Referent Drift: Measuring Persistent Instance Memory Under Scene Changes

Sangwoo Kim[1,3,*] **Seongyoung Moon**[1,2,*] **Dong-oh Kang**[1]
Kimin Yun[1,3,†] **Eun-ju Jeong**[1,†]

[1]Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
[2]Inha University, Incheon, Korea [3]Univ. of Science and Technology (UST), Daejeon, Korea
{swkim2, dongoh, kimin.yun, eunju}@etri.re.kr seongyoungmoon@gmail.com
[*]Equal contribution. [†]Corresponding authors (primary contact: Eun-ju Jeong).

Abstract: In open-vocabulary mobile manipulation, static object-selection accuracy can make a robot's memory appear reliable even when the same symbolic handle refers to a different physical object after the scene changes. We call this failure *referent drift* and introduce Persistent Referent Accuracy (PRA), a history-dependent evaluation of whether a handle remains bound to the same physical instance across a scene intervention. On a Stretch 3 robot, a stateless spatial-ordering resolver reaches 86.7% (13/15) static instance-selection accuracy but 0/14 PRA after a neighbouring object is removed and 0/10 after one is inserted to the target's left. A rank-preserving control inserts to its right instead, adding one object and leaving the bound target untouched, and the resolver returns to 10/10. In these controlled interventions, scene change alone did not induce drift; failure tracked reassignment of the bound target. Appearance-based selection fails already in the static case, degrading toward chance as identical instances are added and recovering only when a discriminative colour cue is available. As a minimal constructive control, a registry storing the binding explicitly (Deterministic Instance Grounding) removes the observed failure in all tested conditions. Appearance answers *what* and current-scene ordering *which now*; neither answers *which same instance over time*, so static selection accuracy does not test referential memory: persistence must be evaluated under interventions that change the current-scene interpretation while leaving the bound object in place.

Keywords: Robot Memory, Persistent Instance Identity, History-Dependent Evaluation

1 Introduction

Manipulation requires object references to survive the robot's own actions. A robot can answer the same object query correctly twice and still refer to two different physical objects. We call this *referent drift*: the symbolic handle stays the same while the physical entity it denotes changes. Once a handle is bound to a physical instance, later uses should preserve that binding until it is explicitly invalidated, not silently reassign it from the current scene. This is a memory requirement, not a perception one, because the later referent is fixed by a past binding that the current scene does not carry. Identical products remove appearance as a shortcut. An ordinal such as "can #2" (Figure 1) is a persistent task handle established at t_0, not a fresh description of whichever can is second at each query. It is a proxy for a task-level binding such as "the third item of order 1234", which stays with one object until invalidated. Interchangeable units do not make the binding arbitrary: any record a task slot carries (e.g. where the unit went) refers to the unit bound to it, so a silent substitution completes the task while invalidating it. Re-indexing is reasonable under a current-scene reading; under the persistent-handle reading it is referent drift, and static evaluation cannot separate the two because they agree while the scene is unchanged.

CoRL 2026 Workshop on Memory for Robot Foundation Models (RoboMem), Austin TX, USA. Non-archival.

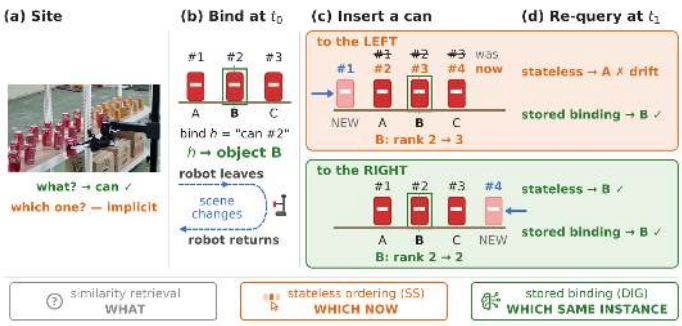


Figure 1: **Current-scene indexing versus persistent task binding.** (a) Appearance identifies the category, not the instance. (b) A handle h is bound to B at t_0; after an observation gap the scene has changed while B stays in place. (c,d) Inserting on the left renumbers the cans, so SS, reading the ordinal against the new scene, returns A while DIG retains the binding and returns B; inserting on the right leaves B at rank 2 and both agree. Neither reading is intrinsically incorrect: SS answers *which one now*, DIG *which same instance as before*, and PRA evaluates the latter.

