# SAME OBJECT, OR ITS TWIN? DIAGNOSING INSTANCE ASSOCIATION ACROSS SCENE REVISITS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Persistent object maps must maintain physical identity across changes in position and viewpoint. An incorrect association can reflect absent observational cues, difficulty reading them from features, or limitations in how matching uses them. We test these explanations through simulated paired revisits of otherwise identical chairs with artificial persistent color markers. Representations share oracle masks, views and the full candidate map. In the primary strong-color swap condition with correlated cameras, an RGB descriptor recovers most identities. A development-fitted linear probe decodes marker colors from MASA-R50 features, yet maximum view-pair cosine matching frequently selects the exchanged partner. Holding view features and candidates fixed, normalized-mean aggregation increases correct associations from 2 to 10 of 36 queries. Penalizing differences in the probe's continuous output raises this count to 15. Aggregation accounts for most of both this recovery and the accuracy loss on other map objects, while readout gains vary across representations and camera conditions. This partial recovery shows that evaluating readable identity cues requires testing how fixed features are aggregated and compared under full-map candidate competition.

## 1 INTRODUCTION

Object-centric maps fuse observations into object-level records (Gu et al., 2024; Werby et al., 2024; Maggio et al., 2024). A chair can move during an unobserved interval while another takes its place. Category and location can remain correct while history is attached to the wrong individual.

Visual embeddings support association (Oquab et al., 2024; Li et al., 2024). Their performance reflects both recoverable information and multi-view comparison. Final accuracy conflates these factors. Can changing aggregation and comparison recover identity from fixed features with readable cues?

Controlled paired revisits provide this test (Figure 1). Unmarked exact twins establish an observation-level ambiguity boundary under our sensing model. Artificial persistent colors follow identity through screen-space overlays, supplying distinguishing evidence beyond that boundary. With shared oracle masks, views and full candidate maps, RGB matching tests input sufficiency and a linear probe tests marker decoding. Binary marker decoding alone does not establish full-map identity recovery. We therefore vary aggregation and penalize query–candidate differences in continuous probe outputs with view features and candidates fixed.

MASA features support correct marker decoding alongside wrong partner selection. In the primary comparison, aggregation and readout partially recover identities, with aggregation driving most recovery and ordinary-object accuracy loss (Figure 2).

We make three contributions.

- **Controlled cue comparisons.** Paired revisits separate observation ambiguity from marked inputs that support physical-identity association.
- **Readable cues and failed matches.** Same-query analysis reveals correct MASA marker decoding alongside failed full-map identity ranking.
- **Fixed-feature interventions and their limits.** Aggregation and readout changes expose partial recovery, condition-dependent gains and ordinary-object accuracy costs.



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.

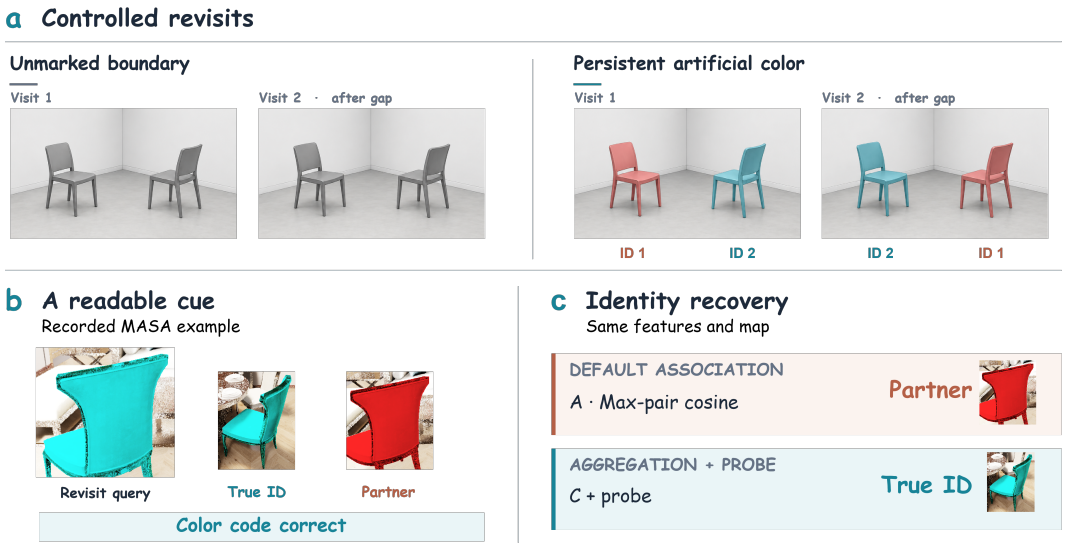


Figure 1: **A readable cue becomes useful for identity association.** (a) AI-generated conceptual illustration with explanatory identity labels. (b,c) Recorded MASA inputs and saved decisions for a query scored against the complete map. The color code is decoded correctly. Default association selects the partner. Changing aggregation and adding the probe recovers the true identity from fixed features. Figure 2 separates the interventions and reports population effects. Appendix C.4 gives the example selection.

## 2 Related Work

**Feature readout and task use.** Linear probes measure how readily a property can be decoded from fixed representations (Alain & Bengio, 2018). Controlled language experiments show that such recoverability can coexist with task irrelevance (Ravichander et al., 2021). We examine this distinction in cross-visit instance association by pairing marker decoding with full-map identity ranking. With visual features and candidates fixed, we vary aggregation and incorporate the learned readout into association scores. These interventions test whether accessible cues improve identity ranking and expose the competition that remains among other map candidates.

**Open-vocabulary mapping and scene graphs.** Persistent maps assign identity during fusion. ConceptGraphs uses a visual-geometric threshold (Gu et al., 2024), ConceptFusion stores open-set features (Jatavallabhula et al., 2023), and Hydra, HOV-SG, Clio, and Embodied VideoAgent organize the resulting memory (Hughes et al., 2022; Werby et al., 2024; Maggio et al., 2024; Fan et al., 2025). Video models propagate identity through streaming masks or appearance matches (Ravi et al., 2025; Carion et al., 2026; Li et al., 2024). A revisit gap interrupts temporal continuity and makes these associations depend on cues shared across visits. We isolate this step to test how aggregation and comparison use instance-specific evidence.

**Association uncertainty.** Multiple-hypothesis tracking retains association histories (Reid, 1979), and joint probabilistic data association marginalizes gated assignments (Bar-Shalom & Fortmann, 1988). Object-level SLAM uses semantic landmarks and probabilistic matches (Salas-Moreno et al., 2013; Bowman et al., 2017; Doherty et al., 2019). Exchangeable labels create related inference ambiguities (Jasra et al., 2005), while labeled random finite sets propagate multi-object hypotheses (Vo et al., 2014). These frameworks represent uncertainty and exploit distinctive measurements or temporal dynamics. Our controlled design separates observation-level ambiguity from recoverable cues that an association score fails to exploit.

**Benchmarks for changing scenes.** ScanNet annotates static indoor reconstructions (Dai et al., 2017). 3RScan adds natural changes and annotated correspondence alternatives (Wald et al., 2019),

while Rescan propagates instance segmentation through sparse rescans (Halber et al., 2019). Controlled revisits let us introduce identity cues while retaining a known physical correspondence. 3RScan extends the evaluation to naturally ambiguous annotations.

**Selective and set-valued prediction.** Selective prediction trades coverage for risk through rejection (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Set-valued classification returns plausible label sets under error control (Sadinle et al., 2019). Both can represent association uncertainty. Our auxiliary evaluation examines commitment reliability under shared budgets and stated costs. Open-set recognition addresses inputs outside known categories (Scheirer et al., 2013), while physical identity also introduces uncertainty among individuals within a category.

## 3 Physical Identity and Its Observation Boundary

The analysis establishes an observation-level boundary for the cue-use experiments. Exact twins instantiate this boundary under the sensing model below. Persistent markers supply evidence beyond it. Proofs and decision conditions appear in Appendix A.

**Setup.** Two visits are separated by an unobserved interval. Each object $g \in \mathcal{G}$ has an asset $A(g)$ and visit-$t$ pose with centroid $p_t(g)$. Twins $T \subseteq \mathcal{G}$ share an asset, with $|T| = K \geq 2$. Swapping $a, b \in T$ exchanges their complete poses. Physical correspondence remains $a \mapsto a$ as locations change.

A fixed, possibly stochastic sensing process generates $O = (O_1, O_2)$ from the multisets $S_t = \{\{(A(g), \mathrm{pose}_t(g)) : g \in \mathcal{G}\}\}$. Identity enters observations through asset and pose. Registration error is bounded by $\varepsilon$. A decision links a visit-1 record (MERGE), registers a new individual (NEW), or leaves identity unresolved (ABSTAIN) (Chow, 1970).

**Proposition 1** (Two-world observational equivalence)**.** *From a valid scene configuration, let $W_1$ apply a final $\mathrm{swap}(a, b)$ and let $W_0$ omit it. Under the sensing model above, the worlds share their observation law while their slot identities differ by $(a\, b)$. The likelihood ratio is $1$ almost surely on their common support, so posterior odds equal prior odds. For any rule $\delta$ that names a record for each exchanged slot, its expected within-pair 0–1 errors satisfy $\mathrm{err}_{W_1}(\delta) + \mathrm{err}_{W_0}(\delta) \geq 1$.*

Equal likelihoods leave posterior preferences to the prior. Position illustrates how such a preference becomes a systematic partner error.

**Assumption 1** (Relative margin)**.** *For each swapped twin, the centroid distance from its visit-2 position to its partner's visit-1 position, plus $2\varepsilon$, is smaller than that to every other same-category visit-1 candidate, native chairs included.*

**Proposition 2** (Systematic inversion under a quasi-static prior)**.** *A* pose-monotone *rule selects the nearest same-category predecessor whenever it is unique. Under Assumption 1, it selects the partner on every swapped slot. For a prior supported on $\{W_0, W_1\}$ with $\pi_0 > \pi_1$ and a decoder restricted to these two correspondences, the Bayes rule under 0–1 loss selects $W_0$ almost surely. Both rules therefore have conditional accuracy $0$ and partner error $1$ on $W_1$.*

Persistent markers change the observation law by making appearance identity-dependent. They allow us to test cue use beyond exact-twin ambiguity. Lemma 1 gives the exchangeable-prior chance baseline. Corollary 1 specifies when a reject action preserves unresolved identity.

## 4 Controlled Revisit Design

Each simulated episode opens a Kujiale interior in NVIDIA Isaac Sim,[1] records visit 1, applies registered edits, and records visit 2. Exact twins share a native chair's USD mesh and texture. The generator retains physical identity as duplicate count, displacement, swaps, insertion, removal and lighting change. We use this infrastructure for two complementary experiments with distinct association protocols.

[1] https://docs.isaacsim.omniverse.nvidia.com/4.5.0/

The core fixed-feature diagnosis introduces persistent identity cues. Six scenes provide 18 scene-placement groups, each with no-swap and pure-swap worlds under correlated and independent cameras. The twins receive identity-specific screen-space color overlays across strengths and paired context/neutral backgrounds. Cameras, depth and target masks match across worlds, while independently captured RGB differs. The diagnosis uses GT instances, unchanged views and complete candidate maps at zero and strong color. Every scoring rule independently ranks all map records with the same output semantics. Masked RGB supplies a same-input association control. Separate rerun-SAM3 and observation controls assess sensitivity to perception and background (Appendix B.2).

The auxiliary policy comparison instead uses the original unmarked swap panel, which balances scenes and duplicate counts and applies a displacement before the final pose exchange under exact alignment. Policies share a per-episode SAM3–DINOv2 perception cache (Carion et al., 2026; Oquab et al., 2024) while retaining their own gates and assignment constraints. SAM3 segments scene-vocabulary concepts, DINOv2-L supplies mask-pooled features, and depth provides 3D centroids. A fixed within-visit procedure groups detections by category, centroid and appearance, then merges fragments (Appendix B). Multiple records can represent one physical object. Evaluation maps selected records to annotated identities, with ground-truth identity fields reserved for generation, calibration and evaluation.

For a real-data comparison, we convert 3RScan reference/rescan pairs (Wald et al., 2019) to the same interface using annotated instances. The moved-lookalike panel contains 88 queries. Released mappings define a designated correspondence, and ambiguity annotations provide additional plausible matches. We separate designated agreement, directly annotated alternatives and other choices. These labels support correspondence analysis. Independent verification of physical identity is unavailable for ambiguous objects. Calibration pairs are held out from evaluation (Appendix B.9).

Across the simulated evaluations, outcomes distinguish correct, specified-partner and other incorrect merges, correct or wrong NEW, and ABSTAIN. Unknown labels remain explicit. The auxiliary panels also compare events with uniform-within-group selection, whose probability is $1/K$ for the designated ID or a specified partner and $(K-1)/K$ for any other member. A second baseline samples same-category map records, retaining fragments. Auxiliary intervals resample simulation episodes or real reference scans and recompute query-specific baselines. Per-scene and leave-one-scene-out results describe sensitivity within the fixed scenes. The following section defines the shared scoring interface and grouped uncertainty for the core diagnosis.

## 5 Fixed-Feature Association Diagnostics

The diagnosis separates three questions. A simple color descriptor tests whether the shared inputs support identity association. A binary probe tests marker-code readout from learned features. Reusing its continuous output in full-map association tests whether this information can correct identity ranking. The latter intervention holds the encoder, unit view features, entity membership, view selection and candidate map fixed.

We first vary how views contribute to a match. Let $u_{ir}$ be entity $i$'s unit feature in view $r$, and $m_i$ its mean over views. For query $q$ and map candidate $j$, we compare

$$S_A(q,j) = \max_{r,s} u_{qr}^\top u_{js}, \qquad S_B(q,j) = m_q^\top m_j, \qquad S_C(q,j) = \frac{m_q^\top m_j}{\|m_q\|\|m_j\|}.$$

A retains the best view pair, B averages every pair cosine, and C normalizes the entity means before comparison. Since $S_B = \|m_q\|\|m_j\|S_C$, B also weights similarity by within-entity agreement across views. Each rule independently ranks every map record, retaining fragments. Exact top ties or unavailable comparisons remain unresolved.

We then test whether a marker-specific readout can improve these rankings. For each representation, a ridge probe fitted on strong-color development observations produces a continuous output $r(i)$ from the normalized entity mean. The same model applies to both visits, all candidates and every test condition. We compare

$$S_{a,\beta}(q,j) = S_a(q,j) - \beta \left( \frac{r(q) - r(j)}{\sigma_{\text{dev}}} \right)^2, \qquad a \in \{A, B, C\},$$

where $\sigma_{\mathrm{dev}}$ is the standard deviation of development training outputs. A development-only objective balancing marked-target and ordinary-object accuracy selects one weight per aggregation and representation, held fixed across test conditions. Distance-only matching isolates the penalty term. Ordinary candidates remain in association competition and are excluded from binary marker accuracy.

To distinguish the learned direction from a generic score perturbation, each of 32 fixed random directions receives the same scaling and development-selection budget as the learned readout. No-color, unswapped and independent-camera conditions test the scope of recovery. Accuracy retains all target slots, including missing true candidates. Ordinary-query outcomes are reported separately. Paired intervals resample scene-placement groups while keeping all associated conditions together. Appendix B.4 gives the coefficient grid, tie rules, unavailable-score handling and fixed-probe uncertainty scope.

## 6 Experimental Results

We first examine marked revisits and the fixed-feature interventions. Unmarked swaps then connect the diagnosis to the observation boundary. Decision costs and natural-lookalike annotations establish complementary limits on interpreting an association.

### 6.1 From Color Evidence to Identity Association

Persistent colors supply identifying evidence beyond the exact-twin construction. Our primary comparison retains all 36 target slots under swaps and correlated cameras. Representations share GT-instance masks, views and the complete candidate map (Figure 2).

Masked RGB identifies every strong-color target whose true candidate is present. Learned-feature maximum-cosine matching rarely recovers these same identities. Yet the fixed probe decodes the marker from MASA features on most queries (Figure 2a). The DINO probe provides weaker evidence of transferable decoding. Binary marker classification and complete identity association have different prediction targets. Their contrast motivates putting the readout into the association score, where it competes with every map record.

Changing aggregation alone recovers a substantial part of the missed identities. Both mean-based rules recover more MASA identities than maximum-pair matching (Figure 2b). Recovery occurs with and without normalizing the entity means, implicating the combination of view evidence in the original failure. DINO also benefits from pairwise averaging, with a different ordering of B and C. Aggregation also accounts for most of the ordinary-object loss in the primary MASA A-to-C+probe comparison. Thus view combination affects both the targeted correction and its main cost on the surrounding map.

B and C share mean-feature directions but differ in their weighting by feature norms. For a fixed query, B scales each candidate's directional similarity by its mean-feature norm, which reflects agreement across that candidate's views. B incurs a larger ordinary-object loss than C for both representations in the primary condition (Appendix C.4). Thus combining views changes candidate competition through both direction and cross-view consistency. Normalizing the mean removes this candidate-dependent factor.

The readout provides an additional, representation-dependent correction. For MASA, the development-selected penalty adds a positive paired gain over C, recovering missed identities without introducing a new marked-target error in this condition (Figure 2b). Its accuracy exceeds every random-direction control given the same selection budget. The corresponding DINO intervention leaves target accuracy unchanged. The readout adds a smaller ordinary-object loss than the preceding aggregation change. At a common weight, the learned readout also improves MASA association under the other aggregations, so its usefulness extends beyond normalized-mean matching. Larger penalties eventually reduce ordinary-object accuracy in both representations. The complete grid appears in Appendix C.4, with development-selected weights fixed throughout evaluation.

The remaining errors reveal the limits of a binary marker. The MASA readout reduces partner inversions, but some decisions move to other incorrect map records. On queries with comparable true and partner candidates, the readout changes MASA's mean true-minus-partner margin from negative to

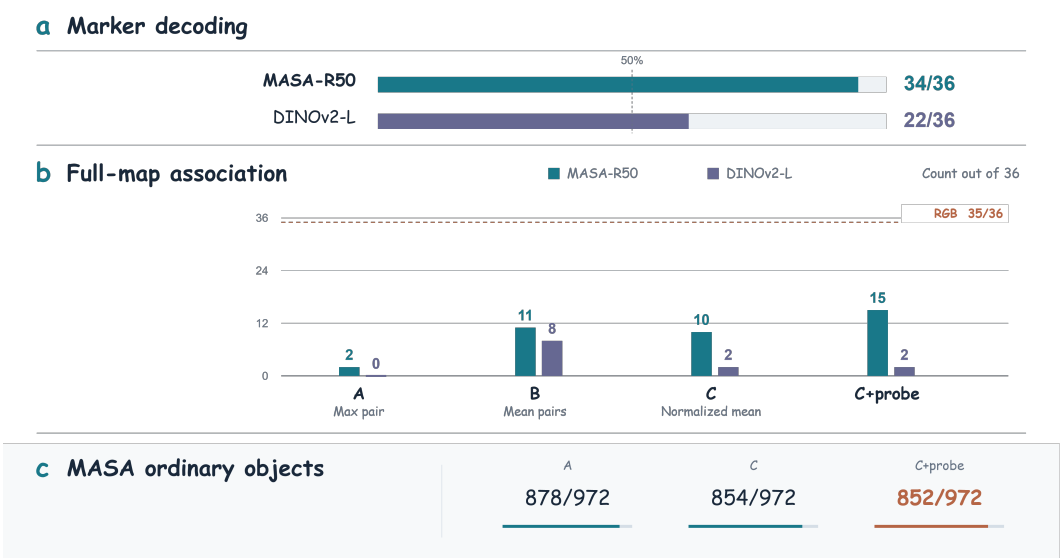


Figure 2: **Readable markers support partial identity recovery, with ordinary-object costs.** (a) Binary marker-code prediction on visit-2 targets. (b) Full-map identity association with fixed features under A, B, C and C+probe. Both tasks retain all target slots. (c) MASA ordinary-object correctness. The paired 95% interval for MASA's C+probe target-accuracy gain over C is [2.8, 27.8] percentage points, resampling scene-placement groups. Full controls appear in Appendix C.4.

positive. On the corresponding available support, the mean margin against the strongest wrong candidate remains negative, and mean true-candidate rank does not improve (Appendix C.4). Matching by readout distance alone predominantly selects these other objects. The marker direction can favor the true individual over its partner while leaving other map candidates competitive. Its value in full-map association therefore depends on how it combines with the remaining visual evidence.

