sigma}_t = \Sigma_s + Q_{s\to t}, \tag{B.3}$$

where $Q_{s\to t}$ accumulates uncertainty growth. This local prior restricts plausible reassociation; it does not guarantee accurate extrapolation through abrupt motion or prolonged occlusion.

B.2 Radial-motion consistency under robot motion

Radar radial velocity measures relative motion along a line of sight and must not be equated with $\mathbf{v}_t$. To make the reference convention explicit, let $\mathbf{o}_t^w$ and $\mathbf{v}_{\mathrm{rad},t}^w$ be the radar origin and its world-frame velocity. With positive Doppler defined as increasing range, the translational prediction is

$$\widehat{\boldsymbol{\ell}}_t^w = \frac{\widehat{\mathbf{p}}_t - \mathbf{o}_t^w}{\|\widehat{\mathbf{p}}_t - \mathbf{o}_t^w\|}, \qquad \widehat{V}_t^{\mathrm{rad}} = (\widehat{\boldsymbol{\ell}}_t^w)^\top(\widehat{\mathbf{v}}_t - \mathbf{v}_{\mathrm{rad},t}^w). \tag{B.4}$$

For a radar rigidly mounted on the base, its velocity includes the lever-arm term

$$\mathbf{v}_{\mathrm{rad},t}^w = \mathbf{v}_{b,t}^w + \boldsymbol{\omega}_{b,t}^w \times \left(R_{w\leftarrow b}(t)\mathbf{r}_{\mathrm{rad}}^b\right), \tag{B.5}$$

where $\mathbf{r}_{\mathrm{rad}}^b$ is the sensor offset from the base origin. The observed and predicted values must use the same Doppler sign and ego-motion convention. If observations are already ego-motion compensated, the corresponding prediction is $(\widehat{\boldsymbol{\ell}}_t^w)^\top\widehat{\mathbf{v}}_t$; compensation must not be applied twice. The residual in Eq. (3.2) compares the valid mask-level radial statistic with this current-view prediction.

B.3 Missing evidence and identity acceptance

The MAD terms in Eq. (5) are computed separately over valid depth and radial-velocity support. A numerically small dispersion from insufficient support is not, by itself, evidence of a coherent object. Position validation requires enough valid depth to localize the candidate. Unavailable or unreliable radial and thermal cues contribute no consistency term; they are not inserted as zero residuals and counted as supporting evidence.

The online update follows four steps:

1. Construct the synchronized PMV observation, update scene context, and predict the previously accepted target state.
2. Generate an acquisition candidate or propagate the existing mask; extract its valid physical statistics and evaluate the applicable scores.
3. Accept a state update only after semantic and physical checks. Otherwise retain the prediction, regenerate candidates from the original query, and apply Eq. (7) with an explicit rejection option.
4. Supply the accepted state and confidence to the controller. An unresolved identity or unavailable localization inhibits capture.

Ranking candidates is not sufficient for acceptance: even the highest-ranked candidate may be incompatible with the maintained instance. Confidence c_t regulates this decision and subsequent execution, but is not a probability that the target identity is correct.

C Confidence-Guided Approach and Capture

C.1 Viewpoint selection and coordinated alignment

For a candidate base position $\mathbf{x}$ and heading ψ, the visibility term in Eq. (8) is

$$C_{\mathrm{vis}}(\mathbf{x},\psi) = 1 - \mathbb{I}[\mathbf{p}_t \in \mathrm{FOV}(\mathbf{x},\psi)] \exp\left[-\frac{d(\mathbf{x},\mathbf{p}_t)^2}{\sigma_{\mathrm{vis}}^2}\right]. \tag{C.1}$$

The cost discourages excessive distance and unfavorable viewing direction; field-of-view membership alone does not certify an unoccluded line of sight. Manipulability is evaluated relative to a

precomputed reachability volume in the base frame. Geometry and semantic costs provide complementary collision and traversability context.