The information needed to recover the original referent comes from the past binding, not from the current scene alone. We first show that appearance-based retrieval does not resolve same-category instances even statically (Sec. 4.1), then ask whether stateless spatial ordering preserves an established reference across scene changes. Deploying a semantic-memory stack [1] among repeated products showed that category retrieval does not fix which physical instance a task reference should keep; Deterministic Instance Grounding (DIG) therefore assigns identity once, from spatial structure, and stores it in a registry that later queries resolve by lookup.

The same gap is built into the systems this registry replaced. Many open-vocabulary mobile-manipulation (OVMM) pipelines re-ground language against the current scene, where static success does not establish that referential history was retained. Such systems pair an open-vocabulary detector [2, 3] or a vision-language embedding [4] with a spatial map and resolve instructions by similarity search [1, 5] or scene-graph planners [6], and are scored on whether the right category is recovered. Our protocol isolates the failure when a task-level instance reference is re-derived from current-scene cues rather than retained. Identity, where addressed, rides on appearance or motion [7, 8, 9], weak cues when instances neither differ nor move. Tracking evaluates identity consistency across observations (e.g. HOTA [10]) and persistent scene memories [11] keep object identifiers from spatial and appearance cues; both can associate across a departure and return when the target stays put. PRA asks a narrower question: whether the task-level binding, once established, is the one the system acts on after an intervention, whatever association mechanism it uses. Downstream policies [12, 13, 14] assume a designated target, and history-dependent benchmarks [15, 16] score task completion, so a policy that re-grounds to a different but interchangeable instance still succeeds and the drift goes unrecorded. We score the binding itself: whether an established reference stays attached to the same physical instance after a scene change.

Contributions. (i) *Vocabulary and failure mode:* we separate semantic recognition, current-scene instance resolution, and persistent physical-instance identity, and name the silent reassignment referent drift. (ii) *Evaluation:* PRA is a history-dependent evaluation of whether a handle remains bound to the same physical instance across a scene intervention. (iii) *Controlled diagnosis:* a control that adds one identical object on either side of the bound target rules out scene change itself as the explanation. (iv) *Constructive control:* explicitly storing the binding with DIG removes the observed failure in the controlled Stretch 3 experiments. The contribution is not that stored lookup outperforms re-indexing by construction, but that static evaluation cannot reveal which referential semantics a robot implements.

2 Why Current-Scene Cues Do Not Guarantee Persistent Identity

The referent of "the can I selected earlier" is defined by history, not by the current visual configuration; three gaps follow.

The similarity-retrieval objective used here contains no explicit instance comparison. In the deployed stack [1], f is a SigLIP 2 image/text encoder [4] applied to detected crops and to the query string, and a query q is answered by $\arg\max_i \langle f(o_i), f(q) \rangle$ over stored observations; each item is scored independently, so the ordinal is not enforced.

Category-level recognition does not resolve instance identity. An open-vocabulary detector [2] localises categories but does not order instances within one: over a 15-category set, colour queries scored 0.84–0.93 against 0.17–0.32 for brand-level ones, yet detections stayed unordered (Appendix I).

Two metrics. *Instance Selection Accuracy* (ISA) is the fraction of static queries resolved to the intended object. *Persistent Referent Accuracy* (PRA) measures whether a symbolic handle stays bound to the same physical entity across a scene change: (i) bind handle h to object o at t_0; (ii) apply an intervention that does not remove or move o itself; (iii) re-query h at t_1. A trial succeeds if the selected object is o. The intervention removes a left neighbour or inserts an identical object to the left or right of the bound target; remove-left is the deployment-faithful condition, the insertions diagnostic controls. A method can therefore score highly on ISA and at zero on PRA (Sec. 4.2).