The broader conditions show why the primary recovery is conditional. Averaging loses correct links in unswapped independent-camera revisits. Under independent cameras with swaps, the readout improves MASA's point estimate, with a paired interval spanning zero. Camera variation changes both baseline difficulty and the evidence for the added readout. These results motivate checking ordinary queries and camera changes alongside a targeted correction.

The separate rerun-perception, background and viewing-direction controls retain frequent partner errors under their respective inputs (Appendix C.5). We next return to the auxiliary unmarked panel to examine association preferences without an added identity cue.

### 6.2 Association at the Unmarked Boundary

The auxiliary swap panel compares CG-style visual-geometric matching, EVA-style category gating, nearest-neighbor appearance, cosine–Hungarian assignment, single-exemplar and TIER proxies, and three probabilistic policies (PDA-DP, PDA-MM, PDA-CONST) (Gu et al., 2024; Fan et al., 2025; Li et al., 2024; Mu et al., 2016; Doherty et al., 2019; Frey et al., 2019). These adaptations isolate association components. The TIER hedge restores rejection from the proxy's possibly-same band. A separate source component retains NEW with a possibly-same edge. MASA-R50 supplies a second representation and a source component adapted to whole-visit entity batches. Appendix B specifies their relation to the original systems. The diagnostic CUE variants additionally test cue fusion and rejection, as detailed below.

On queries from exchanged objects, every forced-commit policy favors the specific partner over the correct identity (Table 1). Treating every selection of an unannotated candidate as correct still gives accuracy below the uniform-within-group baseline. Ground-truth position selects the partner on every evaluated swapped slot, satisfying the geometric premise for inversion. With perceived

Table 1: **Swaps favor partner errors, while natural revisits require a different interpretation.** Simulation scores physical identity and the specific exchanged partner. 3RScan scores designated agreement and selection of another container member, which may be an annotated alternative. Unknown marks merges to unannotated candidates. Remaining actions include other errors and, for selective rules, abstention. Table 3 gives the full partition. The last row gives event-specific uniform-within-group baselines.

| | Simulation exchanged pair | | | 3RScan lookalike container | |
|---|---|---|---|---|---|
| Policy | Correct ↑ | Swapped-partner error ↓ | Unknown | Designated agreement | Other container member |
| CG-style | 0.019 | 0.906 | 0.019 | 0.852 | 0.057 |
| PDA-MM | 0.019 | 0.962 | 0.009 | 0.852 | 0.114 |
| PDA-DP | 0.019 | 0.925 | 0.009 | 0.693 | 0.159 |
| Cosine–H | 0.019 | 0.811 | 0.028 | 0.330 | 0.534 |
| NN | 0.028 | 0.802 | 0.038 | 0.398 | 0.534 |
| EVA-style | 0.028 | 0.792 | 0.009 | 0.534 | 0.375 |
| TIER proxy | 0.047 | 0.792 | 0.028 | 0.341 | 0.455 |
| Single-exemplar | 0.038 | 0.689 | 0.019 | 0.205 | 0.534 |
| PDA-const | 0.000 | 0.660 | 0.009 | 0.341 | 0.148 |
| CUE | 0.000 | 0.226 | 0.000 | 0.330 | 0.216 |
| CUE+POS | 0.009 | 0.774 | 0.000 | 0.761 | 0.068 |
| Uniform-within-group baseline | 0.330 | 0.330 | | 0.253 | 0.747 |

inputs, both appearance-only and position-only controls favor the partner. Adding position to CUE reduces abstention and increases partner selection (Table 3). Appendix C.1 gives full outcomes and event-specific chance baselines.

Replacing DINOv2 with MASA-R50 while retaining entity membership and detector boxes preserves frequent partner errors under both cosine–Hungarian and the adapted source component (Appendix C.9). In the DINO-based CUE model, partner preference already appears before calibration (Appendix C.3). These observations motivate examining the reliability and cost of the resulting commitments.

### 6.3 SELECTIVE RISK AND DECISION COST

The auxiliary policies raise a separate question about commitment reliability. This section evaluates their fixed confidence scores and calibrated rejection rules on the broader simulated ambiguity slice and held-out 3RScan queries. These populations and policies differ from the marked fixed-feature diagnosis in Section 6.1. Figure 3 therefore compares auxiliary policy decisions and does not evaluate C+probe.

Among these policies, CUE combines DINOv2 appearance, category consistency and a within-visit $\mathrm{Sim}(3)$-invariant neighborhood descriptor. Isotonic fits on designated easy-cell episodes map each cue to a clipped log-likelihood ratio (Zadrozny & Elkan, 2002). The model adds these to a NEW prior and normalizes over map records and NEW. It commits above fixed thresholds and otherwise leaves identity unresolved (Chow, 1970; El-Yaniv & Wiener, 2010). CUE+POS adds absolute cross-visit position. Both variants require comparable cross-visit coordinates because the NEW prior counts map candidates within 3 m of the query. Appendix B.7 gives its calibration and decision rules.

We first connect conditional swap errors to decision cost by scoring each fixed prediction under both world assignments. The cached control retains one observation and policy computation at each spatial slot. A hypothetical swap frequency $q$ combines the two costs as

$$R_\pi(q) = (1-q)R_\pi(W_0) + qR_\pi(W_1).$$

In an uninformative two-identity model with error cost $C$ and reject cost $c$, static, swap and reject actions cost $Cq$, $C(1-q)$ and $c$. Deferral has the lowest cost when $c < C\min\{q, 1-q\}$. Its

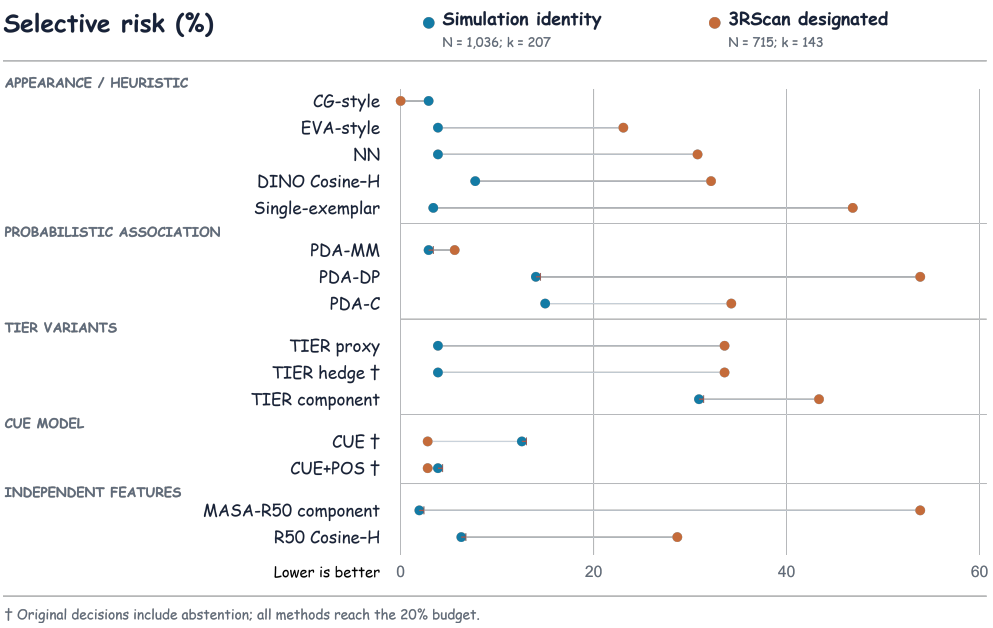


Figure 3: **Equal commitment budgets reveal different rankings across datasets.** The auxiliary policies share labelable queries and a budget within each dataset. Simulation counts identity errors. 3RScan counts designated-correspondence disagreements, including annotated alternatives. Gray lines connect dataset estimates. Short orange extensions bound unknown-label outcomes. The fixed-feature readout intervention is evaluated separately in Figure 2. Appendix C.6 gives paired differences and complete curves.

value therefore depends on both scene dynamics and the relative rejection penalty. We vary $q$ as a sensitivity parameter (Appendix C.2).

To compare confidence rankings empirically, equal budgets separate the number of commitments from their reliability. Policies rank available commitments on the same action-independent labelable queries using fixed confidence scores and label-independent tie keys. They share an integer budget wherever attainable. Selective rules stop at their coverage ceiling. At approximately 20% coverage on the simulated ambiguity slice, CG-style matching incurs fewer wrong merges than CUE (Figure 3). Calibration-selected thresholds use the same held-out set, cost grid and conservative treatment of unknown candidate labels. Appendix C.6 gives exact budgets, full outcomes and paired differences.

Under the cost setting in Table 33, thresholded PDA-MM attains lower simulated cost than both the thresholded and expected-cost CUE rules. On 3RScan, the CUE+POS expected-cost rule attains lower designated-correspondence cost than thresholded CG-style matching. Score quality and the relative error and rejection penalties jointly determine these comparisons. The full cost grid includes all-abstain and shows their sensitivity to the operating costs. To interpret the real-data costs, we next examine the lookalike annotations.

### 6.4 NATURAL LOOKALIKES

Designated agreement and annotation support answer different questions in the real lookalike panel (Table 1). Every query also has a directly annotated alternative. Selecting this alternative disagrees with the designated correspondence while remaining consistent with the ambiguity annotation. A nearer lookalike can itself be an annotated alternative, as Figure 4 illustrates. These comparisons concern released correspondence labels. Physical identity remains unverified for ambiguous objects. Appendix C.7 separates designated matches, annotated alternatives, unsupported links and abstentions across the full evaluation set.

Keeping the retained queries and decisions fixed, accepting directly annotated alternatives reduces disagreement for nearest-neighbor and single-exemplar matching. The CUE disagreements at the common budget remain unchanged because they are NEW decisions on existing objects (Table 37). This comparison isolates label interpretation from ranking and coverage. Alternative-aware evaluation distinguishes annotation-supported links from unsupported ones, while physical-identity verification would require additional evidence.

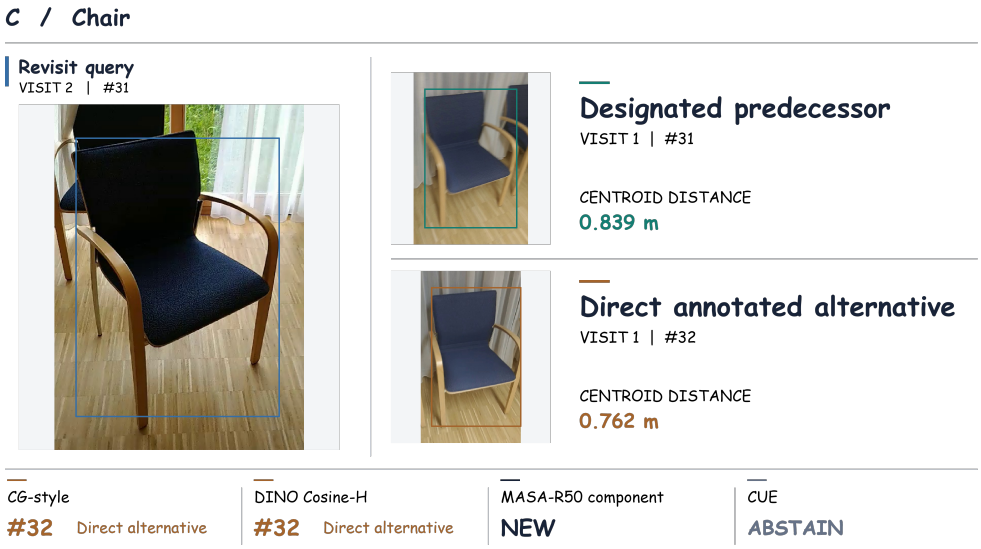


Figure 4: **A closer lookalike can be an annotated alternative.** The query, designated predecessor and direct alternative illustrate correspondence ambiguity without verifying physical identity. Distances use aligned annotated centroids. Cases are selected by visibility and image size independently of policy outcomes (Appendix C.7).

## 7 CONCLUSION

Controlled revisits separate cue availability, marker decoding and complete identity association. Changing aggregation and adding a marker readout partially recover missed identities from fixed features and an unchanged candidate map. Their representation-dependent gains and ordinary-object losses delimit the improvement. Evaluating visual representations for persistent identity therefore requires testing how their features are aggregated and compared. Observation ambiguity, decision costs and annotation semantics provide complementary boundaries for interpreting the resulting identity claims.

Limitations. The evidence covers chairs in six fixed interiors and artificial screen-space colors. Probe development uses one separate placement in a scene shared with evaluation, limiting scene-generalization claims. RGB differences persist between independently captured worlds. Oracle masks and viewing-direction selection define diagnostic controls. Route changes also affect geometry, membership and prototypes. Independent physical-identity verification is unavailable for ambiguous 3RScan objects. Deployment ambiguity and swap frequencies remain unknown. Immediate decision costs describe the stated cost grid. Stateful multi-visit experiments are needed to measure downstream resolution, history contamination and delay.

### ETHICS STATEMENT

This study uses simulated scenes and released 3RScan data, with no new human-subject collection. Applications of unresolved identity must account for the costs of delayed decisions and incorrect links.

### REPRODUCIBILITY STATEMENT

Appendices A–C provide proofs, protocols, complete outcomes and uncertainty analyses. Experiment artifacts link commands, configurations and tables to their inputs. Rendering reproducibility is assessed separately from deterministic cached analysis.

### AI USE STATEMENT

AI tools assisted study design, implementation, result analysis and interpretation, and manuscript and figure preparation. The authors take responsibility for the final content and reported results.

## REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and Data Association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press, Boston, 1988. ISBN 0-12-079760-7.

Sean L. Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J. Pappas. Probabilistic data association for semantic SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1722–1729, 2017. doi: 10.1109/ICRA.2017.7989203.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*. OpenReview.net, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2432–2443. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.261.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, pp. 2632–2652, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0b408293619f725fd30162af057e531a-Paper-Conference.pdf.

Kevin J. Doherty, Dehann Fourie, and John J. Leonard. Multimodal semantic SLAM with probabilistic data association. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2419–2425, 2019. doi: 10.1109/ICRA.2019.8794244.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied VideoAgent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6342–6352. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00598.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. doi: 10.1214/aos/1176342360.

Kristoffer M. Frey, Ted J. Steiner, and Jonathan P. How. Efficient constellation-based map-merging for semantic SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1302–1308, 2019. doi: 10.1109/ICRA.2019.8794452.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

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 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Maciej Halber, Yifei Shi, Kai Xu, and Thomas A. Funkhouser. Rescan: Inductive instance segmentation for indoor RGBD scans. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 2541–2550. IEEE, 2019. doi: 10.1109/ICCV.2019.00263.

Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. In *Robotics: Science and Systems XVIII*, New York City, NY, USA, 2022. doi: 10.15607/RSS.2022.XVIII.050.

Ajay Jasra, Christopher C. Holmes, and David A. Stephens. Markov chain monte carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005. doi: 10.1214/088342305000000016.

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 B. Tenenbaum, Celso Miguel de Melo, K. Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Robotics: Science and Systems XIX*, Daegu, Republic of Korea, 2023. doi: 10.15607/RSS.2023.XIX.066.

Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segù, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18963–18973. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01794.

Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. doi: 10.1007/s11263-020-01375-2.

Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3D scene graphs. *IEEE Robotics and Automation Letters*, 9(10):8921–8928, 2024. doi: 10.1109/LRA.2024.3451395.

Beipeng Mu, Shih-Yuan Liu, Liam Paull, John Leonard, and Jonathan P. How. SLAM with objects using a nonparametric pose graph. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4602–4609, 2016. doi: 10.1109/IROS.2016.7759677.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. doi: 10.1080/10618600.2000.10474879.

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, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé 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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Donald B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979. doi: 10.1109/TAC.1979.1102177.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016 Workshops*, volume 9914 of *Lecture Notes in Computer Science*, pp. 17–35. Springer, 2016. doi: 10.1007/978-3-319-48881-3_2.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341.

Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H. J. Kelly, and Andrew J. Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1352–1359, 2013. doi: 10.1109/CVPR.2013.178.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Ba-Ngu Vo, Ba-Tuong Vo, and Dinh Phung. Labeled random finite sets and the Bayes multi-target tracking filter. *IEEE Transactions on Signal Processing*, 62(24):6554–6567, 2014. doi: 10.1109/TSP.2014.2364014.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 7657–7666. IEEE, 2019. doi: 10.1109/ICCV.2019.00775.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation. In *Robotics: Science and Systems XX*, Delft, The Netherlands, 2024. doi: 10.15607/RSS.2024.XX.077.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151.

# SUPPLEMENTARY APPENDIX

**Same Object, or Its Twin? Diagnosing Instance Association Across Scene Revisits**

Companion to the main paper. Section, figure and table numbers follow the combined manuscript. References are provided in the main paper.

### APPENDIX ROADMAP

Appendix A proves the identifiability and decision statements and connects their assumptions to implementation checks. Appendix B specifies generation, perception, association, calibration, and evaluation protocols. Appendix C organizes additional evidence by the main-text claims, including uncertainty bounds, intervention controls, real-data cases, and independent representations.

## A PROOFS, SCOPE, AND EXECUTABLE CHECKS

The formal statements concern the sensing model in Section 3. The empirical checks use physical-identity events from the frozen run `sam3_dinoL`. We distinguish model implications, geometric premises verified on recorded scenes, and measured policy behavior.

### A.1 APPEARANCE SYMMETRY

**Lemma 1** (Appearance tie)**.** *Suppose appearance observations depend on the generating object only through its asset and relative viewpoint, the candidate set is restricted to $T$, and the joint appearance observation has the same conditional law under all $K$ possible predecessor correspondences. Assume a label-exchangeable prior over which twin occupies the queried slot. Then (i) any appearance-measurable rule has accuracy at most $1/K$, with equality when it always selects within $T$, and (ii) uniform random selection within $T$ names any fixed twin, including the swapped partner, with probability $1/K$. Averaging over queries gives the corresponding reference $\mathbb{E}[1/K]$.*

*Proof of Lemma 1.* Condition on an appearance observation. By assumption, its likelihood is equal under all $K$ possible predecessor correspondences. A label-exchangeable prior then gives a uniform posterior on $T$. Any selected candidate in $T$ has posterior probability $1/K$ of being the correct identity. Selecting an outside candidate or NEW cannot increase correctness on these existing-object queries. This proves the accuracy bound.

The probability of selecting a specified partner concerns the output distribution. Uniform sampling within $T$ assigns probability $1/K$ to each label. A sampler that sometimes returns an outside output and otherwise selects uniformly within $T$, conditional on the observation, names each twin with probability $\mathbb{P}(\delta \in T)/K$. Averaging the forced uniform-selection probabilities over queries gives $\mathbb{E}[1/K]$. □

The label-exchangeable prior is an assumption of this reference. Knowledge of which objects the generator swaps can restrict the relevant identities to a smaller set. The oracle-restricted row computes $1/K$ analytically to supply a reference distribution. Candidate-record and candidate-identity references in Appendix C.1 account for the actual retrieved set and its fragments. Perceived-query and physical-slot averages differ because the recorded perception can produce several queries for one slot.

### A.2 TWO OBSERVATIONALLY EQUIVALENT WORLDS

*Proof of Proposition 1.* The two worlds have the same visit-1 configuration. On visit 2, the swap exchanges the complete poses of objects with the same asset. Consequently the multiset of asset–pose pairs is identical in both worlds. Applying the same sensing law to this multiset gives

$$\mathbb{P}(O \mid W_1) = \mathbb{P}(O \mid W_0).$$

Bayes' rule therefore leaves the posterior odds equal to the prior odds.