Over a discrete planning horizon, coordinated base and arm commands minimize

$$\min_{\{\mathbf{u}_i^b,\mathbf{u}_i^a\}_{i=0}^{N-1}} \sum_{i=1}^{N} \left[\|\mathbf{e}_i^{\mathrm{ee}}\|_{W_e}^2 + \left\|\mathbf{e}_i^b\right\|_{W_b}^2 + \lambda_{\mathrm{col}} C_{\mathrm{col},i} \right], \tag{C.2}$$

subject to the robot's motion, kinematic, and collision constraints. Here $\|\mathbf{e}\|_W^2 = \mathbf{e}^\top W \mathbf{e}$; $\mathbf{e}_i^{\mathrm{ee}}$ measures error relative to the desired capture configuration and $\mathbf{e}_i^b$ measures error relative to the desired base stance. The optimization couples repositioning and reaching, while confidence determines whether execution may proceed to capture.

C.2 Static grasping and dynamic interception

For a quasi-static target, $\|\mathbf{v}_t\| < \tau_v$, candidate grasps from local target geometry are ranked by

$$g^* = \arg\max_{g \in \mathcal{G}_t^{\mathrm{feas}}} \left[S_{\mathrm{geom}}(g) - \lambda_e S_{\mathrm{env}}(g) - \lambda_k S_{\mathrm{kin}}(g) \right], \tag{C.3}$$

where the terms measure geometric suitability, environmental interference, and proximity to kinematic limits. The feasible set excludes candidates violating hard constraints.

Moving targets use the short-horizon interception rule in Eq. (9). The target trajectory and end-effector trajectory must be expressed in the same frame, and candidate interception times must remain reachable and collision-free. Local arm correction uses the damped-Jacobian form

$$\dot{\mathbf{q}}_a = J_a^\dagger \left[K_p(\mathbf{p}^* - \mathbf{p}^{\mathrm{ee}}) + K_v(\mathbf{v}_t - \dot{\mathbf{p}}^{\mathrm{ee}}) \right], \qquad J_a^\dagger = J_a^\top (J_a J_a^\top + \lambda_J^2 I)^{-1}. \tag{C.4}$$

The task-space signals and Jacobian use a common reference with base motion accounted for. This local correction operates within the coordinated controller and its limits; it does not independently guarantee feasibility. The plan is revised as new accepted observations arrive.

C.3 Coverage score and outcome verification

A primitive a defines a time-varying capture region $\mathcal{C}_a(\Delta)$. For a specified normalized predictive spatial weighting $\mu_{t,\Delta}(\cdot \mid \mathcal{B}_t)$, its time-averaged coverage is

$$S_{\mathrm{cov}}(a \mid \mathcal{B}_t) = \frac{1}{H} \int_0^H \mu_{t,\Delta}(\mathcal{C}_a(\Delta) \mid \mathcal{B}_t)\, d\Delta. \tag{C.5}$$

The score averages predicted occupancy of the reachable capture region. The weighting may summarize a nominal prediction or a spread of candidate positions; Σ_t alone does not specify its distribution. Neither the weighting nor S_{cov} is assumed statistically calibrated. The score measures spatial coverage over time, not the probability of correct association, closure, and sustained retention.

Primitive selection balances coverage and interaction costs:

$$a^* = \arg\max_{a \in \mathcal{A}_t^{\mathrm{feas}}} \left[S_{\mathrm{cov}}(a \mid \mathcal{B}_t) - \lambda_e S_{\mathrm{env}}(a) - \lambda_k S_{\mathrm{kin}}(a) \right]. \tag{C.6}$$

The interface accommodates grasping, interception, and available enclosure or sweeping tools through their capture regions. The reported evaluation aggregates task outcomes; it does not establish separate success rates for each primitive.

Recovery is invoked when $c_t < \tau_c$ or localization is unavailable. Approach or alignment continues for $\tau_c \le c_t < \tau_a$. Capture requires $c_t \ge \tau_a$, an accepted identity, and feasibility checks. An attempt is verified against commanded-instance capture and retention; unsuccessful attempts return to alignment or recovery while the episode remains active.