3 Two Reference Resolvers: Stateless Sorting (SS) and DIG

To evaluate PRA we instantiate two reference semantics with the same detector and spatial cue. *Stateless Sorting* (SS) treats the ordinal as a current-scene description, re-detecting and re-sorting at every query so that "#2" is whatever is currently second. *Deterministic Instance Grounding* (DIG) treats it as a persistent handle, storing the instance assigned at t_0 and reusing it later.

DIG is intentionally simple. It tests whether retaining the task-to-instance binding suffices to remove referent drift under the persistent-handle semantics; it is a constructive control, not a proposed memory or tracking method. PRA itself is mechanism-agnostic: an explicit registry, retained-observation matching, or learned memory may all satisfy it. At build time, detections of a category are ordered by a spatial key and each receives a persistent identifier, $R : (\text{class}, \text{id}) \mapsto (x, y, \theta)$, where (x, y, θ) is the pre-grasp base pose for that instance. A query is resolved by lookup in $O(1)$. Bindings stay fixed while the registry is valid. When the configuration is judged stale, the registry is rebuilt ($\approx$30 s per shelf).

4 Experiments

Experiments run on a Stretch 3 robot on a laboratory shelf with the manipulation stack of the order-picking deployment that motivated the study. In the primary remove-left condition the robot leaves the rack to pick and deposit the neighbour, then returns (Figure 2). Identical products remove appearance as a cue; each can carries a small index marker hidden from all robot perception and used only for experimenter-side scoring.

Protocol. At t_0 the resolver under test resolves the handle once, the robot approaches the selected can, and its marker is recorded as the bound referent o. At t_1 the selected object is the can the gripper is aligned to after the approach; the trial succeeds if its marker equals o. Trials with no selection at t_1 count as failures, reported separately from drift (Table A2). Wilson 95% intervals are reported. We ask whether appearance alone gives instance identity (Q1) and whether static success implies the handle stays bound after a scene change (Q2).

4.1 Q1: Does similarity-based selection provide physical-instance identity?

With identical cans, similarity retrieval falls toward chance as instances are added: 90%, 30% and 15% at $N = 2$, 3, 4 against chance of 50%, 33%, 25% (the $N = 4$ batch differed in setup). The distinct-can control isolates the cause: with the scene fixed and only the query changed, the same can scores 0% as "can #2" but 70% as "red can #2" (Table A1, Figure A1). With no term enforcing the ordinal, retrieval follows appearance and does not reliably resolve the requested can; this motivates the memory question, which Q2 tests.

Table 1: **Referent drift follows rank reassignment rather than scene change alone.** Static row is ISA, intervention rows PRA; rank is the target's re-derived ordinal after the intervention. SS fails in both rank-changing conditions and passes the rank-preserving control; DIG is a constructive control (stored lookup, no query-time perception).

Metric	Condition	Target rank	SS	DIG (control)
ISA	Static (no change)	$2 \rightarrow 2$	13/15	15/15
PRA	Remove left neighbour	$2 \rightarrow 1$	**0/14**	14/14
	Insert left neighbour	$2 \rightarrow 3$	**0/10**	10/10
	Insert right neighbour	$2 \rightarrow 2$	10/10	10/10



Figure 2: **Primary condition on real hardware: remove-left.** The robot picks and carries away the left neighbour; on return, re-indexing (SS) addresses a different can while the stored binding (DIG) addresses the original.

4.2 Q2: Does current-scene re-indexing preserve an earlier task binding?

Table 1 gives the result. SS is strong statically (13/15) but keeps the original referent in 0/14 trials once the neighbour is removed, while DIG does so in all 14 trials. Inserting an identical can to the target's left shifts its rank from 2 to 3 and SS scores 0/10; inserting to the right leaves the rank unchanged and SS recovers to 10/10. SS is correct under current-scene semantics throughout (it returns the object now at rank 2), so its PRA failure is not a sorting error: re-indexing does not preserve an earlier binding, and failure occurred only when the target's re-derived rank changed.