Couple the worlds to produce the same observation. At a fixed exchanged slot, the true identity is $a$ in one world and $b$ in the other. Let $q_a(O)$ and $q_b(O)$ be the probabilities that the rule selects these physical identities, summing over all records of each identity. These events are mutually exclusive, so $q_a + q_b \leq 1$. The sum of error probabilities is

$$(1 - \mathbb{E}q_a) + (1 - \mathbb{E}q_b) \geq 1.$$

□

A deterministic renderer satisfying this sensing model would produce identical observations. A stochastic renderer has equal conditional laws under the same model. Equality of two independently sampled images is an additional property to measure. Renderer state, instance-dependent effects, or a visible persistent marker can change the observation law and require their own audit. The cached counterfactual control fixes the inputs exactly and changes the world labels to test the decision tradeoff conditional on those inputs.

### A.3 Quasi-static inversion

*Proof of Proposition 2.* Assumption 1 makes the partner's visit-1 record the unique nearest same-category candidate after allowing for the registration error. Every pose-monotone rule selects that record. Its physical identity differs from the query's, giving conditional correctness zero and partner assignment one on $W_1$.

For the Bayesian statement, observational equivalence preserves prior probabilities $(\pi_0, \pi_1)$. With $\pi_0 > \pi_1$ and support on the two existing-object correspondences, the MAP rule under 0–1 loss selects the $W_0$ correspondence. On the swapped pair, this correspondence assigns both observations to their partners. □

The clean recorded panel satisfies the geometric premise on all 86 physical slots, with a minimum measured relative margin of 0.129 m. That check uses ground-truth positions and the full same-category geometric candidate set. Perceived centroids, category errors and missing records change the inputs of the empirical policies, so their measured rates are reported separately.

### A.4 Threshold and cost decisions

**Corollary 1** (Preserving an unresolved identity)**.** *(a) Any rule restricted to* MERGE *and* NEW *has worst-case within-pair accuracy at most* $1/2$ *over* $\{W_0, W_1\}$ *and cannot determine which world holds. (b) At an exact top tie among* $m \geq 2$ *candidates with* $p_{\text{new}} \leq 1/2$*, a merge threshold above* $1/2$ *and a strict NEW threshold of* $1/2$ *leave the identity unresolved. (c) An absolute-position cue can break the tie in favor of the swapped partner and trigger a partner merge once its likelihood-ratio gap crosses the decision threshold.*

*Proof of Corollary 1.* (a) Proposition 1 gives identical observation laws and an error sum of at least one. Thus at least one world has error at least $1/2$. A NEW output is also wrong for these existing objects and cannot improve the bound.

(b) Write the candidate posterior as

$$p_j = (1 - p_{\text{new}}) \frac{e^{\ell_j}}{\sum_k e^{\ell_k}}.$$

At an exact $m$-way top tie, $\max_j p_j \leq (1 - p_{\text{new}})/m \leq 1/2$ for $m \geq 2$. A merge threshold above $1/2$ cannot be exceeded. With $p_{\text{new}} \leq 1/2$ and a strict NEW threshold of $1/2$, the NEW branch also remains inactive, yielding ABSTAIN.

For an approximate top tie of spread at most $\delta$, with every outside candidate scoring no higher than its maximum,

$$\max_j p_j \leq \frac{1 - p_{\text{new}}}{1 + (m-1)e^{-\delta}}.$$

This bound specifies a sufficient condition for rejection once compared with the merge threshold. The condition on $p_{\text{new}}$ remains a separate premise.

(c) Adding a position LLR changes the relative logits. When it makes the partner the top candidate and raises its posterior above the merge threshold, the decision becomes a partner merge. □

For costs $(C_{\mathrm{om}}, C_{\mathrm{os}}, c_{\mathrm{abs}})$, the posterior expected costs are

$$C_{\mathrm{om}}(1 - \max_j p_j), \qquad C_{\mathrm{os}}(1 - p_{\mathrm{new}}), \qquad c_{\mathrm{abs}}.$$

The rejection action is a minimizer when both committed-action costs are at least $c_{\mathrm{abs}}$. At the main costs $(5, 1, 0.25)$, this requires $\max_j p_j \leq 0.95$ and $p_{\mathrm{new}} \leq 0.75$, with equality handled by the implementation's stated tie rule. These are posterior decision conditions. Their empirical utility depends on calibration and the quality of the selected queries.

**Scope.** Appearance symmetry assumes equal joint observation laws. Recorded camera paths and embeddings can retain view-conditioned differences, which the empirical cue controls measure. Geometric inversion additionally requires the relative margin and a pose-monotone rule. Fusion, gating and assignment constraints can change the decision. Appendix B.8 specifies event references and uncertainty, and Appendix B.11 records implementation checks.

# B Experimental Protocols and Implementation Details

This appendix records the generator, perception, policy, decision-rule, and evaluation settings. Generator configurations retain scene and placement seeds. Intermediate artifacts include episode data, perception caches, per-policy decisions and metric tables. Association and evaluation read frozen inputs.

## B.1 Episode generator

**Scenes and assets.** Episodes use six locally curated, commercially sourced Kujiale interiors converted to USD. Each scene contains a few hundred per-object meshes whose filename prefix gives the category, room polygons with room types, and an HDR environment. A probe pass inventories the scene's categories, prim paths, world-space bounding boxes and room assignments. The grid expander chooses target instances from this inventory, which covers objects inside room polygons.

**Capture.** Generation runs in Isaac Sim 4.5.0 with Isaac Lab 2.1.1. Each visit captures 60 poses at $960 \times 540$ resolution and camera height $\approx 1.45$ m: roughly 75% on rings of radius 1.4/2.2/3.0 m centered on the target's horizontal position, with yaw facing the target and pitch set to look at the target's height, and roughly 25% on a room orbit and lateral sweep for context. Visit 2 reuses the visit-1 poses offset 0.15 m toward the target with a $+5^\circ$ yaw change, so viewpoints differ across visits by construction. Each frame stores RGB, float32 depth, an instance-id segmentation with a pixel-to-prim mapping, and the camera pose in OpenCV convention. An object counts as visible in a frame when it covers at least 200 segmentation pixels.

**Twin placement.** Twins are placed by rejection sampling inside the target's room polygon with a deterministic per-seed RNG. A candidate pose must lie at least 0.6 m from every registered object and every twin already placed, measured between object centers. The boundary margin can step down from 0.6 m to 0.45 m and then 0.3 m when a small room cannot host the requested placements. The first twin slot inherits the native target chair's pose, placing its clearances outside the sampler guarantee. The recorded panel measures a minimum twin-twin visit-1 distance of 0.602 m and a minimum twin-to-native same-category distance of 0.649 m. These checks evaluate placement endpoints. Mesh intersections and transport dynamics remain untested.

**Edit operations.** Edits apply sequentially: move (translate the first twin by $d$ along a seeded heading), swap (exchange the current world transforms of the last two twins), remove (deactivate the prim of the movable non-target object nearest the room centroid), insert (spawn one extra box object at a sampled clear pose), lighting (apply a fixed intensity scale to every light prim in the stage), and texture (multiply visit-2 RGB by a fixed tint inside one twin's instance mask). Sampled move targets must satisfy the same 0.6 m object clearance. Each historical swap-panel episode composes move(0.5 m) before the swap. Because move displaces the first twin while swap exchanges the last

two, the moved twin belongs to the swap pair when $K{=}2$ and stays outside it when $K \geq 3$, matching the composition in Section 4.

**Historical tint correction.** The main-grid texture arm used a USD `displayColor` tag ignored by material-bound meshes and was a no-op. A separate corrected control applies a visit-2 RGB multiplier $(0.85, 0.80, 0.72)$ inside one twin's mask in 42 paired episodes, then reruns SAM3–DINOv2. This nonpersistent intervention is distinct from the two-visit markers below. The archived outcomes retain the fixed 84 target slots, including a target that loses its eligible chair query after tinting.

**Grid composition.** The historical main grid crosses the full $(K, d)$ product $\{1, 2, 3, 6\} \times \{0, 0.5, 1.0, 2.5\}$ m with all other factors off, plus four one-factor-at-a-time arms anchored at the $K{=}3$, $d{=}0.5$ m reference cell: swap, removal-plus-insertion, global lighting, and the audited texture-tag control. This gives 20 factor cells per scene (16 core cells plus 4 arms). Over six scenes, four cells of one small dining-room scene failed the configured 0.6 m center-clearance placement check, leaving 116 episodes plus two smoke episodes (one $K{=}3$, $d{=}0.5$ swap cell and one $K{=}1$, $d{=}0$ calibration cell). The swap extension arm adds 36 episodes ($K \in \{2, 3, 6\}$, two seeds, six scenes) for the swap panel, and 42 seed-matched tint counterparts form the paired appearance panel. Episode-prefix guards exclude both extension sets from grid tables. The data schema also supports a load-time visibility factor, and all reported experiments use full visibility.

### B.2 Persistent-cue and viewpoint protocol

**Frozen factors.** The design specifies six scenes and placement seeds 9101, 9102 and 9103, giving 18 scene-placement groups. Each group has no-swap and swap worlds under correlated and independently sampled camera paths, for 72 scientific episode IDs. Every world uses $K = 2$ and zero additional displacement. The color-strength grid is $\lambda \in \{0, 0.05, 0.15, 0.40, 0.85\}$, with 0.85 designated as the strong control. Context and neutral-background embeddings are paired at each strength. A separate seed 9099 supports engineering checks and is excluded from this population.

**Persistent screen-space color.** A group-specific assignment gives the two physical targets different RGB multipliers, $(1, 1-\lambda, 1-\lambda)$ and $(1-\lambda, 1, 1)$. The assignment follows physical identity across both visits and worlds (Figure 5). It is randomized between groups and applied inside the rendered instance masks. Zero strength preserves every source pixel. These are artificial persistent appearance cues, applied after rendering. The audit records visible pixels, actual channel changes and any changes outside the target masks. Marker codes and physical-identity labels remain in generation and evaluation. Association receives sanitized entity records and observed features.

**Capture pairing.** Each scene-placement group begins with one excluded no-swap warmup that captures both visits. Its verified first visit is shared by the four scientific arms, whose second visits are captured independently. Four preassigned capture-order templates balance which world comes first across the 18 groups. All visits contain 60 frames. The independent camera regime selects 60 poses from a fixed 240-pose, slot-anchored pool, and paired worlds use the same selected poses. The paths depend on scene geometry and slots, so exchanging physical identities leaves the camera sequence unchanged. All 36 formal world/camera pairs pass camera, depth and target-mask equality checks. Among 2160 paired visit-2 frames, 0 are RGB-identical. Mean absolute RGB difference is 4.599 on the 0–255 scale. These captures provide paired geometry with measured image variation. The equality of observation laws in Proposition 1 is a model condition.

**Perception routes and background control.** The GT-instance route supplies masks, categories and centroids, using up to five evenly spaced visible frames per entity with at least 400 foreground pixels. It isolates association with oracle instance geometry. Full SAM3 reruns segmentation and aggregation at every strength, retaining up to eight detector-score-selected prototypes per entity. Route differences therefore include geometry, membership and view selection. Within each route and strength, the background intervention fixes crops, masks, candidates and membership. It replaces pixels outside the foreground with RGB $(127, 127, 127)$ before recomputing DINOv2 features. Full-SAM3 membership can vary across strengths. Verified first-visit features may be shared within a group. Execution records distinguish inference from reuse.

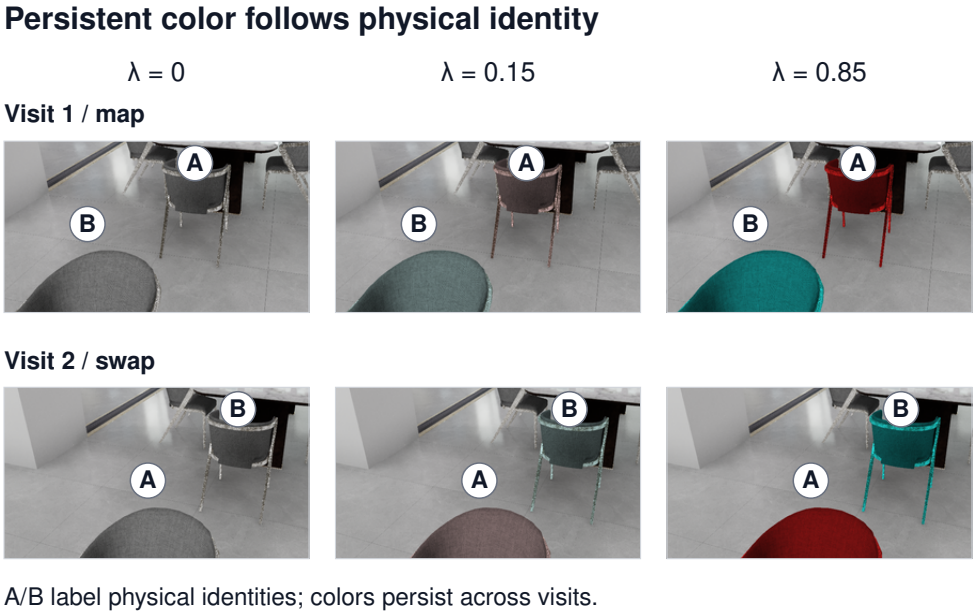


Figure 5: Persistent screen-space color in the first predeclared scene-placement group. A/B follow the physical chairs through the swap. Three strengths share one crop and a capture index selected by maximum minimum target-mask area. The foreground chair is truncated by the original image boundary.

**Oracle view control.** Within the GT-instance route, an additional diagnostic selects target crops from the existing rendered frames using object-relative camera bearings computed from GT poses. A feature-independent greedy match selects two to five four-way view tuples spanning both targets and visits, with pairwise angular spread at most $30^\circ$. The comparison includes the original prototypes, matched views and an equal-count control with evenly spaced frame-order selection. Non-target features, candidate geometry and entity membership stay fixed. Groups without sufficient matching views retain an explicit unavailable status. This controls viewing direction. Camera distance and roll are not jointly matched.

**Readout and analysis.** A fixed color readout uses GT masks and the two target candidates. It averages $(R - G)/(R + G + 1)$ over foreground pixels, takes the median over up to five selected views, and chooses the closest visit-1 scalar descriptor. Numerical ties within $10^{-12}$ remain unresolved. Readout-correct control abstention conditions actual policy abstention on a correct oracle-mask readout, with non-detection separate. Primary outcomes weight physical slots equally. Fragments share their slot's weight and missing detections remain in the population. Paired marker, background, camera and world effects use 5,000 bootstrap draws with seed 606. Slot- and episode-weighted analyses resample scene-placement groups, while scene-weighted sensitivity resamples the six scenes. Margins report candidate support and denominators. The nine predeclared primary policies include CUE and CUE+POS. EVA-style and PDA-DP remain exploratory. Calibration, thresholds and contrasts are frozen before formal extraction.

**Selective and view populations.** T5 matched-risk comparisons rank the same observed target records within each condition. Membership may vary across strengths and perception routes. Oracle-view effects use eligible paired support, while the main outcome curves retain all physical slots. These three populations have separate denominators in the accompanying tables.

### B.3 SAME-INPUT REPRESENTATION AND READOUT CONTROLS

**Common inputs.** The additional control uses all 72 formal GT-instance episodes at $\lambda \in \{0, 0.85\}$ with context intact. It replays the same treated RGB, 10% crop expansion, masks, boxes, selected frames, prototype indices and entity membership used for DINO. Each reconstructed crop stream must match the original input hash. Both visits retain all visible map entities. Marker codes enter the pixel generator and later evaluation. The association interface receives images, features and sanitized entity records.

**Representations and ranking.** Mean RGB averages the foreground pixels in each crop and normalizes the resulting three-vector. A secondary descriptor first divides each pixel by $R+G+B+1$, then averages and normalizes. MASA-R50 reads the same pre-resize RGB crop and the original box translated into that crop. Its native resize, normalization, backbone, adapter and ROI head produce a 256-dimensional feature. DINO uses its original 518-square resize and mask-pooled patch features. These encoders share source observations while retaining their native encoding procedures. MASA's ROI encoding is evaluated as a component, separately from its native tracking system.

Every representation uses the maximum cosine over the query/candidate prototype pairs and the same all-map candidate set. A same-category restriction is reported separately. The diagnostic ranker has no learned threshold. It selects a unique maximum, returns ABSTAIN for exact ties or unavailable prototype comparisons, and returns NEW for an empty candidate set. This common rule isolates feature-dependent retrieval behavior. Original NN and CUE retain their own thresholds and calibration in the separate policy results.

**Development-fitted feature probe.** A binary readout predicts the generator's marker code from the normalized mean of unit prototypes for each observed target entity. At each strength, ridge regression with penalty 1 and an unpenalized intercept fits only the four original development arms from one placement in `kujiale_0003`, seed 9099. Both visits contribute 16 entity observations of two physical targets. Centering uses these training vectors, and the prediction threshold is zero. Evaluation retains the two visit-2 target slots in every formal episode. Unavailable features remain incorrect in the unconditional readout rate. An entity containing a zero prototype is unavailable under this probe's normalization rule. This code diagnostic measures transfer from one development placement. Its output is separate from full-candidate identity association.

**Stage readout and uncertainty.** The original CUE log-likelihood ratios, NEW prior and posterior are reconstructed from their saved DINO inputs. Each stage reports the true-minus-partner maximum over identity records, combining cue values per candidate before taking that maximum. A shared candidate prior preserves the ordering among remembered candidates and affects competition with NEW. Reported flags distinguish exact score collapse, sign reversal and clipping. Fixed-slot outcome rates retain non-detections and unknown candidates. Paired contrasts use 5,000 scene-placement bootstrap draws with seed 606. Costs of observed decisions use $(5, 1, 0.25)$ for wrong MERGE, wrong NEW and ABSTAIN. A non-detected slot has no imputed action or cost.

Table 2: Units in the same-input mechanism experiment. Each scene-placement group contains both worlds and both camera regimes. Strength changes reuse those captures.

| Unit | Development | Formal evaluation |
|---|---|---|
| Scenes | 1 | 6 |
| Scene-placement groups | 1 | 18 |
| World/camera episode arms | 4 | 72 |
| Visit layouts | 8 | 144 |
| Crop observations across two strengths | 2,640 | 63,268 |
| Visit-2 physical slots per strength | 8 | 144 |
| Slots in each world/camera stratum | 2 | 36 |

### B.4 AGGREGATION AND FIXED-READOUT INTERVENTIONS

**Score support and edge cases.** The cached intervention applies the A, B and C scores in Section 5, keeping every selected view, entity membership and visit-1 candidate from the same-input experiment. Each rule independently ranks the complete map, retaining candidate fragments as separate records. Exact top ties and wholly unavailable comparisons return ABSTAIN. An empty map returns NEW. Zero view vectors are excluded consistently across the new rules, and a mean norm at most $10^{-12}$ makes C and its readout unavailable. The consumed feature population has no zero view vectors. View counts, mean norms and the maximizing view pair are retained for interpretation.

**One readout across conditions.** For each of MASA and DINO, the ridge model is fitted once on strong-color development observations with penalty 1. The model, centering and intercept are shared across both strengths, worlds and camera regimes. Its continuous output uses the normalized entity mean. The penalty in Section 5 scales output differences by the population standard deviation of development training outputs. Distance-only matching retains the negative squared term. Both visits and all ordinary map candidates receive the same readout. Only marked objects enter binary-code accuracy. An unavailable readout or a nonfinite or near-zero scale makes positive-$\beta$ comparisons unavailable. At $\beta = 0$, the base score is recovered exactly.

**Development selection and controls.** Each representation, aggregation and readout selects one $\beta$ from $\{0, 0.01, 0.03, 0.1, 0.3, 1, 3\}$. Within each development condition, target-slot accuracy and observed labelable non-target accuracy receive equal weight. The objective then averages the eight world, camera and strength conditions. Exact objective ties select the smaller $\beta$. This uses the same single development placement as the original probe. For each representation, 32 Gaussian directions with seed 17092026 are normalized to the fitted weight norm. Each retains the same centering and intercept and receives its own development scale and the identical selection budget. All directions and grid outcomes are reported. The complete original strength-specific probes remain as a separate historical comparison.