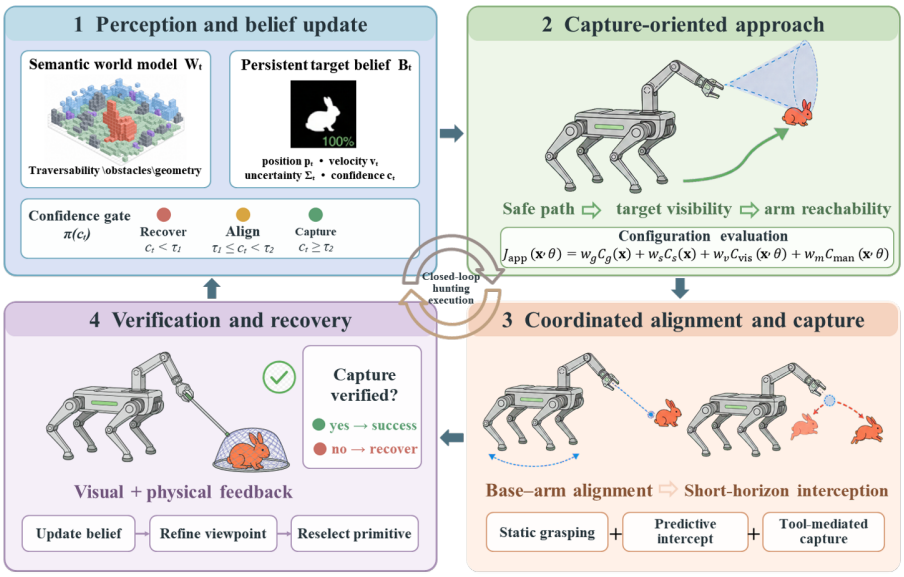


Figure 9: Capture loop: update the map and target state, approach, align, select a feasible primitive, and verify retention. The legacy P_{cap} notation denotes the coverage score in Eq. (C.5).

D Experimental Protocols and Statistical Definitions

D.1 Platform and comparison conditions

Table 4 summarizes the reported settings. The mouse and monster are primarily stationary; rabbit and dog targets additionally follow prescribed trajectories. These are inanimate targets. Robot and target poses, obstacles, and motion directions vary across eight physical layouts. Marsim–Unity is used for controller tuning and safety validation; the reported mapping, tracking, capture, and ablation results use the real robot.

Mapping comparisons use matched trajectories, map resolution, semantic backbone, and SLAM/odometry estimates. The high-density LiDAR reference is constructed offline and registered to the common world frame; it is unavailable to online mapping methods. This reference is distinct from any auxiliary online LiDAR measurements.

KineMan-RL, BirdieRL, and DQ-Net denote the task-adapted implementations evaluated on the same platform, as specified in Section 4.1. Layouts, initial conditions, trajectories, stopping rules, and safety constraints are matched, while each implementation retains its required observation channels. Consequently, these are system comparisons; they do not isolate representation quality under identical sensor inputs.

D.2 Mapping metrics

For the evaluated semantic classes, 3D mIoU averages $\mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$. Chamfer-L1 denotes the bidirectional mean of unsquared nearest-neighbor Euclidean distances between reconstructed and reference point sets, reported in centimeters. Target recall measures the fraction of reference target instances recovered by the map. Free-space F1 summarizes precision and recall for free-space predictions. Latency is the reported per-frame mapping processing time, not the duration of a complete perception–action cycle. Continuous mapping results are summarized over real-robot sequences as mean $\pm$ standard deviation.

D.3 Occlusion and identity metrics

Occlusion ratios use projected target area from the calibrated RGB mask, with aligned range observations supporting verification. Occluder motion and exposure duration follow matched profiles;

Table 4: Implementation settings and separate evaluation protocols.

Item	Setting
Platform	Unitree Go2 quadruped with PiPER arm
Primary observations	RGB, FIR, and 4D mmWave radar
LiDAR	Offline mapping reference; auxiliary online ranging in selected experiments
Perception / planning / control	$\geq$ 30 Hz / 10 Hz / 200 Hz
Moving-target speed	0.2–1.0 m/s
Capture protocol	4 targets $\times$ 4 conditions $\times$ 30 episodes; $n = 480$ per method
Capture criterion	Capture within 60 s; retain for $\geq$ 3 s; no collision or safety-limit violation
Occlusion protocol	4 targets $\times$ 4 conditions $\times$ 4 occlusion levels $\times$ 10 repetitions; $n = 640$ per method
Occluded target area	20%, 40%, 60%, or 80%
Recovery criterion	Correct identity and 3D state restored within 5 s and stable for 10 frames

distractor identities and relative positions are randomized, and trials are interleaved across methods and conditions. Each condition–occlusion point pools four targets and ten repetitions ($n = 40$); pooling conditions gives $n = 160$ per occlusion level.