Two SS failures were recognition failures: a base rotation at t_1 left the remaining cans incompletely observed and no instance was selected (DIG avoids this only by not perceiving at query time); among complete-detection trials, drift was 12/12 for SS and 0/14 for DIG.

5 Conclusion

Static selection accuracy does not establish persistent instance identity: a system can resolve a fresh current-scene ordinal correctly while silently changing the referent of an earlier handle, and the rank-preserving control shows that this follows rank reassignment rather than scene change. The protocol is cheap and mechanism-agnostic (bind, intervene without moving the target, re-query); any system claimed to remember objects can be scored this way, and we suggest reporting PRA alongside static accuracy.

Two limits remain. The comparison is between a resolver that discards the binding and one that stores it; since the target stays put, a history-aware spatial matcher (retain the t_0 detection, match it to t_1 detections) may suffice here, and evaluating it is needed to show what PRA reveals beyond stateless re-indexing. For the same reason the protocol does not yet separate persistent identity from fixed-location lookup. Both are addressed by intervening on the target itself (observable relocation or removal), scored as re-resolution, abstention, or invalidation, where DIG's stale-binding cost (Appendix G) would bite first.

Acknowledgments

This work was supported in part by the Institutional Research Program of the Electronics and Telecommunications Research Institute (ETRI) [26ZB1100, Development of Embodied Self-Improving Artificial Intelligence and Robotics Technology, 10%], and in part by the Institute of Information &

communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220907, Development of AI Bots Collaboration Platform and Self-organizing AI) [90%].

References

[1] P. Liu, Z. Guo, M. Warke, S. Chintala, C. Paxton, N. M. M. Shafiullah, and L. Pinto. DynaMem: Online dynamic spatio-semantic memory for open world mobile manipulation. In *ICRA*, 2025.

[2] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan. YOLO-World: Real-time open-vocabulary object detection. In *CVPR*, 2024.

[3] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

[4] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[5] P. Liu, Y. Orru, J. Vakil, C. Paxton, N. M. M. Shafiullah, and L. Pinto. Demonstrating OK-Robot: What really matters in integrating open-knowledge models for robotics. In *RSS*, 2024.

[6] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf. SayPlan: Grounding large language models using 3D scene graphs for scalable robot task planning. In *CoRL*, 2023.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[9] H. K. Cheng and A. G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *ECCV*, 2022.

[10] J. Luiten, A. Ošep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:548–578, 2021.

[11] Y. Fan, X. Ma, R. Su, J. Guo, R. Wu, X. Chen, and Q. Li. Embodied VideoAgent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *ICCV*, pages 6342–6352, 2025.

[12] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

[13] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023.

[14] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*, 2023.

[15] Y. Dai, H. Fu, J. Lee, Y. Liu, H. Zhang, J. Yang, C. Finn, N. Fazeli, and J. Chai. RoboMME: Benchmarking and understanding memory for robotic generalist policies. In *ICML*, 2026.

[16] N. Chung, T. Hanyu, T. Nguyen, H. Le, F. Bumgarner, D. M. H. Nguyen, K. Vo, K. Yamazaki, C. Rainwater, T. Kieu, A. Nguyen, and N. Le. Rethinking progression of memory state in robotic manipulation: An object-centric perspective. In *AAAI*, volume 40, pages 3407–3415, 2026. doi:10.1609/aaai.v40i5.37337.

Appendix

A Full Results with Confidence Intervals

Table A1 reports every condition with success counts and Wilson 95% confidence intervals. Trial counts are modest by design: each trial requires a physical scene reset and a full navigate–observe–grasp cycle on real hardware.

Table A1: **Complete results.** Wilson intervals are used rather than normal approximations because several cells are at the boundary (0% or 100%), where the normal approximation is degenerate. The identical-can rows and the SS-study row are separate batches and differ in the query issued (the distinct-can control is plotted in Figure A1): the $N = 2$ and $N = 3$ rows used "red can #2", where every instance is a red can so the colour token narrows nothing; the $N = 4$ row queried each of "can #1" to "can #4" five times; the SS-study row used "can #2". Spacing between cans is given in parentheses; the $N = 4$ batch was run at 10 cm rather than 20 cm, so the drop from $N = 3$ to $N = 4$ confounds instance count with spacing.