**Populations and uncertainty.** Target identification retains every physical slot, including missing true map candidates. Fragments share their slot's weight. Non-target outcomes use the observed labelable query population, with unknown identities separate. Paired changes in margins and ranks require finite scores in both compared rules and report their support separately. Confidence intervals use 5,000 shared bootstrap draws of the 18 scene-placement groups, seed 1709, preserving every condition within each draw. These intervals condition on the fitted readout and the six existing scenes. A prospective scene-independent evaluation and variation across development placements remain separate questions.

### B.5 SHARED PERCEPTION PASS

**Segmentation and embeddings.** SAM3 (Carion et al., 2026) runs per frame with text-concept prompts drawn from the episode vocabulary (the scene inventory categories). We keep a mask when the predictor score is at least 0.35 and the mask area at least 400 px, and store it in compressed RLE form. DINOv2-L with registers (Oquab et al., 2024; Darcet et al., 2024) embeds each detection by pooling patch tokens under the mask on its bounding-box crop, expanded by 10% for context and resized to $518 \times 518$ so the patch grid is exactly $37 \times 37$, followed by $\ell_2$ normalization. 3D centroids take the median valid depth inside the mask and backproject through the intrinsics and the per-frame camera pose.

**Entity aggregation (fragment merging).** Within each visit, detections are processed in detection-id order, and each joins the best qualifying cluster: qualifying means the same category, centroid distance below 0.5 m, and maximum cosine above 0.5 against a cluster member, and best means the highest such cosine. A detection that qualifies nowhere starts a cluster. Two deterministic post-passes repair the fragments the frame-order pass leaves behind. The first repeatedly merges the closest same-category cluster pair whose centroids lie within 0.45 m and whose best cross-cluster cosine exceeds 0.30, a distance chosen below the 0.6 m spacing that separates two twins. The loose cosine gate accommodates opposite-side views of one chair. The second absorbs cross-category crumbs, since the segmenter labels a chair's cushion or side as a different concept: cluster $B$ folds

into cluster $A$ when their centroids lie within 0.25 m, their best cross-cluster cosine exceeds 0.60, and $B$ is small enough to be a fragment ($n_B \leq \max(3,\, 0.15\, n_A)$). An entity's centroid is the median of its member centroids, its prototype set keeps up to eight member embeddings chosen by descending detector score with at most one per frame, and clusters with fewer than two observations are then dropped as flicker. This aggregation is shared verbatim by every policy.

**Evaluation-only bookkeeping.** Each detection records the ground-truth id whose instance mask it overlaps by majority (at least 0.5 of mask pixels). Between 0.2 and 0.5, the highest-overlap id is recorded with its overlap. Below 0.2, no id is recorded. Detector recall per episode is the fraction of ground-truth entities of interest categories with at least one matching detection, and cells with recall below 0.85 are flagged. Ground-truth ids, purity and visibility support calibration and evaluation. A fixed-input audit clears or randomizes these fields before association and checks the resulting decisions.

### B.6 BASELINE CONSTANTS

All policies operate on the same interface: given the visit-1 map entities and the visit-2 perceived entities, produce for every visit-2 entity a visit-1 entity, NEW, or ABSTAIN, together with a score matrix and an optional per-decision confidence. The thresholds below are the values used in all reported runs.

- CG (ConceptGraphs-style (Gu et al., 2024)): score $= 0.5 \cdot \mathrm{sem} + 0.5 \cdot \mathrm{geo}$, where sem is the maximum cosine over prototype pairs and $\mathrm{geo} = \exp(-\|\Delta\mathrm{centroid}\|/1.0\,\mathrm{m})$. Greedy row-wise argmax selects the candidate, with NEW if the best score falls below 0.6.
- EVA (Embodied VideoAgent-style (Fan et al., 2025)): entities whose category is in a movable set (chair, box, plant, and similar) are dynamic and the rest are static. Both lists require the same category and an appearance maximum cosine of at least 0.55, and the static list adds a 0.75 m position gate. Each query is tried against the static list first and becomes NEW if both fail.
- NN: maximum cosine over all prototypes with greedy argmax, returning NEW below 0.62.
- Cosine–Hungarian (cache identifier `masa`): appearance-only Hungarian assignment on negative DINOv2 cosine with acceptance threshold 0.6, enforcing one-to-one matches. This local policy is a shared-feature association proxy.
- Single-exemplar (cache identifier `sam3p`): each visit-1 entity contributes its single highest-scoring DINOv2 exemplar embedding, a visit-2 entity takes the best-matching exemplar by cosine, and the match is accepted above a strict 0.7. Below that the entity becomes NEW. The policy measures exemplar-based association on the common entity cache.
- TIER: the long-gap tier of a hand-tuned legacy robot-memory associator. Its first two tiers cover in-shot track continuity and fused matching inside an 8 s active window, both inapplicable across a revisit gap, so the ported rule is the third: appearance-only prototype matching with at most five exemplars per entity and a top-10 candidate cut, merging above 0.80 cosine, with a "possibly same" band from 0.55 to 0.80 that the forced-commit output contract turns into NEW.
- PDA-DP: a Chinese-restaurant prior over visit-1 entities (entity $j$ weighted by its observation count, NEW by concentration $\alpha$=2.0) times a Gaussian position likelihood ($\sigma_{\mathrm{pos}}$=0.75 m) and an appearance factor $\exp(4(\cos -1))$, committed by argmax (Mu et al., 2016). This baseline combines a Dirichlet-process prior with a committed argmax output. It uses a fixed concentration, while CUE scales $\alpha_c$ with neighborhood density (Appendix B.7).
- PDA-MM: a local maximum-score mixture proxy. Each hypothesis is the same Gaussian position likelihood times a categorical agreement factor (0.9 when the categories match, 0.1 otherwise), and a constant null hypothesis at log-density $-5$ supplies NEW.
- PDA-CONST: constellation joint compatibility over same-visit neighbor geometry. Same-category pairs above a loose 0.4 appearance gate are considered in descending appearance order, and a pair is accepted only if its same-visit distances agree within 0.35 m with every

already-accepted pair. Observations that fit no jointly compatible pair become NEW (Frey et al., 2019).

The CG-style, NN, EVA-style, single-exemplar, PDA-DP and PDA-MM rules allow many-to-one assignment. Cosine–Hungarian and PDA-const enforce one-to-one assignment. These choices affect both candidate competition and the errors the output can express.

The TIER hedge wrapper restores an unresolved output for its original possibly-same band. A separate native-component evaluation executes the original long-gap association code from a fixed source commit. It preserves NEW plus a possibly-same edge as two linked output states. Both versions retain the shared DINO entity cache, which isolates the association component. Appendix C.6 reports their distinct decision semantics.

**Independent MASA-R50 representation.** We use the official ResNet-50 checkpoint, preprocessing, feature adapter and 256-dimensional ROI tracking head (Li et al., 2024) on the cached detector boxes. Entity membership and prototype selection remain fixed. The entity-batch component represents each entity by its highest-detector-score exemplar and executes the source bidirectional-softmax/cosine scoring and detector-score-priority greedy assignment. It uses match, object and new-track thresholds of 0.5, 0.01 and 0.1. Evaluation freezes visit-1 memory and matches the whole revisit entity batch. The source video tracker instead forms each competing query batch from one frame. Temporal retention, image-distance gating, distractor NMS and memory updates are outside this component interface. R50 cosine–Hungarian retains all cached prototypes and the original cosine–Hungarian rule, changing its feature family. Both variants are reported separately.

For selective evaluation, the entity-batch component ranks merges by their accepted matching score. Its NEW ranking score is 0.5 minus the best available matching score. Untracked queries map to ABSTAIN. These local ranking quantities are passed to the same held-out calibration protocol as the other policies. Source commit, checkpoint SHA256, runtime versions, feature-frame identities and output hashes are retained in the execution manifests.

### B.7 CUE MODEL AND DECISION RULES

The display names CUE and CUE+POS denote the cue model and its position-augmented variant. Reproduction files retain the historical policy identifiers `ours` and `ours+pos`. The display-name change leaves all policies, parameters, and results unchanged.

**Cues.** CUE scores each candidate pair with appearance (maximum cosine between prototype sets), category consistency, and a constellation descriptor. The descriptor collects sorted log distance-ratios to the $k{=}4$ nearest same-visit entities and their categories, then compares descriptors by $L_1$ distance with a category-overlap term. Uniform scaling and rigid transformations within each visit preserve this descriptor. CUE+POS adds absolute-position proximity $\exp(-d/2.0)$ for cross-visit centroid distance $d$. Both complete cue-model variants also use cross-visit coordinates in the NEW prior.

**Calibration.** Each cue is mapped to $P(\text{same} \mid \text{cue})$ by isotonic regression (Zadrozny & Elkan, 2002) fitted on the designated easy-cell episodes (the $K{=}1$, $d{=}0$ cells and the accepted real pairs of Appendix B.9). The persisted curves remove the class base rate to obtain a log-likelihood ratio. Clipping each LLR to $\pm 6.0$ bounds its contribution. Relative cue influence still depends on the other scores.

**New-object prior.** A two-level prior, inspired by new-cluster allocation in Dirichlet-process models (Ferguson, 1973; Neal, 2000), assigns unit weight to the existing-object hypothesis and weight $\alpha_c = \alpha_0(1 + n_c)$ to NEW. Here $\alpha_0 = 0.05$ and $n_c$ counts same-category map entities within 3 m of the query. For $M > 0$ candidates, $P(\text{NEW}) = \alpha_c/(1 + \alpha_c)$ and $P(\text{MERGE}(j)) = 1/[M(1 + \alpha_c)]$. This normalization makes NEW mass increase with local density. When the candidate set is empty, the rule assigns all mass to NEW.

**Decision rules.** Per-cue LLRs are summed with the prior into a posterior over remembered candidates and NEW. The threshold rule commits MERGE if the maximum posterior exceeds

$\tau_{\text{merge}} = 0.7$, outputs NEW if $p_{\text{new}}$ is largest and exceeds $\tau_{\text{new}} = 0.5$, and otherwise outputs ABSTAIN. The expected-cost variant minimizes posterior expected costs with main values $(C_{\text{overmerge}}, C_{\text{oversplit}}, c_{\text{abstain}}) = (5, 1, 0.25)$. Its rejection region follows the arithmetic in Appendix A. The default per-decision score is the maximum posterior. Margin and normalized-entropy variants retain their own reachable commitment ceilings.

**Cue fusion.** The cue model adds calibrated cue LLRs. Appearance and neighborhood geometry can carry correlated information, so individual cue calibration does not establish calibration of their sum. The matched-coverage and held-out cost evaluations measure the resulting decision quality.

**Parameter provenance.** This revision freezes the inherited hyperparameters. Historical development records describe changing $\alpha_0$ from 1.0 to 0.05 after inspecting early full-grid outputs. A manifest identifying the original parameter-selection split is unavailable. The revised simulation evaluation is calibration-disjoint within the legacy development scenes.

### B.8 METRICS PROTOCOL

**Identity and correspondence events.** In simulation, generation labels map each perceived record to a physical identity, and all records of the query's identity count as correct candidates. The other events are the specified partner, another twin, an outside-group merge, wrong NEW, and ABSTAIN. A truly new query permits correct NEW. In 3RScan, the same action partition is evaluated against the designated released correspondence. Additional annotation-supported alternatives are reported separately in Appendix B.9. Unknown candidate labels remain explicit, and query labelability is determined before reading policy actions.

**Chance baselines.** Uniform-within-group selection samples the $K$ annotated IDs. Simulation correctness and selection of a specified partner have probability $1/K$. In 3RScan, designated agreement has probability $1/K$ and selection of any other container member has probability $(K-1)/K$. The common candidate-record baseline samples same-category map records. Candidate-ID baselines collapse known fragments before sampling. Each policy's gating and assignment constraints define its own scoring support. Unknown candidate labels give intervals for compatible event probabilities.

**Selective decisions.** For each policy, we rank its eligible native commitments on the same labelable evaluation universe using its fixed score. Common integer budgets define matched coverage. Native abstentions remain unavailable at budgets beyond the policy's ceiling. A no-reject scorer is a separate configuration. Ties use a deterministic query-key rule independent of correctness. Simulation reports wrong merges and wrong NEW decisions, while 3RScan reports disagreement with the designated correspondence. Merges to unknown candidates remain separate, with bounds on committed-decision risk and total-query cost.

**Calibration and cost.** Isotonic fits, threshold selection and score semantics use the designated calibration split. All-abstain is a constant-cost comparator. Thresholds minimize the same conservative calibration objective across policies. Evaluation outcomes do not enter threshold fitting in this revision. Inherited hyperparameters retain the development provenance above. Shared partial-AURC intervals end at a common reachable budget. Appendix C.6 gives the denominators and splits.

**Uncertainty.** Paired bootstrap draws keep all policies and their query-specific references together, recomputing both events and references in each draw. Simulation resamples episodes and reports descriptive scene and leave-one-scene-out analyses. 3RScan clusters related rescans by reference scene. Wider scene-population inference requires more than the six simulated scenes. The displayed intervals measure effects relative to the matching event reference. Unknown-identity bounds are propagated inside each resample. The scripts persist seeds and resample settings.

**Supplementary diagnostics.** Historical map-level over-merge/over-split, IDF1x, AssA and committed-merge ECE remain in the baseline archive. IDF1x and AssA adapt standard identity and association metrics (Ristani et al., 2016; Luiten et al., 2021). ECE uses 15 equal-width bins (Guo et al., 2017). Historical grids include calibration episodes and method-specific AURC ranges. Revised tables use physical events, held-out decisions and common comparison ranges.

### B.9 3RSCAN PROTOCOL

**What the labels establish.** The dataset provides both designated instance mappings and transformations between ambiguous instances. Its description explicitly allows multiple plausible mappings when the individual correspondence is unknown (Wald et al., 2019, Section 3.1 and Figure 3). We join each saved action to the released metadata and distinguish designated agreement, a directly annotated alternative, a path-only connection through other annotated pairs, membership only in the same outer container, and a merge with no such support. NEW and ABSTAIN remain separate. Direct-pair support describes the annotation relation. The official task additionally evaluates pose errors and symmetries, which our ID-only actions do not predict. A path through several pairs is reported as a sensitivity condition. None of these categories independently verifies the physical individual.

**Episode construction.** Each of the 28 ambiguous reference/rescan pairs (Wald et al., 2019) becomes one episode. Visit 1 is the reference scan and visit 2 the rescan, whose geometry is transformed into the reference frame using the released alignment. Annotated object IDs and oriented-bounding-box centroids define entities. The designated correspondence uses an explicit rigid-instance mapping when available. Otherwise it retains the same ID when that reference object exists and is absent from the removed list. Only objects with neither counterpart receive NEW. Metadata supplies $K$, the largest ambiguity-container size containing a moved instance (else 1), displacement from the rigid transform, and the group category.

**Perception route.** The reported run synthesizes the perception cache from ground-truth instances: DINOv2 embeddings are computed on instance crops from sampled color frames, with instance masks obtained by projecting the annotated mesh vertices, and centroids come from the annotated instances. This isolates association behavior from real-detector noise while keeping the cache format identical to simulation.

**Corpus and calibration selection.** The evaluation table covers the 28 ambiguous pairs with $n$=715 decisions per policy and 7 to 66 eligible decisions per pair. Real-data calibration reserves pairs with no ambiguity annotations and no recorded rigid motion. Six candidates were downloaded. An automatic factor filter accepted four and rejected two whose rescans contain inserted objects (3 and 1 insertions, respectively). Perception recall on the accepted pairs is 0.40/1.0/1.0/1.0. The persisted selection list excludes all six reserve pairs from evaluation. Dataset-provided cross-scan alignment supplies the coordinates for the absolute-position cue of CUE+POS.

### B.10 POSE-NOISE PROTOCOL

The historical noise sweep perturbs visit-2 detection geometry once per episode, with translation scales $\sigma_t \in \{0, 0.05, 0.15, 0.50\}$ m paired with rotation scales $\sigma_R \in \{0, 2, 5, 10\}^\circ$. It then reruns aggregation and fits the associated calibration curves. All 11 policies completed the four levels over 118 episodes. Changes in that sweep combine coordinate perturbation, entity aggregation and calibration. The fixed-input audit in Appendix C.8 holds entity membership, embeddings and calibration constant.

### B.11 COMPUTE AND REPRODUCIBILITY

Experiments use a shared node with eight NVIDIA RTX A6000 GPUs. Each rendering or offline perception process uses one GPU, while association and statistics run on the host from cached features. Episode generation uses an Isaac Sim 4.5.0 / Isaac Lab 2.1.1 container. Run tags identify the historical caches: `sam3_dinoL` (main grid and panel), `sam3_dinoL_n05/n15/n50` (pose-noise levels), and `3rscan_gt` (real anchor). Revision runs record timed development checks, prospective compute budgets and actual execution times separately.

The persistent-cue full-SAM3 feature calls took 34.070 wall-clock hours on one GPU, including reuse checks. The same-input MASA control encoded 23,652 unique inputs on two GPUs in 629 seconds per shard, restoring 65,908 prototype rows across development and test. This encoding time excludes CPU preparation and the development smoke. Shared cosine comparisons and feature probes run on CPU. Execution manifests retain the remaining stage timings.

Generation records episode seeds and renderer settings. Association and evaluation use saved perception caches, calibration and resampling seeds. Rendering reproducibility is assessed from the saved observations. CPU tests cover schema round-trips, visibility, policy behavior, calibration, prior scaling, matched budgets and intervention invariants, including fragmented records, missing identities and counterfactual worlds. The command `evalx/verify_c2_prop.py --revision-run rev_20260906_a` passes nine checks covering physical-event accounting, labels and candidate identities, event references, unknown-identity bounds, paired intervals, legacy accounting, the geometric margin and decision arithmetic.

# C Additional Experimental Evidence

## C.1 Events, candidates and uncertainty

Events and chance baselines follow Appendix B.8. Any record of the true simulated predecessor counts as correct. The exchanged twin defines the specific-partner event. In 3RScan, designated agreement and selection of any other container member remain separate from annotation-supported alternatives. Candidate-record baselines sample the common same-category map, retaining fragments. Candidate-ID baselines collapse known fragments. Actual policy gates can differ. A missing designated candidate has zero selection probability, an empty set has no uniform-candidate baseline, and unknown labels yield bounds.

The simulated panel has 106 queries, 43 episodes and 86 physical slots. The same-category candidate set contains the true identity for 103 queries. The real panel has 88 queries from 27 distinct reference scans and contains the designated candidate for 84 queries. A higher-level site partition is unavailable. Intervals use paired episode/reference resampling and recompute each query's baseline in every draw. For unknown predictions, outer intervals combine lower-event/larger-baseline and upper-event/smaller-baseline endpoints. Per-scene and leave-one-scene-out results describe sensitivity within the fixed scene collection.