Persistent lock and post-occlusion recovery are distinct events. Lock loss occurs when identity is absent or confidence remains below threshold for more than 0.5 s. Recovery requires the correct identity and metric state within 5 s, followed by ten stable frames, regardless of formal lock loss. Thus recovery is evaluated over all 640 trials, not only those classified as lock failures. These metrics need not be complements: identity may remain assigned while fresh metric-state validation is unsuccessful. The reported identity-switch rate counts trials containing a switch to an incorrect instance.

Recovery-time summaries include successful recoveries only. Confidence AUC is the normalized area under a confidence trace over the evaluation interval, $T^{-1}\int_{t_0}^{t_0+T} c(t)\,dt$. It describes score persistence and does not establish probability calibration. Cross-method interpretation also depends on the declared confidence scales and evaluation window.

D.4 Capture outcomes and statistical treatment

All 480 capture episodes per method remain in the success denominator. Target loss, infeasible approaches, timeouts, unsuccessful closure or enclosure, failed retention, and constraint violations are failures. For episode i, let y_i indicate success and t_i denote its recorded capture time when successful. The cumulative success curve is

$$F_{\mathrm{cap}}(\tau) = \frac{1}{N}\sum_{i:y_i=1} \mathbb{I}[t_i \leq \tau]. \tag{D.1}$$

Failed episodes contribute zero throughout, so the endpoint equals the final success rate. This empirical curve is distinct from the controller's coverage score. Capture-time distributions in the ablation figure are conditional on success and should be interpreted together with the associated success counts.

For k successes among n trials, $\widehat{p} = k/n$, the 95% Wilson interval is

$$\frac{\widehat{p} + z^2/(2n) \;\pm\; z\sqrt{\widehat{p}(1-\widehat{p})/n + z^2/(4n^2)}}{1 + z^2/n}, \qquad z = 1.96. \tag{D.2}$$

The ablation table uses two-sided Fisher exact tests against the full system, with Holm correction across four comparisons. These pooled count-based analyses do not model dependence induced by shared layouts or trajectories. They also do not identify the causal effect of an individual physical cue.

Table 5: Complete mapping results (mean ± standard deviation). mIoU, recall, and F1 are percentages; Chamfer-L1 is in cm and latency in ms.

Condition	Method	mIoU ↑	Chamfer ↓	Recall ↑	Free-space F1 ↑	Latency ↓
Indoor normal	RGB-only	74.2 ± 2.8	6.8 ± 0.7	79.1 ± 3.2	82.4 ± 2.6	31.6 ± 1.7
	RGB-D	80.6 ± 2.3	4.9 ± 0.5	85.0 ± 2.8	87.8 ± 2.1	38.9 ± 2.0
	Late fusion	85.1 ± 2.0	4.1 ± 0.4	89.6 ± 2.4	90.3 ± 1.9	56.8 ± 2.7
	Ours	91.0 ± 1.5	2.8 ± 0.3	95.5 ± 1.6	94.7 ± 1.3	29.8 ± 1.6
Indoor low-light	RGB-only	50.4 ± 4.2	12.1 ± 1.3	52.7 ± 4.7	65.9 ± 3.8	31.4 ± 1.9
	RGB-D	62.5 ± 3.6	8.4 ± 0.9	65.2 ± 4.1	74.8 ± 3.2	38.7 ± 2.1
	Late fusion	72.9 ± 3.1	6.2 ± 0.7	78.7 ± 3.5	82.8 ± 2.7	57.2 ± 3.0
	Ours	87.3 ± 2.1	3.4 ± 0.4	92.9 ± 2.0	91.9 ± 1.8	30.6 ± 1.8
Outdoor normal	RGB-only	69.8 ± 3.2	8.9 ± 0.9	72.6 ± 3.6	78.5 ± 3.0	32.3 ± 1.9
	RGB-D	76.1 ± 2.8	6.9 ± 0.7	79.7 ± 3.2	83.4 ± 2.6	39.6 ± 2.2
	Late fusion	81.6 ± 2.5	5.5 ± 0.6	85.7 ± 2.8	87.0 ± 2.3	58.1 ± 3.2
	Ours	88.7 ± 1.8	3.6 ± 0.4	93.2 ± 1.9	92.5 ± 1.6	31.2 ± 1.9
Outdoor low-light	RGB-only	42.3 ± 4.7	16.0 ± 1.7	43.8 ± 5.0	59.1 ± 4.1	32.0 ± 2.0
	RGB-D	55.1 ± 4.0	11.8 ± 1.2	59.1 ± 4.3	68.8 ± 3.5	39.3 ± 2.3
	Late fusion	68.2 ± 3.5	8.6 ± 0.9	73.3 ± 3.8	77.9 ± 3.1	58.8 ± 3.4
	Ours	83.8 ± 2.5	4.7 ± 0.5	89.4 ± 2.4	88.7 ± 2.1	32.1 ± 2.0