Metric	Setting	Method	Succ./Trials	Rate (%)	Wilson 95% CI
ISA	Static, N=2 ident. (20 cm)	Similarity ("red can #2")	9/10	90.0	59.6–98.2
		DIG	10/10	100.0	72.2–100.0
ISA	Static, N=3 ident. (20 cm)	Similarity ("red can #2")	3/10	30.0	10.8–60.3
		DIG	10/10	100.0	72.2–100.0
ISA	Static, N=4 ident. (10 cm)	Similarity ("can #n")	3/20	15.0	5.2–36.0
		DIG	20/20	100.0	83.9–100.0
ISA	Static, N=3 distinct	Similarity ("can #2")	0/10	0.0	0.0–27.8
		Similarity ("red can #2")	7/10	70.0	39.7–89.2
		DIG ("can #2")	10/10	100.0	72.2–100.0
ISA	Static, N=3 (SS study)	Similarity ("can #2")	0/15	0.0	0.0–20.4
		Stateless Sorting	13/15	86.7	62.1–96.3
		DIG	15/15	100.0	79.6–100.0
PRA	Remove left	Stateless Sorting	0/14	0.0	0.0–21.5
		DIG	14/14	100.0	78.5–100.0
PRA	Insert left (rank 2 → 3)	Stateless Sorting	0/10	0.0	0.0–27.8
		DIG	10/10	100.0	72.2–100.0
PRA	Insert right (rank unchanged)	Stateless Sorting	10/10	100.0	72.2–100.0
		DIG	10/10	100.0	72.2–100.0

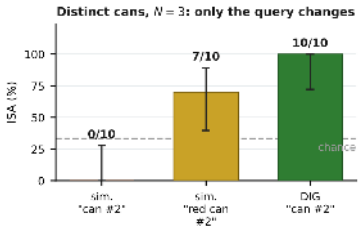


Figure A1: **Q1 distinct-can control.** The scene is fixed and only the query changes; the ordinal alone returns nothing correct, a colour cue recovers most of the accuracy, and the stored binding is exact. Counts and intervals in Table A1.

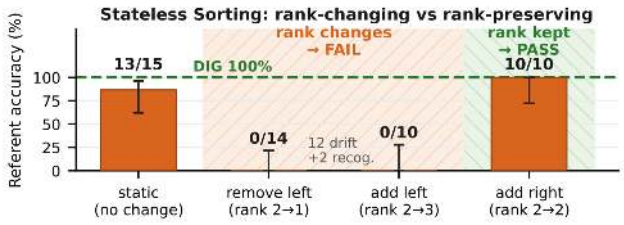


Figure A2: **Stateless Sorting across the static baseline and three intervention conditions (Table 1 as a chart).** Hatched bands mark rank-changing (fail) and rank-preserving (pass) interventions; the dashed line is DIG at 100%.

B Multi-Step Protocol



Figure A3: **Insertion conditions on real hardware (unedited detector output).** Green marks the bound instance in (i) and the can Stateless Sorting returns in (ii), (iii). In the remove-left batch (Figure 2) the leftmost can sits at the frame edge in every trial, a consequence of the robot's start pose; these insertion trials used a start pose shifted left so that all cans are fully in view.

C Failure Decomposition

Table A2 lists the primary cause for every trial. Figure A4 states the protocol and Figure 2 shows the remove-left condition on real hardware. Figure A3 shows the two insertion conditions, and Figure A2 plots the four-condition result.

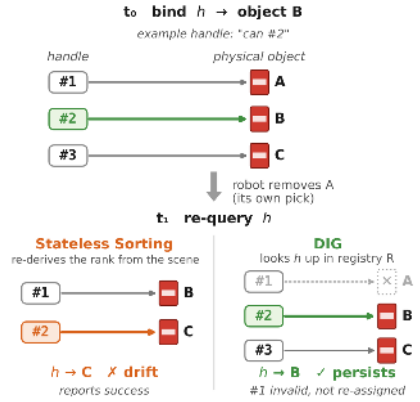


Figure A4: **Referent drift under the remove-left protocol.** A handle is bound at t_0, the scene changes without moving the bound object, and the same handle is re-issued at t_1.