Table 3: Simulation panel event rates. Partner names the specified exchanged identity. Unknown denotes a selected record without an assigned physical identity.

| Policy | Correct | Specific partner | Abs. | Unknown |
|---|---|---|---|---|
| CG-style | 0.019 | 0.906 | 0.000 | 0.019 |
| PDA-MM | 0.019 | 0.962 | 0.000 | 0.009 |
| PDA-DP | 0.019 | 0.925 | 0.000 | 0.009 |
| Cosine–H | 0.019 | 0.811 | 0.000 | 0.028 |
| NN | 0.028 | 0.802 | 0.000 | 0.038 |
| EVA-style | 0.028 | 0.792 | 0.000 | 0.009 |
| TIER proxy | 0.047 | 0.792 | 0.000 | 0.028 |
| Single-exemplar | 0.038 | 0.689 | 0.000 | 0.019 |
| PDA-const | 0.000 | 0.660 | 0.000 | 0.009 |
| CUE | 0.000 | 0.226 | 0.774 | 0.000 |
| CUE+POS | 0.009 | 0.774 | 0.208 | 0.000 |

Table 4: Simulation event differences from event-specific oracle references: $1/K$ for correctness and the specified partner, and $(K-1)/K$ for any other twin. Difference bounds account for unknown candidate identities. Outer 95% intervals additionally reflect paired episode resampling.

| Policy | Event | Difference bounds | Outer 95% CI |
|---|---|---|---|
| CG-style | Correct | $[-0.311, -0.292]$ | $[-0.353, -0.242]$ |
| CG-style | Specific partner | $[+0.575, +0.594]$ | $[+0.516, +0.652]$ |
| CG-style | Any other twin | $[+0.236, +0.255]$ | $[+0.176, +0.311]$ |
| PDA-MM | Correct | $[-0.311, -0.302]$ | $[-0.353, -0.252]$ |
| PDA-MM | Specific partner | $[+0.632, +0.642]$ | $[+0.583, +0.686]$ |
| PDA-MM | Any other twin | $[+0.292, +0.302]$ | $[+0.240, +0.350]$ |
| PDA-DP | Correct | $[-0.311, -0.302]$ | $[-0.353, -0.258]$ |
| PDA-DP | Specific partner | $[+0.594, +0.604]$ | $[+0.539, +0.653]$ |
| PDA-DP | Any other twin | $[+0.264, +0.274]$ | $[+0.205, +0.332]$ |
| Cosine–H | Correct | $[-0.311, -0.283]$ | $[-0.355, -0.228]$ |
| Cosine–H | Specific partner | $[+0.481, +0.509]$ | $[+0.388, +0.593]$ |
| Cosine–H | Any other twin | $[+0.170, +0.198]$ | $[+0.095, +0.275]$ |
| NN | Correct | $[-0.302, -0.264]$ | $[-0.350, -0.201]$ |
| NN | Specific partner | $[+0.472, +0.509]$ | $[+0.383, +0.588]$ |
| NN | Any other twin | $[+0.151, +0.189]$ | $[+0.076, +0.261]$ |
| EVA-style | Correct | $[-0.302, -0.292]$ | $[-0.350, -0.228]$ |
| EVA-style | Specific partner | $[+0.462, +0.472]$ | $[+0.372, +0.555]$ |
| EVA-style | Any other twin | $[+0.142, +0.151]$ | $[+0.063, +0.224]$ |
| TIER proxy | Correct | $[-0.283, -0.255]$ | $[-0.335, -0.190]$ |
| TIER proxy | Specific partner | $[+0.462, +0.491]$ | $[+0.377, +0.574]$ |
| TIER proxy | Any other twin | $[+0.132, +0.160]$ | $[+0.058, +0.227]$ |
| Single-exemplar | Correct | $[-0.292, -0.274]$ | $[-0.340, -0.223]$ |
| Single-exemplar | Specific partner | $[+0.358, +0.377]$ | $[+0.262, +0.472]$ |
| Single-exemplar | Any other twin | $[+0.066, +0.085]$ | $[-0.008, +0.160]$ |
| PDA-const | Correct | $[-0.330, -0.321]$ | $[-0.367, -0.273]$ |
| PDA-const | Specific partner | $[+0.330, +0.340]$ | $[+0.234, +0.434]$ |
| PDA-const | Any other twin | $[-0.009, +0.000]$ | $[-0.073, +0.067]$ |
| CUE | Correct | $-0.330$ | $[-0.367, -0.294]$ |
| CUE | Specific partner | $-0.104$ | $[-0.188, -0.002]$ |
| CUE | Any other twin | $-0.443$ | $[-0.531, -0.351]$ |
| CUE+POS | Correct | $-0.321$ | $[-0.360, -0.281]$ |
| CUE+POS | Specific partner | $+0.443$ | $[+0.339, +0.535]$ |
| CUE+POS | Any other twin | $+0.104$ | $[+0.017, +0.184]$ |

### C.2 TWO-WORLD COUNTERFACTUAL CONTROL

The cached control aligns queries by spatial slot. Removing the final swap changes the identity assigned to the slot while preserving the input used for the recorded output. The control uses one recorded observation and one policy computation. All fragments and both world labels of a base episode are resampled together. Wrong merges, wrong NEW and ABSTAIN cost 5, 1 and 0.25. Unknown candidate predictions produce cost bounds. Per-slot reporting averages fragments within each slot before aggregating episodes.

The risk–frequency curve reweights fixed world-specific risks over the declared $q$ grid. Policy thresholds remain frozen. The analytic two-identity lower envelope is an oracle decision curve with knowledge of $q$, labeled separately from empirical multi-candidate policies. It supplies a hand-checkable connection between the conditional swap result and the deployment mixture.

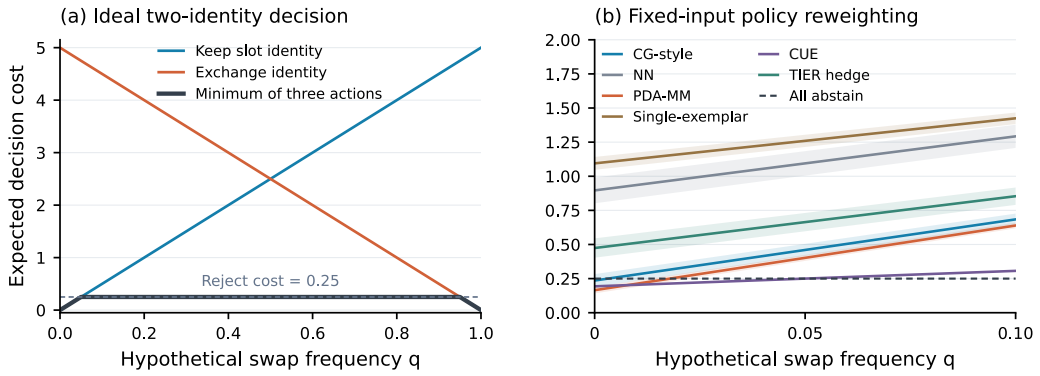


Figure 6: World-frequency sensitivity. (a) Ideal two-identity actions at wrong-link cost 5 and rejection cost 0.25. (b) Fixed cached policies under counterfactual slot labels, shown near low swap frequencies. Shading bounds unknown candidate identities. The complete frequency grid, all policies and paired cluster intervals accompany the source CSV.

### C.3 SAME-INPUT COLOR AND REPRESENTATION EVIDENCE

The controls in Section 6.1 use the protocol and units in Appendix B.3. All 72 formal episodes and both strengths are retained. Each world/camera stratum has 36 fixed physical query slots in 18 scene-placement groups. All queries are detected in this GT-instance route. At either strength, one query in each stratum lacks a true visit-1 candidate and two lack joint true/partner support. Identity accuracy keeps the full 36-slot denominator. True-minus-partner margins condition on the 34 slots with both candidates. The complete candidate maps also include objects outside the exchanged pair.

**World and camera sensitivity.** Tables 5 and 6 retain every outcome of the common all-map rule. Strong RGB and chromaticity recover 35/36 identities in every world/camera stratum. Learned-feature rankings perform well in the unswapped world. With the exchange present, DINOv2 yields no correct identities under either camera path, while MASA-R50 recovers 2/36 under correlated cameras and 4/36 under independent cameras. The contrasting world results describe the dependence of these rankings on scene continuity. They do not estimate the prevalence of either world in deployment.

**Paired decision-cost differences.** Table 7 compares feature choices on the same primary-condition queries at strong color. All four configurations commit on every observed query, so coverage is identical. Wrong links cost 5 and correct links cost zero. RGB reduces per-query cost by 4.861 relative to DINO, with paired interval $[-5.000, -4.583]$. MASA's smaller reduction has an interval reaching zero. These differences characterize the common cosine rule on this stratum. They are separate from the calibration-fitted selective policies in Appendix C.6.

Table 5: Same-input all-map association in the swapped world. Every row sums to 36 fixed query slots. Partner denotes the exchanged individual and Other an additional incorrect identity. Abs., Unk. and Miss denote abstention, unknown candidate identity and non-detection. Corr. and Indep. identify camera paths.

| Camera | Features | $\lambda$ | Correct | Partner | Other | NEW | Abs. | Unk. | Miss |
|---|---|---|---|---|---|---|---|---|---|
| Corr. | Mean RGB | 0 | 1 | 13 | 22 | 0 | 0 | 0 | 0 |
| Corr. | Mean RGB | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | Chromaticity | 0 | 0 | 12 | 24 | 0 | 0 | 0 | 0 |
| Corr. | Chromaticity | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | DINOv2-L | 0 | 0 | 35 | 1 | 0 | 0 | 0 | 0 |
| Corr. | DINOv2-L | 0.85 | 0 | 35 | 1 | 0 | 0 | 0 | 0 |
| Corr. | MASA-R50 | 0 | 0 | 35 | 1 | 0 | 0 | 0 | 0 |
| Corr. | MASA-R50 | 0.85 | 2 | 33 | 1 | 0 | 0 | 0 | 0 |
| Indep. | Mean RGB | 0 | 0 | 11 | 25 | 0 | 0 | 0 | 0 |
| Indep. | Mean RGB | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | Chromaticity | 0 | 2 | 9 | 25 | 0 | 0 | 0 | 0 |
| Indep. | Chromaticity | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | DINOv2-L | 0 | 0 | 35 | 1 | 0 | 0 | 0 | 0 |
| Indep. | DINOv2-L | 0.85 | 0 | 35 | 1 | 0 | 0 | 0 | 0 |
| Indep. | MASA-R50 | 0 | 0 | 33 | 3 | 0 | 0 | 0 | 0 |
| Indep. | MASA-R50 | 0.85 | 4 | 28 | 4 | 0 | 0 | 0 | 0 |

Table 6: Same-input all-map association in the unswapped world. The event partition and denominator match Table 5. Partner now names the other designated twin without implying an exchange occurred.

| Camera | Features | $\lambda$ | Correct | Partner | Other | NEW | Abs. | Unk. | Miss |
|---|---|---|---|---|---|---|---|---|---|
| Corr. | Mean RGB | 0 | 12 | 3 | 21 | 0 | 0 | 0 | 0 |
| Corr. | Mean RGB | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | Chromaticity | 0 | 14 | 0 | 22 | 0 | 0 | 0 | 0 |
| Corr. | Chromaticity | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | DINOv2-L | 0 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | DINOv2-L | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | MASA-R50 | 0 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Corr. | MASA-R50 | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | Mean RGB | 0 | 8 | 1 | 27 | 0 | 0 | 0 | 0 |
| Indep. | Mean RGB | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | Chromaticity | 0 | 7 | 6 | 23 | 0 | 0 | 0 | 0 |
| Indep. | Chromaticity | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | DINOv2-L | 0 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | DINOv2-L | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |
| Indep. | MASA-R50 | 0 | 33 | 0 | 3 | 0 | 0 | 0 | 0 |
| Indep. | MASA-R50 | 0.85 | 35 | 0 | 1 | 0 | 0 | 0 | 0 |

Table 7: Paired changes from DINO under the common all-map rule at strong color in the swap/correlated-camera condition. Accuracy differences are percentage points, and cost differences are per query. Intervals resample the 18 scene-placement groups, retaining paired slots and methods.

| Features minus DINO | $\Delta$ correct (pp) | 95% CI | $\Delta$ cost | 95% CI |
|---|---|---|---|---|
| Mean RGB | $+97.222$ | $[+91.667, +100.000]$ | $-4.861$ | $[-5.000, -4.583]$ |
| Chromaticity | $+97.222$ | $[+91.667, +100.000]$ | $-4.861$ | $[-5.000, -4.583]$ |
| MASA-R50 | $+5.556$ | $[+0.000, +13.889]$ | $-0.278$ | $[-0.694, +0.000]$ |

**Development-fitted feature probe.** Table 8 gives code prediction on all formal visit-2 queries. Every representation supplies a valid entity feature on this population. MASA's primary-condition accuracy increases from 18/36 without color to 34/36 with color. Its strong-color accuracy interval is $[86.1, 100]\%$, while common-cosine identity accuracy is 2/36. On 32 of the 36 queries, the probe predicts the correct marker code while cosine association selects the twin partner. Color-code prediction and identity association are distinct tasks with different training information and candidate structures. The probe establishes recoverable marker information for this readout, and the common rule tests its use in cross-visit ranking. DINO's primary-condition probe interval includes 50%, leaving its decoding ability under other readouts unresolved. All probes use a single development placement. Their intervals describe the fixed test scenes and do not capture variation over training placements.

Table 8: Development-fitted binary marker-code prediction. Counts use all 36 query slots per stratum. The 95% accuracy intervals are percentages. World 0 is unswapped and world 1 swapped. Separate fixed probes are fitted at each strength using only the development episodes.

| World | Camera | Features | $\lambda = 0$ | 95% CI | $\lambda = .85$ | 95% CI |
|---|---|---|---|---|---|---|
| 0 | Corr. | Mean RGB | 20/36 | [41.7, 69.4] | 36/36 | [100.0, 100.0] |
| 0 | Corr. | Chromaticity | 21/36 | [50.0, 66.7] | 36/36 | [100.0, 100.0] |
| 0 | Corr. | DINOv2-L | 22/36 | [41.7, 80.6] | 26/36 | [55.6, 86.1] |
| 0 | Corr. | MASA-R50 | 17/36 | [30.6, 63.9] | 35/36 | [91.7, 100.0] |
| 0 | Indep. | Mean RGB | 18/36 | [36.1, 63.9] | 36/36 | [100.0, 100.0] |
| 0 | Indep. | Chromaticity | 20/36 | [50.0, 63.9] | 36/36 | [100.0, 100.0] |
| 0 | Indep. | DINOv2-L | 19/36 | [36.1, 69.4] | 27/36 | [63.9, 86.1] |
| 0 | Indep. | MASA-R50 | 17/36 | [36.1, 58.3] | 35/36 | [91.7, 100.0] |
| 1 | Corr. | Mean RGB | 16/36 | [30.6, 58.3] | 36/36 | [100.0, 100.0] |
| 1 | Corr. | Chromaticity | 15/36 | [33.3, 50.0] | 36/36 | [100.0, 100.0] |
| 1 | Corr. | DINOv2-L | 12/36 | [16.7, 50.0] | 22/36 | [47.2, 75.0] |
| 1 | Corr. | MASA-R50 | 18/36 | [33.3, 66.7] | 34/36 | [86.1, 100.0] |
| 1 | Indep. | Mean RGB | 18/36 | [36.1, 63.9] | 36/36 | [100.0, 100.0] |
| 1 | Indep. | Chromaticity | 16/36 | [36.1, 50.0] | 36/36 | [100.0, 100.0] |
| 1 | Indep. | DINOv2-L | 18/36 | [33.3, 66.7] | 22/36 | [44.4, 77.8] |
| 1 | Indep. | MASA-R50 | 19/36 | [41.7, 63.9] | 36/36 | [100.0, 100.0] |

**Original cue-model stages.** Table 9 follows the original DINO-based CUE computation on its same-category candidate set. A negative value favors the partner. At both strengths, every evaluable raw appearance margin is negative. Isotonic calibration preserves all 34 signs. Neither clipping nor an appearance-margin collapse occurs. Category evidence ties the two identities. Constellation evidence further favors the partner, and summation preserves this order. The equal prior across existing candidates leaves their relative logits unchanged for a fixed candidate set. Its density-dependent NEW mass affects posterior confidence and competition with NEW. The complete pipeline consequently still requires cross-visit coordinates for that prior.

**Candidate restriction.** The prespecified same-category comparison retains the original features and all four representations. Restricting candidates leaves DINO and MASA correctness unchanged in every world, camera and strength stratum. Unmarked color-descriptor outcomes can change, and some wrong decisions move between partner and other candidates. Complete event partitions remain in the accompanying machine-readable results. Tables 5 and 6 report the full-map task used for the central diagnosis.

Table 9: Mean true-minus-partner margins in the original DINO cue model, swap/correlated-camera condition. Means and negative counts use the 34 slots with both candidates. Two further slots have undefined margins. Raw similarities, LLRs and posterior probabilities retain their separate units.

| Stage | Mean ($\lambda = 0$) | Negative | Mean ($\lambda = .85$) | Negative |
|---|---|---|---|---|
| Raw appearance | $-0.1418$ | 34 | $-0.1152$ | 34 |
| Appearance LLR | $-1.3031$ | 34 | $-0.9618$ | 34 |
| Category LLR | $+0.0000$ | 0 | $+0.0000$ | 0 |
| Raw constellation | $-1.3798$ | 34 | $-1.3798$ | 34 |
| Constellation LLR | $-3.9413$ | 34 | $-3.9413$ | 34 |
| Summed LLR | $-5.2444$ | 34 | $-4.9031$ | 34 |
| Logit | $-5.2444$ | 34 | $-4.9031$ | 34 |
| Posterior | $-0.8171$ | 34 | $-0.8414$ | 34 |

### C.4 Full-map aggregation and readout interventions

These interventions use the fixed features and complete candidate maps specified in Appendix B.4. Every target row retains 36 physical slots, including the slot without a true map candidate. A denotes maximum view-pair cosine, B mean view-pair cosine, and C normalized-mean cosine. The same strong-color probe is used across both worlds, camera regimes and strengths. The original strength-specific probes remain in Table 8.

**Recorded example in Figure 1.** Panels (b,c) use query `v2_r0010` in scene `kujiale_0008`, placement 9101, with a swapped world, correlated cameras and color strength 0.85. Its complete map contains 87 records. The example is selected from the five primary-condition queries for which the fixed probe decodes the query, true predecessor and partner correctly, A and C select the partner, and C+probe selects the true identity. Among these recoveries, we choose the largest minimum visible mask area across the three entities, breaking ties by episode and query ID. This post-hoc readability selection illustrates a recovery. Population effects use the full evaluation set. Each displayed crop is the largest-mask-area view within that entity's original encoded view set, while scores use all cached views. The true candidate ranks 2, 8, 2 and 1 under A, B, C and C+probe. With the development-selected $\beta = 0.1$, its C score changes from 0.743 to 0.740, while the partner changes from 0.779 to 0.639 and the strongest other candidate scores 0.731. Evaluation identities select and explain the example but do not enter scoring. Panel (a) is an AI-generated conceptual illustration, separate from these recorded inputs.

**Recovery and its limits.** In the primary swap/correlated-camera condition, C improves MASA correctness by +22.2 points over A, with paired interval $[+5.6, +41.7]$. This change rescues nine errors and introduces one. Adding the selected marker penalty rescues five more errors without losing a correct target. The latter gain exceeds the point estimates from all 32 random directions in this condition. The two interventions affect different comparisons, and their effects also depend on the penalty weight.

The no-color and unswapped conditions expose the scope of the improvement. Under independent cameras, replacing A with C reduces strong-color unswapped MASA correctness from 35 to 31. Adding the readout increases swapped strong-color correctness from 19 to 21, but its paired interval includes zero. Thus a successful primary contrast does not establish a uniformly better association rule.