Table 6: Post-occlusion recovery pooled across conditions ($n = 160$ per level). Counts correspond to the reported balanced-protocol recovery rates.

Occluded area	Ours	DQ-Net	Gain (pp)
20%	155/160 (96.9)	142/160 (88.8)	8.1
40%	149/160 (93.1)	121/160 (75.6)	17.5
60%	134/160 (83.8)	88/160 (55.0)	28.8
80%	105/160 (65.6)	45/160 (28.1)	37.5

E Detailed Quantitative Results

E.1 Complete mapping results

Table 5 provides all four mapping conditions, including the normal-illumination settings omitted from the compact main-text table. The mIoU margins over late fusion are 5.9 and 7.1 percentage points under normal indoor and outdoor illumination, increasing to 14.4 and 15.6 points under low illumination. Geometry, target recall, and free-space metrics provide complementary evidence about the representation's spatial utility.

E.2 Recovery by occlusion severity

Table 6 pools the balanced trials by occlusion level. The counts sum to the 543/640 and 396/640 recovery outcomes reported in Table 2. The larger gap at severe occlusion supports evaluating recovery frequency alongside conditional recovery time.

E.3 Capture outcomes by target and condition

Table 7 retains the target-level counts, and Table 8 summarizes the pooled results. The overall gain over DQ-Net is 14.6 percentage points; outdoors at night it is $(106 - 77)/120 = 24.2$ percentage points. In that condition, mouse capture increases from 16/30 to 25/30. These comparisons assess the full pipeline, including sensing, association, planning, and execution.

Table 7: Capture-success counts by target and condition. Each target entry is out of 30; condition totals are out of 120.

Condition	Method	Mouse	Rabbit	Dog	Monster	Total / 120
Indoor day	KineMan-RL	21	24	25	26	96
	BirdieRL	23	25	26	27	101
	DQ-Net	24	26	27	28	105
	Ours	28	28	29	29	114
Indoor night	KineMan-RL	14	17	19	21	71
	BirdieRL	16	18	21	22	77
	DQ-Net	18	20	23	24	85
	Ours	26	27	27	28	108
Outdoor day	KineMan-RL	20	22	24	25	91
	BirdieRL	21	23	25	26	95
	DQ-Net	23	25	26	27	101
	Ours	26	27	29	28	110
Outdoor night	KineMan-RL	12	14	17	18	61
	BirdieRL	14	16	19	20	69
	DQ-Net	16	18	21	22	77
	Ours	25	26	27	28	106

Table 8: Pooled capture outcomes. Confidence intervals use all 480 episodes.

Method	Successful / failed	Success (%)	95% Wilson interval (%)
KineMan-RL	319 / 161	66.5	[62.1, 70.5]
BirdieRL	342 / 138	71.3	[67.0, 75.1]
DQ-Net	368 / 112	76.7	[72.7, 80.2]
Ours	438 / 42	91.3	[88.4, 93.5]

E.4 Module replacements and failure interpretation

Table 9 specifies the replacement used when each module is removed, complementing the target-wise success counts in Table 3. Hardware, scene layouts, episode limits, and safety criteria follow the same capture protocol, allowing the configurations to be compared under consistent evaluation conditions. The full system achieves a median capture time of 17.8 s (IQR: 15.34–20.37 s). All reported medians are computed over successful episodes only and should therefore be interpreted alongside success rates: a shorter conditional capture time does not necessarily indicate better overall performance.