Table A2: **Primary failure cause per trial.** The two SS recognition failures occurred when a small base rotation left the target scene incompletely observed, so no instance was selected. They are retained in the PRA denominator rather than excluded. Among the 12 trials with complete detections, all 12 exhibited referent drift.

Condition	Cause	Count
SS, remove-left (n=14)	Referent drift (index re-derived after removal)	12
	Recognition failure under viewpoint rotation	2
Similarity retrieval, static (n=15)	No ordinal constraint; highest-scoring view selected	15
DIG (all conditions)	—	0

D Baseline and Registry Implementation

Figure A5 contrasts the three strategies described below.

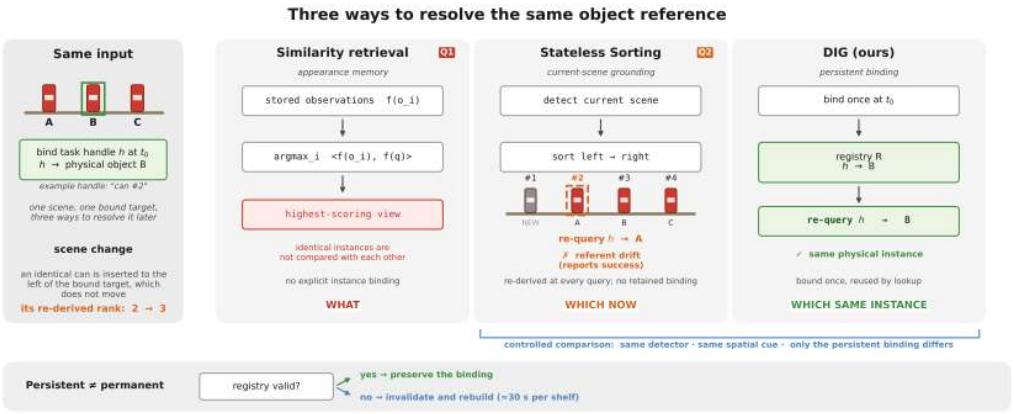


Figure A5: **Three reference-resolution strategies.** Similarity retrieval resolves appearance, Stateless Sorting re-derives the current-scene ordinal, and DIG retains the task-to-instance binding for later lookup. When judged stale, the registry is rebuilt.

Similarity retrieval. Observations are stored as per-frame crops of detected regions. Each crop and each text query is embedded with a SigLIP 2 image/text encoder [4], and the target is the stored crop maximising cosine similarity to the query embedding. Ordinal qualifiers in the instruction are passed through in the query string; they have no corresponding term in the objective. Ties are broken by the higher detector confidence.

Stateless Sorting (SS). At every query, the open-vocabulary detector [2] is run on the current observation, detections of the requested category are projected to world coordinates, and the k-th detection along the shelf axis (x, increasing left to right) is selected. No state is carried between queries; ties in x (within 2 cm) are broken by depth.

DIG. Identical to SS at build time, but the ordering is written once into the registry together with the pre-grasp base pose. Queries are dictionary lookups; the detector is not run at query time.

Query strings. A query names a detector class and an ordinal within it, written here as "can #2" or "red can #2". The class token narrows the candidate set only when it distinguishes the objects present: in the identical-can conditions every instance is a red can, so the token selects the whole set and the ordinal carries the entire burden. In the distinct-can condition only one can is red, so the same token becomes a discriminating appearance cue—which is what the 0%/70% contrast measures.

Static failures. The two SS failures in the static condition have the same mechanism as the interventions. In one the head camera did not capture can #1, so the ordering was re-derived over two cans

and rank 2 addressed a different can; the other shows the same substitution without a recorded note. Neither is a control or grasp error.

Trial counts. Block sizes follow the sessions in which they were run rather than a power analysis. The static and remove-left blocks were run in sessions of 15 and 14–16 trials; the insertion conditions were a later batch of 10 each. Wilson intervals in Table A1 reflect these counts.