Table 10: MASA identity recovery across all conditions, correct targets out of 36. Probe columns use the development-selected penalty for each aggregation. Distance retains only continuous-readout distance. World 0 is unswapped and world 1 swapped.

| World | Camera | $\lambda$ | A | B | C | A+probe | B+probe | C+probe | Distance |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Corr. | 0 | 35 | 27 | 35 | 35 | 27 | 34 | 3 |
| 0 | Corr. | 0.85 | 35 | 35 | 35 | 35 | 35 | 35 | 14 |
| 0 | Indep. | 0 | 33 | 20 | 25 | 33 | 20 | 26 | 1 |
| 0 | Indep. | 0.85 | 35 | 28 | 31 | 35 | 29 | 33 | 9 |
| 1 | Corr. | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 0 |
| 1 | Corr. | 0.85 | 2 | 11 | 10 | 3 | 11 | 15 | 5 |
| 1 | Indep. | 0 | 0 | 5 | 4 | 0 | 5 | 3 | 0 |
| 1 | Indep. | 0.85 | 4 | 15 | 19 | 7 | 16 | 21 | 6 |

Table 11: DINO identity recovery across all conditions, with the same definitions and fixed denominator as Table 10.

| World | Camera | $\lambda$ | A | B | C | A+probe | B+probe | C+probe | Distance |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Corr. | 0 | 35 | 24 | 33 | 35 | 26 | 33 | 3 |
| 0 | Corr. | 0.85 | 35 | 29 | 35 | 35 | 31 | 35 | 18 |
| 0 | Indep. | 0 | 35 | 19 | 26 | 35 | 19 | 25 | 2 |
| 0 | Indep. | 0.85 | 35 | 23 | 31 | 35 | 26 | 32 | 12 |
| 1 | Corr. | 0 | 0 | 7 | 1 | 0 | 6 | 1 | 0 |
| 1 | Corr. | 0.85 | 0 | 8 | 2 | 0 | 8 | 2 | 4 |
| 1 | Indep. | 0 | 0 | 8 | 3 | 0 | 7 | 3 | 0 |
| 1 | Indep. | 0.85 | 0 | 12 | 7 | 1 | 12 | 7 | 2 |

Table 12: Paired MASA accuracy changes in the swapped world, in percentage points with 95% intervals. Each contrast retains all 36 target slots and the same 18 scene-placement groups.

| Camera | $\lambda$ | Contrast | $\Delta$ correct (pp) | 95% CI |
|---|---|---|---|---|
| Corr. | 0 | C minus A | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0 | A+probe minus A | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0 | B+probe minus B | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0 | C+probe minus C | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0.85 | C minus A | $+22.2$ | $[+5.6, +41.7]$ |
| Corr. | 0.85 | A+probe minus A | $+2.8$ | $[+0.0, +8.3]$ |
| Corr. | 0.85 | B+probe minus B | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0.85 | C+probe minus C | $+13.9$ | $[+2.8, +27.8]$ |
| Indep. | 0 | C minus A | $+11.1$ | $[+2.8, +22.2]$ |
| Indep. | 0 | A+probe minus A | $+0.0$ | $[+0.0, +0.0]$ |
| Indep. | 0 | B+probe minus B | $+0.0$ | $[+0.0, +0.0]$ |
| Indep. | 0 | C+probe minus C | $-2.8$ | $[-8.3, +0.0]$ |
| Indep. | 0.85 | C minus A | $+41.7$ | $[+25.0, +58.3]$ |
| Indep. | 0.85 | A+probe minus A | $+8.3$ | $[+0.0, +16.7]$ |
| Indep. | 0.85 | B+probe minus B | $+2.8$ | $[+0.0, +8.3]$ |
| Indep. | 0.85 | C+probe minus C | $+5.6$ | $[-5.6, +19.4]$ |

Table 13: Paired DINO accuracy changes in the swapped world. Definitions and shared resampling follow Table 12.

| Camera | $\lambda$ | Contrast | $\Delta$ correct (pp) | 95% CI |
|---|---|---|---|---|
| Corr. | 0 | C minus A | $+2.8$ | $[+0.0, +8.3]$ |
| Corr. | 0 | A+probe minus A | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0 | B+probe minus B | $-2.8$ | $[-8.3, +0.0]$ |
| Corr. | 0 | C+probe minus C | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0.85 | C minus A | $+5.6$ | $[+0.0, +13.9]$ |
| Corr. | 0.85 | A+probe minus A | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0.85 | B+probe minus B | $+0.0$ | $[+0.0, +0.0]$ |
| Corr. | 0.85 | C+probe minus C | $+0.0$ | $[+0.0, +0.0]$ |
| Indep. | 0 | C minus A | $+8.3$ | $[+0.0, +16.7]$ |
| Indep. | 0 | A+probe minus A | $+0.0$ | $[+0.0, +0.0]$ |
| Indep. | 0 | B+probe minus B | $-2.8$ | $[-8.3, +0.0]$ |
| Indep. | 0 | C+probe minus C | $+0.0$ | $[+0.0, +0.0]$ |
| Indep. | 0.85 | C minus A | $+19.4$ | $[+8.3, +30.6]$ |
| Indep. | 0.85 | A+probe minus A | $+2.8$ | $[+0.0, +8.3]$ |
| Indep. | 0.85 | B+probe minus B | $+0.0$ | $[+0.0, +0.0]$ |
| Indep. | 0.85 | C+probe minus C | $+0.0$ | $[-8.3, +8.3]$ |

**Partner avoidance and complete identity ranking.** Under primary-condition C, the MASA marker penalty reduces specific-partner errors from 23 to 11. Five become correct identities and seven become other incorrect map links. With both candidates available, the mean true-minus-partner margin changes from $-0.081$ to $0.067$ over 34 slots. The mean margin against the strongest wrong candidate remains negative, changing from $-0.086$ to $-0.026$ over 35 slots. Mean true-candidate rank changes from 3.94 to 4.03 on this same support. These statistics distinguish successful top-1 corrections from a general improvement in ranking. Distance-only matching selects no twin partners, yet identifies only five targets and assigns the others to different wrong objects. A binary marker can distinguish the twins while leaving substantial competition elsewhere in the map.

**Penalty weight.** Tables 14 and 15 retain every configured coefficient. The development-selected coefficients differ across aggregations. At a common coefficient, the marker penalty also changes A's rankings, so the selected C result cannot be attributed to a C-specific interaction. Increasing the coefficient eventually reduces ordinary-object accuracy for both representations. The reported selected rules keep the development choices throughout evaluation.

Table 14: Complete MASA coefficient grid in the primary condition. Target counts use 36 slots, and non-target accuracies use 972 labelable queries. Asterisks mark the development-selected coefficient separately for each aggregation.

| | Target correct / 36 | | | Non-target lower bound (%) | | |
|---|---|---|---|---|---|---|
| $\beta$ | A | B | C | A | B | C |
| 0 | $2/36$ | $11/36$ | $10/36$ | 90.3 | 74.8 | 87.9 |
| 0.01 | $2/36$ | $11/36^*$ | $10/36$ | 90.3 | $74.9^*$ | 87.8 |
| 0.03 | $3/36^*$ | $12/36$ | $11/36$ | $90.6^*$ | 75.1 | 87.8 |
| 0.1 | $10/36$ | $14/36$ | $15/36^*$ | 90.7 | 75.5 | $87.7^*$ |
| 0.3 | $17/36$ | $14/36$ | $17/36$ | 90.7 | 76.0 | 87.0 |
| 1 | $15/36$ | $11/36$ | $13/36$ | 89.3 | 72.5 | 85.1 |
| 3 | $12/36$ | $12/36$ | $12/36$ | 83.0 | 66.7 | 78.9 |

Table 15: Complete DINO coefficient grid with the same supports as Table 14. Every coefficient is retained, including settings that lower non-target accuracy.

| | Target correct / 36 | | | Non-target lower bound (%) | | |
|---|---|---|---|---|---|---|
| $\beta$ | A | B | C | A | B | C |
| 0 | 0/36 | 8/36 | 2/36 | 88.9 | 81.2 | 87.3 |
| 0.01 | 0/36 | 8/36 * | 2/36 | 88.9 | 81.4 * | 87.2 |
| 0.03 | 0/36 * | 8/36 | 2/36 * | 89.0 * | 82.2 | 87.3 * |
| 0.1 | 0/36 | 5/36 | 4/36 | 88.9 | 82.6 | 87.0 |
| 0.3 | 2/36 | 5/36 | 6/36 | 88.2 | 81.8 | 85.5 |
| 1 | 4/36 | 5/36 | 5/36 | 84.1 | 77.6 | 81.0 |
| 3 | 5/36 | 5/36 | 5/36 | 75.1 | 67.0 | 72.1 |

**Readout on both visits.** The fixed MASA probe correctly decodes 33 of 36 visit-1 slots and 34 of 36 visit-2 slots in the primary condition. One visit-1 target is absent and remains incorrect in the fixed-slot rate. Tables 16 and 17 report both visits under all conditions. Only marked targets enter decoding accuracy. Every ordinary candidate still competes in association. Reused observations, fragments and the two visits are not independent repetitions.

Table 16: Fixed MASA marker decoding on both visits. Correct/fixed retains missing targets. Observed counts visible physical slots. Conditional accuracy uses observed slots, and available mass averages usable fragments within slots.

| World | Camera | $\lambda$ | Visit | Correct / fixed | Observed | Cond. (%) | Avail. mass |
|---|---|---|---|---|---|---|---|
| 0 | Corr. | 0 | 1 | 20/36 | 35 | 57.1 | 35 |
| 0 | Corr. | 0 | 2 | 23/36 | 36 | 63.9 | 36 |
| 0 | Corr. | 0.85 | 1 | 33/36 | 35 | 94.3 | 35 |
| 0 | Corr. | 0.85 | 2 | 35/36 | 36 | 97.2 | 36 |
| 0 | Indep. | 0 | 1 | 20/36 | 35 | 57.1 | 35 |
| 0 | Indep. | 0 | 2 | 18/36 | 36 | 50.0 | 36 |
| 0 | Indep. | 0.85 | 1 | 33/36 | 35 | 94.3 | 35 |
| 0 | Indep. | 0.85 | 2 | 35/36 | 36 | 97.2 | 36 |
| 1 | Corr. | 0 | 1 | 20/36 | 35 | 57.1 | 35 |
| 1 | Corr. | 0 | 2 | 15/36 | 36 | 41.7 | 36 |
| 1 | Corr. | 0.85 | 1 | 33/36 | 35 | 94.3 | 35 |
| 1 | Corr. | 0.85 | 2 | 34/36 | 36 | 94.4 | 36 |
| 1 | Indep. | 0 | 1 | 20/36 | 35 | 57.1 | 35 |
| 1 | Indep. | 0 | 2 | 19/36 | 36 | 52.8 | 36 |
| 1 | Indep. | 0.85 | 1 | 33/36 | 35 | 94.3 | 35 |
| 1 | Indep. | 0.85 | 2 | 36/36 | 36 | 100.0 | 36 |

**Random-direction distribution.** Each direction uses its own development scale and the same coefficient grid and objective as the learned readout. Tables 18 and 19 summarize all 32 directions in every condition. These are descriptive distributions, without treating directions as independent scene observations or using their maximum to select a method.

**Ordinary-object effects.** In the primary condition, changing MASA from A to C rescues 29 non-target errors and loses 53 correct links, reducing correctness from 878 to 854 of 972 queries. The additional readout loses two further correct links. Non-target differences use the observed labelable query population, with paired group resampling. No unknown non-target labels occur in this cached population. Complete decision transitions, candidate scores and finite-support rank changes accompany these summaries.

Table 17: Fixed DINO marker decoding on both visits, with the same units and missing-target accounting as Table 16.

| World | Camera | $\lambda$ | Visit | Correct / fixed | Observed | Cond. (%) | Avail. mass |
|---|---|---|---|---|---|---|---|
| 0 | Corr. | 0 | 1 | 24/36 | 35 | 68.6 | 35 |
| 0 | Corr. | 0 | 2 | 24/36 | 36 | 66.7 | 36 |
| 0 | Corr. | 0.85 | 1 | 24/36 | 35 | 68.6 | 35 |
| 0 | Corr. | 0.85 | 2 | 26/36 | 36 | 72.2 | 36 |
| 0 | Indep. | 0 | 1 | 24/36 | 35 | 68.6 | 35 |
| 0 | Indep. | 0 | 2 | 24/36 | 36 | 66.7 | 36 |
| 0 | Indep. | 0.85 | 1 | 24/36 | 35 | 68.6 | 35 |
| 0 | Indep. | 0.85 | 2 | 27/36 | 36 | 75.0 | 36 |
| 1 | Corr. | 0 | 1 | 24/36 | 35 | 68.6 | 35 |
| 1 | Corr. | 0 | 2 | 12/36 | 36 | 33.3 | 36 |
| 1 | Corr. | 0.85 | 1 | 24/36 | 35 | 68.6 | 35 |
| 1 | Corr. | 0.85 | 2 | 22/36 | 36 | 61.1 | 36 |
| 1 | Indep. | 0 | 1 | 24/36 | 35 | 68.6 | 35 |
| 1 | Indep. | 0 | 2 | 12/36 | 36 | 33.3 | 36 |
| 1 | Indep. | 0.85 | 1 | 24/36 | 35 | 68.6 | 35 |
| 1 | Indep. | 0.85 | 2 | 22/36 | 36 | 61.1 | 36 |

Table 18: All 32 random-direction MASA controls. Minimum, median and maximum target accuracy are percentages over the fixed 36-slot population. Learned is the marker-trained readout with its own development-selected coefficient.

| World | Camera | $\lambda$ | Agg. | Min | Median | Max | Learned |
|---|---|---|---|---|---|---|---|
| 0 | Corr. | 0 | A | 94.4 | 97.2 | 97.2 | 97.2 |
| 0 | Corr. | 0 | B | 66.7 | 75.0 | 75.0 | 75.0 |
| 0 | Corr. | 0 | C | 91.7 | 97.2 | 97.2 | 94.4 |
| 0 | Corr. | 0.85 | A | 97.2 | 97.2 | 97.2 | 97.2 |
| 0 | Corr. | 0.85 | B | 94.4 | 97.2 | 97.2 | 97.2 |
| 0 | Corr. | 0.85 | C | 97.2 | 97.2 | 97.2 | 97.2 |
| 0 | Indep. | 0 | A | 69.4 | 91.7 | 91.7 | 91.7 |
| 0 | Indep. | 0 | B | 52.8 | 55.6 | 55.6 | 55.6 |
| 0 | Indep. | 0 | C | 66.7 | 69.4 | 69.4 | 72.2 |
| 0 | Indep. | 0.85 | A | 86.1 | 97.2 | 97.2 | 97.2 |
| 0 | Indep. | 0.85 | B | 77.8 | 77.8 | 80.6 | 80.6 |
| 0 | Indep. | 0.85 | C | 80.6 | 86.1 | 86.1 | 91.7 |
| 1 | Corr. | 0 | A | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | Corr. | 0 | B | 5.6 | 8.3 | 13.9 | 8.3 |
| 1 | Corr. | 0 | C | 0.0 | 0.0 | 2.8 | 0.0 |
| 1 | Corr. | 0.85 | A | 0.0 | 5.6 | 19.4 | 8.3 |
| 1 | Corr. | 0.85 | B | 30.6 | 30.6 | 33.3 | 30.6 |
| 1 | Corr. | 0.85 | C | 25.0 | 27.8 | 30.6 | 41.7 |
| 1 | Indep. | 0 | A | 0.0 | 0.0 | 5.6 | 0.0 |
| 1 | Indep. | 0 | B | 11.1 | 13.9 | 19.4 | 13.9 |
| 1 | Indep. | 0 | C | 8.3 | 11.1 | 11.1 | 8.3 |
| 1 | Indep. | 0.85 | A | 8.3 | 11.1 | 36.1 | 19.4 |
| 1 | Indep. | 0.85 | B | 41.7 | 41.7 | 44.4 | 44.4 |
| 1 | Indep. | 0.85 | C | 47.2 | 52.8 | 52.8 | 58.3 |

Table 19: All 32 random-direction DINO controls, with the same selection budget and units as Table 18.

| World | Camera | $\lambda$ | Agg. | Min | Median | Max | Learned |
|---|---|---|---|---|---|---|---|
| 0 | Corr. | 0 | A | 86.1 | 97.2 | 97.2 | 97.2 |
| 0 | Corr. | 0 | B | 63.9 | 66.7 | 77.8 | 72.2 |
| 0 | Corr. | 0 | C | 77.8 | 91.7 | 94.4 | 91.7 |
| 0 | Corr. | 0.85 | A | 91.7 | 97.2 | 97.2 | 97.2 |
| 0 | Corr. | 0.85 | B | 69.4 | 80.6 | 88.9 | 86.1 |
| 0 | Corr. | 0.85 | C | 91.7 | 97.2 | 97.2 | 97.2 |
| 0 | Indep. | 0 | A | 58.3 | 97.2 | 97.2 | 97.2 |
| 0 | Indep. | 0 | B | 44.4 | 52.8 | 61.1 | 52.8 |
| 0 | Indep. | 0 | C | 41.7 | 72.2 | 72.2 | 69.4 |
| 0 | Indep. | 0.85 | A | 75.0 | 97.2 | 97.2 | 97.2 |
| 0 | Indep. | 0.85 | B | 58.3 | 63.9 | 75.0 | 72.2 |
| 0 | Indep. | 0.85 | C | 58.3 | 86.1 | 86.1 | 88.9 |
| 1 | Corr. | 0 | A | 0.0 | 0.0 | 5.6 | 0.0 |
| 1 | Corr. | 0 | B | 5.6 | 18.1 | 19.4 | 16.7 |
| 1 | Corr. | 0 | C | 0.0 | 2.8 | 5.6 | 2.8 |
| 1 | Corr. | 0.85 | A | 0.0 | 0.0 | 16.7 | 0.0 |
| 1 | Corr. | 0.85 | B | 13.9 | 22.2 | 25.0 | 22.2 |
| 1 | Corr. | 0.85 | C | 2.8 | 5.6 | 11.1 | 5.6 |
| 1 | Indep. | 0 | A | 0.0 | 0.0 | 8.3 | 0.0 |
| 1 | Indep. | 0 | B | 8.3 | 22.2 | 22.2 | 19.4 |
| 1 | Indep. | 0 | C | 5.6 | 8.3 | 16.7 | 8.3 |
| 1 | Indep. | 0.85 | A | 0.0 | 0.0 | 22.2 | 2.8 |
| 1 | Indep. | 0.85 | B | 16.7 | 33.3 | 38.9 | 33.3 |
| 1 | Indep. | 0.85 | C | 13.9 | 19.4 | 30.6 | 19.4 |

The bootstrap preserves all conditions within each scene-placement group. Its uncertainty is conditional on the fitted readout and the existing six scenes. Scene-independent confirmation, alternative development placements and object-attached cues remain prospective evaluations.