Across the five configurations, 416 episodes fail. Target loss (124), grasp failure (100), and approach timeout (95) together account for 319/416 failures (76.7%), identifying association, physical capture, and timely approach as the principal failure categories. The configuration without persistent tracking records 56 target-loss events, the configuration without mapping records 28 approach timeouts, and the configuration without coordinated execution records 32 grasp failures. These patterns are consistent with complementary module roles: tracking preserves target identity, mapping supplies spatial context for approach planning, and coordinated execution supports base–arm alignment and capture. They connect the aggregate success results to distinct stages of the perception–action loop, although the interacting components prevent a purely independent attribution of failures. In particular, these module-level ablations do not isolate geometric, radial-motion, and thermal consistency individually. The available results also do not establish generalization to unseen target categories or motion regimes, which requires dedicated evaluation beyond the tested conditions.

Table 9: Module replacements, capture outcomes, and conditional median capture time.

Configuration	Replacement	Successes	Failures	Median (s)
Full system	Complete pipeline	438	42	17.8
w/o Fusion	Late fusion replaces pixel alignment	405	75	22.4
w/o Mapping	Geometry-only occupancy replaces mapping	391	89	24.7
w/o Locking	Frame-wise detection replaces association	368	112	29.6
w/o Execution	Shortest-path navigation with arm control	382	98	27.5

F Additional Pursuit Sequences

Figure 10 supplements the quantitative occlusion evaluation with pursuit sequences sampled every 2 s. The examples show viewpoint changes, partial occlusion, and nearby distractors. They illustrate how maintained identity and updated metric state support continued pursuit; they do not replace the complete trial statistics. The representation and execution schematics appear in Figs. 8 and 9, alongside the corresponding implementation details.

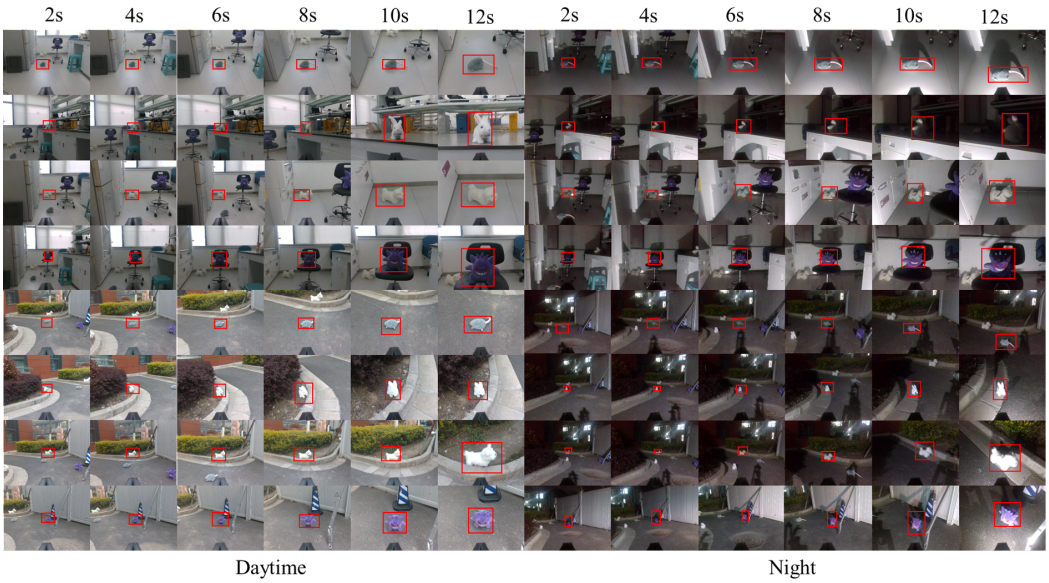


Figure 10: Pursuit sequences sampled every 2 s under viewpoint changes, occlusion, and distractors. Ours in the original legends denotes the proposed system.