Trial accounting. Each PRA trial establishes its own t_0 binding with the resolver under test; the static ISA block (Table A1) is a separate batch and is not reused as t_0 for the intervention trials. The remove-left condition contains 14 SS trials: 12 referent-drift failures with complete detections and 2 recognition failures (Trials 71–72) caused by a small base rotation that resulted in incomplete observation of the remaining cans at t_1, preventing valid instance selection. The latter are included in the PRA denominator but excluded from the referent-drift subset. DIG was run for 14 trials in the same condition.

Ground truth. Each physical can carries a small index marker recorded at t_0 and used to score whether the object approached at t_1 is the same physical instance. The markers are excluded from all robot perception, retrieval, and policy inputs and serve only experimenter-side scoring.

E Registry Construction and Cost

The registry is built from a single detection pass over the shelf. Detections of a requested category are clustered by centroid and sorted along the shelf axis; each cluster receives an index and the corresponding pre-grasp base pose (x, y, θ). Construction completes in $\approx$30 s, dominated by the detection pass rather than by bookkeeping. Query resolution is a dictionary lookup and invokes no vision-language model, so query cost is independent of the number of stored instances. The registry stores one compact pose record per registered instance; by contrast, the dense semantic map it replaced stored per-voxel embeddings over the full workspace.

F Why Not Track Instances Instead?

Tracking is a related but different question. Tracking maintains identity through observation and is the natural alternative when objects move; two properties of this setting make tracking metrics a poor fit for the question asked here, though not inapplicable. First, our objects are static between queries and the robot navigates away between them, so there is no continuous observation to track through. Second, when the robot returns, instance-specific appearance evidence is insufficient in our setting—the products are visually interchangeable at the resolution and viewpoints available to the robot—and re-identification [7] faces the same obstacle. The protocol therefore evaluates retention of a task-level binding across an observation gap; whether a given association mechanism achieves that retention is exactly what PRA scores, and a history-aware spatial matcher is the first such mechanism to test (Sec. 5).

G Limitations in Detail

Spatial key. The current registry sorts along one axis, which suits a single shelf row. Multi-row shelving requires a two-dimensional key, e.g. (class, row, column); this is a straightforward extension we have not yet evaluated. **Static assumption.** Identity is bound at build time and remains valid until the next rebuild. Objects that move between queries without triggering a rebuild will be mis-addressed. In our deployment, the $\approx$30 s rebuild time allows this to be handled operationally.

Scale. Results come from one site, one platform, and one product family, with 10–20 trials per condition. The confidence intervals in Table A1 reflect the uncertainty associated with these modest trial counts. **Context length is not controlled.** Our comparison isolates stateless re-derivation, not context length. Stateless Sorting never sees t_0: it re-derives the ordering from the t_1 scene alone. Retaining observations is not by itself the answer: the deployed stack [1] stores every observation

and retrieves by language similarity, yet scores 0/15 on the static identical-can query (Table A1), so it could not be evaluated for persistence and the intervention comparison is between SS, which is strong statically, and DIG. A system that retained the t_0 observation and matched it explicitly against t_1 could in principle recover the binding, since the bound can does not move. We do not run that baseline, so our results do not establish that the failure survives a longer context window, only that it is not repaired by observing the current scene more completely or by retaining observations for similarity retrieval. Separating stored bindings from recomputed ones requires an explicit-matching baseline and an intervention under which matching is ambiguous (target relocation or removal); both are left to future work, as set out in Sec. 5.

Task length. The protocol spans two query times with one robot action between them. It is history-dependent but not long-horizon: the partial observability here is over referential history rather than over an extended action sequence.

Open question. Whether vision-language-action models are subject to the same failure is a hypothesis we pose rather than answer; we have not examined their training objectives.

Invalidation is asserted, not measured. We report that a shelf registry rebuilds in $\approx$30 s, but we do not evaluate invalidation itself: there is no experiment in which a stale registry is detected and replaced under test. Deciding when a binding has become stale is left open.