Table 20: MASA ordinary-object outcomes in the unswapped world. Base and +probe are accuracy percentages. Rescue and Harm count changes relative to the same aggregation. Unk. retains unannotated candidate choices. The final column gives the paired accuracy-change interval.

| Camera | $\lambda$ | Agg. | $N$ | Base | +probe | Rescue | Harm | Unk. | $\Delta$ 95% CI |
|---|---|---|---|---|---|---|---|---|---|
| Corr. | 0 | A | 972 | 91.0 | 91.2 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Corr. | 0 | B | 972 | 74.2 | 74.3 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Corr. | 0 | C | 972 | 87.9 | 87.4 | 0 | 4 | 0 | $[-0.8, -0.1]$ |
| Corr. | 0.85 | A | 972 | 91.5 | 91.5 | 0 | 0 | 0 | $[+0.0, +0.0]$ |
| Corr. | 0.85 | B | 972 | 75.0 | 75.2 | 2 | 0 | 0 | $[+0.0, +0.5]$ |
| Corr. | 0.85 | C | 972 | 88.4 | 88.1 | 0 | 3 | 0 | $[-0.6, +0.0]$ |
| Indep. | 0 | A | 954 | 83.1 | 83.4 | 3 | 0 | 0 | $[+0.0, +0.7]$ |
| Indep. | 0 | B | 954 | 64.5 | 64.7 | 2 | 0 | 0 | $[+0.0, +0.5]$ |
| Indep. | 0 | C | 954 | 79.2 | 79.1 | 5 | 6 | 0 | $[-1.0, +0.7]$ |
| Indep. | 0.85 | A | 954 | 83.1 | 83.5 | 4 | 0 | 0 | $[+0.1, +0.7]$ |
| Indep. | 0.85 | B | 954 | 64.8 | 65.3 | 5 | 0 | 0 | $[+0.1, +1.1]$ |
| Indep. | 0.85 | C | 954 | 79.7 | 79.8 | 8 | 7 | 0 | $[-0.6, +1.0]$ |

Table 21: MASA ordinary-object outcomes in the swapped world. Base and +probe are accuracy percentages. Rescue and Harm count changes relative to the same aggregation. Unk. retains unannotated candidate choices. The final column gives the paired accuracy-change interval.

| Camera | $\lambda$ | Agg. | $N$ | Base | +probe | Rescue | Harm | Unk. | $\Delta$ 95% CI |
|---|---|---|---|---|---|---|---|---|---|
| Corr. | 0 | A | 972 | 90.4 | 90.7 | 3 | 0 | 0 | $[+0.0, +0.7]$ |
| Corr. | 0 | B | 972 | 74.5 | 74.5 | 0 | 0 | 0 | $[+0.0, +0.0]$ |
| Corr. | 0 | C | 972 | 87.7 | 87.7 | 2 | 2 | 0 | $[-0.5, +0.4]$ |
| Corr. | 0.85 | A | 972 | 90.3 | 90.6 | 3 | 0 | 0 | $[+0.0, +0.6]$ |
| Corr. | 0.85 | B | 972 | 74.8 | 74.9 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Corr. | 0.85 | C | 972 | 87.9 | 87.7 | 0 | 2 | 0 | $[-0.5, +0.0]$ |
| Indep. | 0 | A | 954 | 83.1 | 83.1 | 1 | 1 | 0 | $[-0.3, +0.3]$ |
| Indep. | 0 | B | 954 | 64.3 | 64.5 | 2 | 0 | 0 | $[+0.0, +0.5]$ |
| Indep. | 0 | C | 954 | 79.1 | 79.6 | 8 | 4 | 0 | $[-0.4, +1.2]$ |
| Indep. | 0.85 | A | 954 | 83.0 | 82.9 | 1 | 2 | 0 | $[-0.4, +0.0]$ |
| Indep. | 0.85 | B | 954 | 64.5 | 64.8 | 3 | 0 | 0 | $[+0.0, +0.8]$ |
| Indep. | 0.85 | C | 954 | 79.6 | 79.9 | 8 | 5 | 0 | $[-0.3, +0.9]$ |

Table 22: DINO ordinary-object outcomes in the unswapped world. Base and +probe are accuracy percentages. Rescue and Harm count changes relative to the same aggregation. Unk. retains unannotated candidate choices. The final column gives the paired accuracy-change interval.

| Camera | $\lambda$ | Agg. | $N$ | Base | +probe | Rescue | Harm | Unk. | $\Delta$ 95% CI |
|---|---|---|---|---|---|---|---|---|---|
| Corr. | 0 | A | 972 | 87.9 | 88.1 | 2 | 0 | 0 | $[+0.0, +0.6]$ |
| Corr. | 0 | B | 972 | 80.0 | 80.2 | 2 | 0 | 0 | $[+0.0, +0.5]$ |
| Corr. | 0 | C | 972 | 87.9 | 87.8 | 1 | 2 | 0 | $[-0.6, +0.3]$ |
| Corr. | 0.85 | A | 972 | 88.1 | 88.2 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Corr. | 0.85 | B | 972 | 80.7 | 80.8 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Corr. | 0.85 | C | 972 | 87.8 | 87.6 | 0 | 2 | 0 | $[-0.5, +0.0]$ |
| Indep. | 0 | A | 954 | 85.1 | 85.1 | 5 | 5 | 0 | $[-0.5, +0.5]$ |
| Indep. | 0 | B | 954 | 74.6 | 74.7 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Indep. | 0 | C | 954 | 80.6 | 80.6 | 4 | 4 | 0 | $[-0.3, +0.3]$ |
| Indep. | 0.85 | A | 954 | 86.2 | 85.7 | 1 | 5 | 0 | $[-0.9, +0.0]$ |
| Indep. | 0.85 | B | 954 | 75.6 | 75.8 | 2 | 0 | 0 | $[+0.0, +0.5]$ |
| Indep. | 0.85 | C | 954 | 81.1 | 81.2 | 5 | 4 | 0 | $[-0.2, +0.4]$ |

Table 23: DINO ordinary-object outcomes in the swapped world. Base and +probe are accuracy percentages. Rescue and Harm count changes relative to the same aggregation. Unk. retains unannotated candidate choices. The final column gives the paired accuracy-change interval.

| Camera | $\lambda$ | Agg. | $N$ | Base | +probe | Rescue | Harm | Unk. | $\Delta$ 95% CI |
|---|---|---|---|---|---|---|---|---|---|
| Corr. | 0 | A | 972 | 88.4 | 88.6 | 3 | 1 | 0 | $[-0.2, +0.6]$ |
| Corr. | 0 | B | 972 | 79.9 | 80.6 | 6 | 0 | 0 | $[+0.2, +1.1]$ |
| Corr. | 0 | C | 972 | 87.4 | 87.6 | 5 | 4 | 0 | $[-0.4, +0.7]$ |
| Corr. | 0.85 | A | 972 | 88.9 | 89.0 | 1 | 0 | 0 | $[+0.0, +0.3]$ |
| Corr. | 0.85 | B | 972 | 81.2 | 81.4 | 2 | 0 | 0 | $[+0.0, +0.6]$ |
| Corr. | 0.85 | C | 972 | 87.3 | 87.3 | 4 | 4 | 0 | $[-0.5, +0.5]$ |
| Indep. | 0 | A | 954 | 84.7 | 84.9 | 5 | 3 | 0 | $[-0.4, +0.9]$ |
| Indep. | 0 | B | 954 | 74.7 | 74.8 | 2 | 1 | 0 | $[-0.2, +0.5]$ |
| Indep. | 0 | C | 954 | 81.0 | 80.8 | 2 | 4 | 0 | $[-0.8, +0.3]$ |
| Indep. | 0.85 | A | 954 | 85.8 | 85.8 | 2 | 2 | 0 | $[-0.3, +0.3]$ |
| Indep. | 0.85 | B | 954 | 75.7 | 75.9 | 3 | 1 | 0 | $[-0.2, +0.6]$ |
| Indep. | 0.85 | C | 954 | 81.6 | 81.1 | 2 | 6 | 0 | $[-1.0, +0.2]$ |

### C.5 Persistent cues and direct observation controls

These controls retain each association rule's original candidate and perception support. The oracle-mask readout uses two candidates on 36 primary-condition slots, while the viewing-direction comparison uses 32 eligible slots. GT-instance and full-SAM3 routes retain their own geometry and membership, with query fragments sharing their physical slot's weight. Tables below separate outcome partitions, paired scene-placement intervals and equal-scene sensitivity.

Both formal perception routes complete the 72-episode grid in Appendix B.2. Each route has 720 strength/background conditions and 7,920 condition/policy cells. The fixed physical population contains 15840 slot/condition/policy rows. GT-instance yields 15840 observed query rows, while full-SAM3 yields 25344 because a physical slot can generate several perceived fragments. Each slot contributes unit weight to the outcome curves, divided equally among its observed fragments. Non-detection contributes its own event mass. Development captures and features are excluded.

**Readout support and cue use.** Table 24 reports the fixed oracle-mask color readout over all 36 main-condition slots. With strong color, its 34 correct slots support a direct check of cue availability. The scalar readout uses foreground color and two GT target candidates. Ordinary policies use their own feature scores and candidate gates. Readout-conditioned abstention measures CUE unresolved outputs on these correctly read slots, with non-detection separate. Its information and candidate support are specific to this oracle control.

Table 24: Oracle-mask color readout in the swap/correlated-camera condition. Counts retain all 36 physical slots. Correctness rates and intervals are percentages. Missing denotes insufficient visible support. Intervals resample the 18 scene-placement groups.

| $\lambda$ | Correct | Partner | Missing | Correct rate [95% CI] |
|---|---|---|---|---|
| 0 | 7/36 | 27 | 2 | 19.4 [5.6, 33.3] |
| 0.05 | 33/36 | 1 | 2 | 91.7 [77.8, 100.0] |
| 0.15 | 34/36 | 0 | 2 | 94.4 [83.3, 100.0] |
| 0.4 | 34/36 | 0 | 2 | 94.4 [83.3, 100.0] |
| 0.85 | 34/36 | 0 | 2 | 94.4 [83.3, 100.0] |

Table 25 follows the measurement chain from physical slots through detection, fragmentation and candidate availability. The shared appearance margin uses only slots with both candidates. Among the correctly read slots at strong color, CUE abstention is 23.5% $[11.8, 35.3]$ in GT and 62.7%

$[46.7, 79.2]$ in full SAM3. These percentages condition on the oracle readout's information and candidate set. The complete conditional policy traces remain in the accompanying results.

Table 25: Primary-condition measurement support at every strength. Detected, True cand. and Both cand. count physical slots. Fragments counts observed query records. Margin is the mean true-minus-partner max-prototype cosine score on both-candidate support. Shared fields are checked across all 11 policies before collapse.

| Route | $\lambda$ | Slots | Detected | Fragments | True cand. | Both cand. | Margin |
|---|---|---|---|---|---|---|---|
| GT | 0 | 36 | 36 | 36 | 35 | 34 | -0.142 |
| GT | 0.05 | 36 | 36 | 36 | 35 | 34 | -0.138 |
| GT | 0.15 | 36 | 36 | 36 | 35 | 34 | -0.133 |
| GT | 0.4 | 36 | 36 | 36 | 35 | 34 | -0.126 |
| GT | 0.85 | 36 | 36 | 36 | 35 | 34 | -0.115 |
| SAM3 | 0 | 36 | 35 | 59 | 35 | 34 | -0.172 |
| SAM3 | 0.05 | 36 | 35 | 57 | 35 | 34 | -0.184 |
| SAM3 | 0.15 | 36 | 35 | 60 | 34 | 32 | -0.124 |
| SAM3 | 0.4 | 36 | 35 | 61 | 34 | 32 | -0.092 |
| SAM3 | 0.85 | 36 | 35 | 56 | 35 | 34 | -0.140 |

**Complete states and paired effects.** Tables 26 and 27 retain every primary and exploratory policy at zero and strong color. Tables 28 and 29 give their paired differences and intervals. All intermediate strengths, both worlds, both camera regimes and both backgrounds remain in the complete results. Scene-weighted effects and intervals provide sensitivity to the six-scene population. Full-SAM3 detection, candidate membership, centroids and prototype selection can change with strength, so a cross-route difference includes each of these factors.

GT max-prototype cosine margins change from -0.142 to -0.115 on 34 both-candidate slots, a paired increase of 0.027 $[0.019, 0.036]$. Full-SAM3 endpoint margins are -0.172 and -0.140. Color affects these representations, whose mean true-versus-partner comparison still favors the partner.

Table 26: GT-instance fixed-slot outcome percentages at zero and strong color, swap/correlated-camera/context condition. Each row retains correct, partner, wrong-NEW, other wrong-merge, abstention, unknown-candidate and non-detection mass. Rounding can affect the displayed sum.

| Policy | $\lambda$ | Correct | Partner | NEW | Outside | Abstain | Unknown | Miss |
|---|---|---|---|---|---|---|---|---|
| CG-style | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| CG-style | 0.85 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| NN | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| NN | 0.85 | 0.0 | 97.2 | 0.0 | 2.8 | 0.0 | 0.0 | 0.0 |
| Cosine-H | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| Cosine-H | 0.85 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| Single-exemplar | 0 | 13.9 | 75.0 | 2.8 | 8.3 | 0.0 | 0.0 | 0.0 |
| Single-exemplar | 0.85 | 22.2 | 75.0 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| TIER proxy | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| TIER proxy | 0.85 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| PDA-MM | 0 | 0.0 | 97.2 | 0.0 | 2.8 | 0.0 | 0.0 | 0.0 |
| PDA-MM | 0.85 | 0.0 | 97.2 | 0.0 | 2.8 | 0.0 | 0.0 | 0.0 |
| PDA-C | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| PDA-C | 0.85 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| CUE | 0 | 0.0 | 72.2 | 0.0 | 0.0 | 27.8 | 0.0 | 0.0 |
| CUE | 0.85 | 0.0 | 72.2 | 0.0 | 0.0 | 27.8 | 0.0 | 0.0 |
| CUE+pos | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| CUE+pos | 0.85 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| EVA-style | 0 | 0.0 | 97.2 | 2.8 | 0.0 | 0.0 | 0.0 | 0.0 |
| EVA-style | 0.85 | 0.0 | 97.2 | 0.0 | 2.8 | 0.0 | 0.0 | 0.0 |
| PDA-DP | 0 | 0.0 | 91.7 | 2.8 | 5.6 | 0.0 | 0.0 | 0.0 |
| PDA-DP | 0.85 | 0.0 | 94.4 | 2.8 | 2.8 | 0.0 | 0.0 | 0.0 |

Table 27: Full-SAM3 fixed-slot outcome percentages under the same conditions as Table 26. Rates average fragments within each physical slot.

| Policy | $\lambda$ | Correct | Partner | NEW | Outside | Abstain | Unknown | Miss |
|---|---|---|---|---|---|---|---|---|
| CG-style | 0 | 0.0 | 91.4 | 4.9 | 0.0 | 0.0 | 0.9 | 2.8 |
| CG-style | 0.85 | 0.0 | 88.0 | 2.3 | 6.9 | 0.0 | 0.0 | 2.8 |
| NN | 0 | 0.0 | 90.0 | 2.8 | 4.4 | 0.0 | 0.0 | 2.8 |
| NN | 0.85 | 0.7 | 85.4 | 0.0 | 10.4 | 0.0 | 0.7 | 2.8 |
| Cosine-H | 0 | 0.0 | 87.0 | 3.5 | 6.7 | 0.0 | 0.0 | 2.8 |
| Cosine-H | 0.85 | 0.0 | 85.4 | 3.7 | 7.4 | 0.0 | 0.7 | 2.8 |
| Single-exemplar | 0 | 0.0 | 81.2 | 4.9 | 10.2 | 0.0 | 0.9 | 2.8 |
| Single-exemplar | 0.85 | 3.5 | 77.1 | 4.6 | 12.0 | 0.0 | 0.0 | 2.8 |
| TIER proxy | 0 | 0.0 | 90.0 | 6.2 | 0.9 | 0.0 | 0.0 | 2.8 |
| TIER proxy | 0.85 | 0.0 | 84.0 | 5.3 | 7.9 | 0.0 | 0.0 | 2.8 |
| PDA-MM | 0 | 0.0 | 91.9 | 0.0 | 5.3 | 0.0 | 0.0 | 2.8 |
| PDA-MM | 0.85 | 0.0 | 91.7 | 0.0 | 5.6 | 0.0 | 0.0 | 2.8 |
| PDA-C | 0 | 0.0 | 83.1 | 14.1 | 0.0 | 0.0 | 0.0 | 2.8 |
| PDA-C | 0.85 | 0.0 | 79.9 | 16.0 | 1.4 | 0.0 | 0.0 | 2.8 |
| CUE | 0 | 0.0 | 35.4 | 2.8 | 4.9 | 54.2 | 0.0 | 2.8 |
| CUE | 0.85 | 0.0 | 32.9 | 0.0 | 2.3 | 62.0 | 0.0 | 2.8 |
| CUE+pos | 0 | 0.0 | 80.3 | 7.2 | 2.1 | 7.6 | 0.0 | 2.8 |
| CUE+pos | 0.85 | 0.0 | 84.5 | 0.7 | 3.7 | 8.3 | 0.0 | 2.8 |
| EVA-style | 0 | 0.0 | 84.0 | 9.5 | 3.7 | 0.0 | 0.0 | 2.8 |
| EVA-style | 0.85 | 0.0 | 85.4 | 2.3 | 9.5 | 0.0 | 0.0 | 2.8 |
| PDA-DP | 0 | 0.0 | 85.0 | 0.0 | 12.3 | 0.0 | 0.0 | 2.8 |
| PDA-DP | 0.85 | 0.0 | 87.0 | 0.0 | 10.2 | 0.0 | 0.0 | 2.8 |

Table 28: GT-instance effects of strong versus zero color on the 36 main-condition slots. Each cell gives the change in percentage points above its paired 95% interval. Identical zero results are grouped explicitly below. PDA-DP is exploratory.

| Policy | $\Delta$ correct | $\Delta$ partner | $\Delta$ abstain |
|---|---|---|---|
| | 8.3 | 0.0 | 0.0 |
| Single-exemplar | $[-2.8, 19.4]$ | $[-11.1, 11.1]$ | $[0.0, 0.0]$ |
| | 0.0 | 2.8 | 0.0 |
| PDA-DP | $[0.0, 0.0]$ | $[0.0, 8.3]$ | $[0.0, 0.0]$ |

All three estimates and their 95% intervals are $0.0\,[0.0, 0.0]$ for CG-style, NN, Cosine-H, TIER proxy, PDA-MM, PDA-C, CUE, CUE+pos and EVA-style.

Table 29: Full-SAM3 effects of strong versus zero color on the same 36 physical slots. Each cell gives the effect estimate or unknown-identity bounds above its paired 95% interval, in percentage points. Fragments share each slot's weight.

| Policy | $\Delta$ correct | $\Delta$ partner | $\Delta$ abstain |
|---|---|---|---|
| CG-style | $[-0.9, 0.0]$<br>$[-2.8, 0.0]$ | $-3.5$<br>$[-9.7, 2.8]$ | $0.0$<br>$[0.0, 0.0]$ |
| NN | $[0.7, 1.4]$<br>$[0.0, 4.2]$ | $-4.6$<br>$[-11.6, 2.3]$ | $0.0$<br>$[0.0, 0.0]$ |
| Cosine-H | $[0.0, 0.7]$<br>$[0.0, 2.1]$ | $-1.6$<br>$[-8.3, 5.1]$ | $0.0$<br>$[0.0, 0.0]$ |
| Single-exemplar | $[2.5, 3.5]$<br>$[-2.1, 9.7]$ | $-4.2$<br>$[-14.8, 6.9]$ | $0.0$<br>$[0.0, 0.0]$ |
| TIER proxy | $0.0$<br>$[0.0, 0.0]$ | $-6.0$<br>$[-13.0, 1.4]$ | $0.0$<br>$[0.0, 0.0]$ |
| PDA-MM | $0.0$<br>$[0.0, 0.0]$ | $-0.2$<br>$[-6.5, 5.8]$ | $0.0$<br>$[0.0, 0.0]$ |
| PDA-C | $0.0$<br>$[0.0, 0.0]$ | $-3.2$<br>$[-17.1, 8.3]$ | $0.0$<br>$[0.0, 0.0]$ |
| CUE | $0.0$<br>$[0.0, 0.0]$ | $-2.5$<br>$[-25.0, 19.7]$ | $7.9$<br>$[-13.2, 29.4]$ |
| CUE+pos | $0.0$<br>$[0.0, 0.0]$ | $4.2$<br>$[-3.7, 12.5]$ | $0.7$<br>$[-6.2, 8.1]$ |
| EVA-style | $0.0$<br>$[0.0, 0.0]$ | $1.4$<br>$[-6.0, 8.8]$ | $0.0$<br>$[0.0, 0.0]$ |
| PDA-DP | $0.0$<br>$[0.0, 0.0]$ | $2.1$<br>$[-0.5, 5.6]$ | $0.0$<br>$[0.0, 0.0]$ |

**Neutral background.** Within a route and strength, crop bounds, masks and entity membership are fixed. Only the pixels outside the foreground are replaced before embedding extraction. Tables 30 and 31 report the paired outcome changes at strong color. This design isolates the effect of that background replacement on the recorded pipeline. The accompanying cue-margin tables retain both-candidate support, observed-centroid position scores and appearance/decision conflicts.