Removal of the bound target. By construction PRA applies interventions that leave the bound object in place, so we never test the case where the target itself is taken. There DIG would return a stale binding rather than a re-derived wrong one; whether that is preferable, and how staleness should be signalled, we do not evaluate.

Interventions tested. Two interventions change the target's re-derived rank (removing a left neighbour, inserting a left neighbour) and one leaves it unchanged (inserting a right neighbour). The scene changes in all three, and the target itself is never moved, so the contrast between the first two and the third isolates rank change rather than scene change. The conclusion is drawn within this shelf setting and one product family; we have not varied the spatial key itself.

H Deployment Site



Figure A6: **Real-world deployment of the Stretch 3 platform at the operating order-picking facility.** The platform operates around mobile racks stocking many visually identical units; the registry is rebuilt when rack placement changes. All panels are from the facility; the controlled experiments were run on a laboratory shelf.

The registry is a drop-in replacement for the instance-selection step: the detector, the map, and the pick policy are unchanged.

Deployment notes. The deployment observations come from the facility shown in Figure A6. The same detector, registry, and pick policy used in the controlled experiments run unchanged there; only the registry contents are rebuilt when a rack has moved.

I Detector Characterisation

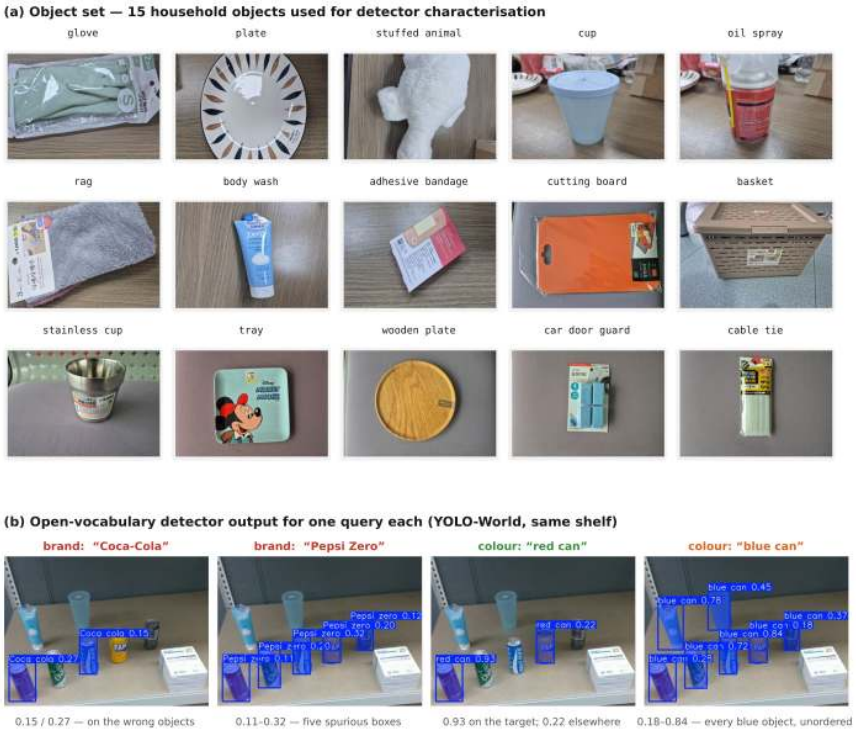


Figure A7: **Object set and open-vocabulary detector output.** (a) The 15-category household object set used for detector characterisation; the identity experiments of Sec. 4 use one product family (cans). (b) Raw YOLO-World [2] output on one shelf for four text queries. Brand-name queries return low, unstable confidences and land on the wrong objects; colour queries are confident but return every matching instance with no ordering—recognition answers *what*, not *which*.

Figure A7 shows the object set and the detector behaviour summarised in Sec. 2. The confidences quoted in the main text are from the representative queries shown. In these examples, colour queries produced stronger detector responses than brand-name queries, and raising the threshold removed the brand-name boxes rather than cleaning them up. This is a descriptive observation, not a causal analysis of the encoder.