Table 30: GT-instance effects of neutral versus context background at strong color, swap/correlated-camera condition. Changes are percentage points, with estimate bounds and paired intervals as in Table 28. Masks, membership and geometry are fixed.

| Policy | $\Delta$ correct | $\Delta$ partner | $\Delta$ abstain |
|---|---|---|---|
| CG-style | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| NN | $0.0$ $[0.0, 0.0]$ | $-2.8$ $[-8.3, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| Cosine-H | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| Single-exemplar | $0.0$ $[-8.3, 8.3]$ | $-2.8$ $[-8.3, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| TIER proxy | $0.0$ $[0.0, 0.0]$ | $-2.8$ $[-8.3, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| PDA-MM | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| PDA-C | $0.0$ $[0.0, 0.0]$ | $-2.8$ $[-8.3, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| CUE | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| CUE+pos | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| EVA-style | $0.0$ $[0.0, 0.0]$ | $-2.8$ $[-8.3, 0.0]$ | $0.0$ $[0.0, 0.0]$ |
| PDA-DP | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ | $0.0$ $[0.0, 0.0]$ |

Table 31: Full-SAM3 effects of neutral versus context background under the same intervention as Table 30. Masks, membership and geometry are fixed within this route and strength. Units and interval notation follow Table 28.

| Policy | Δ correct | Δ partner | Δ abstain |
|---|---|---|---|
| CG-style | 0.0<br>[0.0, 0.0] | 1.4<br>[0.0, 4.2] | 0.0<br>[0.0, 0.0] |
| NN | [−0.5, 0.2]<br>[−4.2, 2.8] | −0.9<br>[−2.8, 0.0] | 0.0<br>[0.0, 0.0] |
| Cosine-H | [−0.7, 0.0]<br>[−2.1, 0.0] | 0.0<br>[0.0, 0.0] | 0.0<br>[0.0, 0.0] |
| Single-exemplar | 1.2<br>[−2.1, 5.6] | −4.6<br>[−12.5, 1.9] | 0.0<br>[0.0, 0.0] |
| TIER proxy | 0.9<br>[0.0, 2.8] | 0.5<br>[−2.8, 4.2] | 0.0<br>[0.0, 0.0] |
| PDA-MM | 0.0<br>[0.0, 0.0] | 0.0<br>[0.0, 0.0] | 0.0<br>[0.0, 0.0] |
| PDA-C | 0.0<br>[0.0, 0.0] | 6.9<br>[0.0, 16.7] | 0.0<br>[0.0, 0.0] |
| CUE | 0.0<br>[0.0, 0.0] | −1.4<br>[−4.2, 0.0] | 1.4<br>[0.0, 4.2] |
| CUE+pos | 0.0<br>[0.0, 0.0] | 0.0<br>[0.0, 0.0] | 0.9<br>[0.0, 2.8] |
| EVA-style | 0.9<br>[0.0, 2.8] | −0.9<br>[−2.8, 0.0] | 0.0<br>[0.0, 0.0] |
| PDA-DP | 0.0<br>[0.0, 0.0] | −1.4<br>[−4.2, 0.0] | 0.0<br>[0.0, 0.0] |

**Viewing-direction support.** Of 72 formal episode IDs, 62 supply two to five eligible matched tuples and ten lack sufficient support. The primary condition retains 32 paired slots from 16 scene-placement groups. Selection uses GT object-relative bearings with non-target features, membership and geometry fixed. Camera distance and roll remain uncontrolled. Table 32 separates matched views from equal-count frame-order selection and reports equal-scene sensitivity on this same eligible population. The CUE partner effect under scene resampling has a zero lower endpoint, retaining uncertainty across the six studied scenes.

Table 32: Viewing-direction controls on 32 eligible strong-color slots. Contrasts I–III are matched minus original, matched minus equal-count, and equal-count minus original, with paired scene-placement intervals. IV repeats I with equal-scene weighting and scene resampling. Cells give percentage-point changes above 95% intervals. The named zero-result group retains all other policies.

| Contrast | Policy | Δ correct | Δ partner | Δ abstain |
|---|---|---|---|---|
| I | Single-exemplar | $-9.4$<br>$[-23.1, 3.1]$ | $9.4$<br>$[-3.1, 23.1]$ | $0.0$<br>$[0.0, 0.0]$ |
| I | CUE | $0.0$<br>$[0.0, 0.0]$ | $12.5$<br>$[2.9, 23.5]$ | $-12.5$<br>$[-23.5, -2.9]$ |
| II | Single-exemplar | $-9.4$<br>$[-23.1, 3.1]$ | $12.5$<br>$[0.0, 26.5]$ | $0.0$<br>$[0.0, 0.0]$ |
| II | CUE | $0.0$<br>$[0.0, 0.0]$ | $12.5$<br>$[2.9, 23.5]$ | $-12.5$<br>$[-23.5, -2.9]$ |
| III | Single-exemplar | $0.0$<br>$[0.0, 0.0]$ | $-3.1$<br>$[-10.0, 0.0]$ | $0.0$<br>$[0.0, 0.0]$ |
| III | CUE | $0.0$<br>$[0.0, 0.0]$ | $0.0$<br>$[0.0, 0.0]$ | $0.0$<br>$[0.0, 0.0]$ |
| IV | Single-exemplar | $-9.7$<br>$[-20.8, 0.0]$ | $9.7$<br>$[0.0, 20.8]$ | $0.0$<br>$[0.0, 0.0]$ |
| IV | CUE | $0.0$<br>$[0.0, 0.0]$ | $12.5$<br>$[0.0, 29.2]$ | $-12.5$<br>$[-29.2, 0.0]$ |

All three estimates and their 95% intervals are $0.0\,[0.0, 0.0]$ for CG-style, NN, Cosine-H, TIER proxy, PDA-MM, PDA-C, CUE+pos, EVA-style and PDA-DP in each of contrasts I–IV.

**Observed-query selective risk.** At strong color in the primary condition, every one of the 11 policies reaches the requested 20% budget and incurs 100% retained-query risk in both routes. This retains seven of 36 GT queries and eleven of 56 full-SAM3 queries, giving 19.4% and 19.6% coverage after integer rounding. The observed-query universe and budget are fixed within each route. Their fragment-level denominator differs from the physical-slot curves and eligible view subset. Non-detections remain in the slot partition without imputed decision cost. The accompanying table archive retains all policy rows, query-universe hashes, tie rules and unknown-label bounds, which coincide here.

### C.6 Selective risk and decision cost

The historical ambiguity slice is `category=chair`, $K \geq 2$, and ($d > 0$ or swap). Its 1116 observations include 80 queries without a predecessor label. Main evaluation fixes the remaining 1036 queries independently of actions. All-query coverage and unknown-query frequency are also reported. Correct merges and correct NEW cost zero. Wrong merges cost $C_{\mathrm{OM}}$, wrong NEW cost $C_{\mathrm{OS}}$, and abstentions cost $c_{\mathrm{abs}}$. Costs are divided by the fixed labeled-query count. A selected candidate of unknown identity yields lower and upper costs.

Calibration uses the original easy-cell transfer protocol. Simulation's seven easy episodes provide 638 observations and share all six scenes with evaluation. They are episode-disjoint from the ambiguity slice. Independent full-grid testing excludes these episodes. A compatibility report retains the historical universe. Real evaluation excludes all six reserve calibration pairs, four of which satisfy the fitting criteria, leaving 715 test queries. Manifests list accepted and reserved calibration groups.

MERGE and NEW retain each policy's recorded confidence semantics. Confidence thresholds, expected-cost variants and fitted cue-model thresholds use the same calibration set and cost grid. The TIER hedge wrapper reads saved possibly-same records. Its ABSTAIN interpretation is separate from the native component, which preserves NEW and the candidate relation. Unknown candidate labels during fitting use the same declared upper-cost objective across methods.

Table 33: Decision cost with merge-error cost 5, NEW-error cost 1 and rejection cost 0.25. Simulation uses 1036 labelable queries and physical-identity labels. 3RScan uses 715 held-out queries and designated-correspondence disagreement. Thresholds use each dataset's calibration set. Brackets bound unknown-candidate costs. All-abstain anchors the rejection cost. CUE+POS adds a position likelihood.

| | Simulation | | 3RScan | |
|---|---|---|---|---|
| Configuration | Cov. | Mean cost | Cov. | Mean cost |
| CG-style + threshold | 0.426 | [0.192, 0.197] | 0.450 | 0.158 |
| EVA-style + threshold | 0.379 | 0.247 | 0.919 | 1.087 |
| NN + threshold | 0.381 | 0.237 | 0.759 | 1.613 |
| Cosine–H + threshold | 0.270 | 0.293 | 0.825 | 1.718 |
| Single-exemplar + threshold | 0.114 | 0.246 | 0.305 | 0.971 |
| TIER proxy + threshold | 0.212 | 0.236 | 0.726 | 1.610 |
| PDA-MM + threshold | 0.409 | [0.187, 0.192] | 0.961 | 0.494 |
| PDA-DP + threshold | 0.000 | 0.250 | 0.069 | 0.409 |
| PDA-C + threshold | 0.437 | 0.283 | 0.669 | 0.670 |
| TIER component + threshold | 0.105 | 0.306 | 0.074 | 0.385 |
| TIER hedge | 0.812 | [0.796, 0.815] | 0.730 | 1.607 |
| CUE | 0.207 | [0.329, 0.338] | 0.733 | 0.695 |
| CUE (expected cost) | 0.031 | 0.242 | 0.509 | 0.324 |
| CUE+pos | 0.768 | [0.369, 0.383] | 0.934 | 0.247 |
| CUE+pos (expected cost) | 0.501 | [0.224, 0.234] | 0.811 | 0.128 |
| All abstain | 0.000 | 0.250 | 0.000 | 0.250 |

Table 34: Equal-integer-budget accounting. Cells contain merge disagreements / NEW disagreements / unknown-candidate merges, evaluated against physical identities in simulation and designated correspondences in 3RScan. The 20% budgets are 207 and 143 queries. The CUE ceilings are 214 and 524. Unavailable budgets are n/a.

| | Simulation | | 3RScan | |
|---|---|---|---|---|
| Policy | 20% | CUE ceiling | 20% | CUE ceiling |
| CG-style | 6/0/0 | 6/0/0 | 0/0/0 | 16/0/0 |
| EVA-style | 8/0/0 | 9/0/0 | 21/12/0 | 102/35/0 |
| NN | 8/0/0 | 8/0/0 | 44/0/0 | 210/0/0 |
| Cosine–H | 16/0/0 | 17/0/0 | 45/1/0 | 203/1/0 |
| Single-exemplar | 7/0/0 | 7/0/0 | 67/0/0 | 315/0/0 |
| TIER proxy | 8/0/0 | 8/0/0 | 48/0/0 | 220/6/0 |
| PDA-MM | 5/1/1 | 6/1/1 | 7/1/0 | 21/1/0 |
| PDA-DP | 29/0/1 | 29/0/1 | 76/1/0 | 167/1/0 |
| PDA-C | 7/24/0 | 8/24/0 | 17/32/0 | 48/234/0 |
| TIER hedge | 8/0/0 | 8/0/0 | 48/0/0 | n/a |
| TIER component | 64/0/1 | 70/0/1 | 62/0/0 | 112/259/0 |
| CUE | 26/0/1 | 27/0/2 | 0/4/0 | 87/14/0 |
| CUE+pos | 8/0/1 | 10/0/1 | 0/4/0 | 6/9/0 |

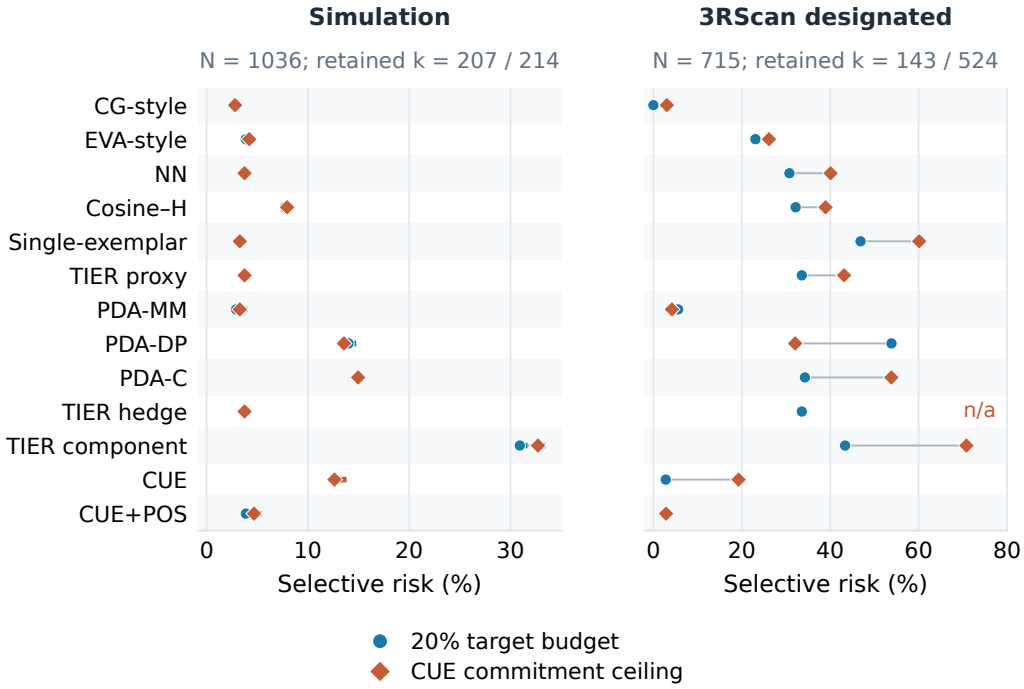


Figure 7: Matched-budget selective risk on the same labelable universe. Native rules stop at their attainable commitment count. Unknown candidate identities produce risk bounds.

At $c_{\text{abs}} = 0.10$, the simulated slice places the CUE expected-cost rule slightly below the best thresholded baseline, with both close to all-abstain. At the other reported levels, thresholded baselines attain lower slice cost. In real data, the CUE+POS expected-cost variant leads at the three larger rejection costs. Independent full-grid tests, paired differences and unknown-label bounds are retained in separate generated outputs.

Partial AURC uses a common attainable interval and a declared normalization. A no-reject scorer curve is a separate configuration. Boundary ties use fixed, ground-truth-independent ordering, with multiplicities recorded at each integer budget. Ranking can vary within a curve: CUE common-interval partial AURC is lower than CG-style matching on the simulated slice, although CG has fewer errors at the CUE endpoint. Other baselines achieve lower partial AURC than either.

### C.7 Real groups and exploratory cases

The full evaluation has 28 distinct reference scans and 715 queries. The lookalike-moved panel has 88 eligible observations from 27 pairs, each with a distinct reference scan. Metadata identifies 91 target entities before visibility filtering. Released mappings supply designated correspondence labels, while ambiguity containers define a retrieval diagnostic. Container sizes range from 2 to 22, with median 4. Query displacement has median 0.144 m and interquartile range 0.058–0.323 m. A higher-level site partition is unavailable, so resampling groups use the reference scan.

The loader forms a container from each released ambiguity entry's member endpoints. Two containers have disconnected annotation graphs, affecting seven panel queries. We retain this historical grouping and additionally score direct released pairs and graph connectivity. All 88 queries have a directly annotated alternative, and 64 have additional alternatives reachable only through a multi-edge path. No saved panel action selects a member found only through a disconnected container. Thus graph splitting leaves those action counts unchanged, while distinguishing direct and path-only support changes their interpretation. Container-uniform baselines remain descriptive sampling probabilities.

Table 35: Annotation-support partition on all 88 moved-lookalike queries. D denotes the designated record, Alt a directly annotated alternative, Path a multi-edge connection only, Ctr a disconnected container relation only, and U a merge without released relation support. Counts sum to 88 in every row. TIER component preserves NEW with its possibly-same edge. MASA-R50 component operates on two entity batches.

| Policy | D | Alt | Path | Ctr | U | NEW | Abs. |
|---|---|---|---|---|---|---|---|
| CG-style | 75 | 4 | 1 | 0 | 2 | 6 | 0 |
| EVA-style | 47 | 18 | 15 | 0 | 4 | 4 | 0 |
| NN | 35 | 26 | 21 | 0 | 6 | 0 | 0 |
| DINO Cosine–H | 29 | 26 | 21 | 0 | 10 | 2 | 0 |
| Single-exemplar | 18 | 26 | 21 | 0 | 15 | 8 | 0 |
| TIER proxy | 30 | 19 | 21 | 0 | 4 | 14 | 0 |
| PDA-MM | 75 | 9 | 1 | 0 | 3 | 0 | 0 |
| PDA-DP | 61 | 14 | 0 | 0 | 13 | 0 | 0 |
| PDA-C | 30 | 8 | 5 | 0 | 2 | 43 | 0 |
| CUE | 29 | 13 | 6 | 0 | 1 | 0 | 39 |
| CUE+POS | 67 | 6 | 0 | 0 | 1 | 0 | 14 |
| TIER hedge | 30 | 19 | 21 | 0 | 4 | 0 | 14 |
| TIER component | 13 | 18 | 14 | 0 | 2 | 41 | 0 |
| MASA-R50 component | 10 | 4 | 7 | 0 | 12 | 55 | 0 |
| R50 Cosine–H | 15 | 8 | 6 | 0 | 1 | 58 | 0 |

Table 36: Real-data event-probability differences from container-uniform baselines, with reference-scan cluster intervals. Values are probability differences. Designated agreement and selection of another container member have distinct baselines.

| Policy | Event | Difference bounds | Outer 95% CI |
|---|---|---|---|
| CG-style | Designated | $+0.599$ | $[+0.504, +0.684]$ |
| CG-style | Other member | $-0.690$ | $[-0.747, -0.633]$ |
| PDA-MM | Designated | $+0.599$ | $[+0.505, +0.682]$ |
| PDA-MM | Other member | $-0.633$ | $[-0.700, -0.571]$ |
| PDA-DP | Designated | $+0.440$ | $[+0.298, +0.554]$ |
| PDA-DP | Other member | $-0.588$ | $[-0.654, -0.497]$ |
| Cosine–H | Designated | $+0.076$ | $[-0.048, +0.218]$ |
| Cosine–H | Other member | $-0.213$ | $[-0.337, -0.095]$ |
| NN | Designated | $+0.145$ | $[+0.035, +0.254]$ |
| NN | Other member | $-0.213$ | $[-0.326, -0.108]$ |
| EVA-style | Designated | $+0.281$ | $[+0.153, +0.396]$ |
| EVA-style | Other member | $-0.372$ | $[-0.474, -0.256]$ |
| TIER proxy | Designated | $+0.088$ | $[-0.021, +0.197]$ |
| TIER proxy | Other member | $-0.292$ | $[-0.412, -0.179]$ |
| Single-exemplar | Designated | $-0.049$ | $[-0.145, +0.037]$ |
| Single-exemplar | Other member | $-0.213$ | $[-0.318, -0.109]$ |
| PDA-const | Designated | $+0.088$ | $[-0.051, +0.246]$ |
| PDA-const | Other member | $-0.599$ | $[-0.683, -0.516]$ |
| CUE | Designated | $+0.076$ | $[-0.025, +0.181]$ |
| CUE | Other member | $-0.531$ | $[-0.633, -0.421]$ |
| CUE+POS | Designated | $+0.508$ | $[+0.400, +0.602]$ |
| CUE+POS | Other member | $-0.679$ | $[-0.735, -0.622]$ |

Group-conditional summaries use each policy's own in-container predictions and remain in the accompanying results. The main comparisons retain the common query population in Table 35, including the MASA-R50 component's 55 NEW decisions on existing objects.

**Label sensitivity at fixed commitment budgets.** We rescore exactly the same retained query keys under designated-only and designated-plus-direct-pair agreement. All 15 rules reach the common budget of 143 out of 715 queries. NN disagreement risk changes from 30.8% to 21.7%. Single-exemplar risk changes from 46.9% to 36.4%. CG-style stays at zero and CUE stays at 2.8%, with all four of its disagreements being NEW decisions. This sensitivity isolates label interpretation while preserving actions, ranking, calibration and coverage.

Table 37: Real-data disagreement risk under two annotation rules at the same 143 retained queries per policy. Direct-pair tolerance accepts a designated match or a directly released alternative relation. This ID-level sensitivity does not evaluate the official pose-estimation task.

| Policy | $k$ | D disagreements | D risk (%) | Alt risk (%) |
|---|---|---|---|---|
| CG-style | 143 | 0 | 0.0 | 0.0 |
| EVA-style | 143 | 33 | 23.1 | 18.2 |
| NN | 143 | 44 | 30.8 | 21.7 |
| DINO Cosine–H | 143 | 46 | 32.2 | 24.5 |
| Single-exemplar | 143 | 67 | 46.9 | 36.4 |
| TIER proxy | 143 | 48 | 33.6 | 23.1 |
| PDA-MM | 143 | 8 | 5.6 | 5.6 |
| PDA-DP | 143 | 77 | 53.8 | 49.0 |
| PDA-C | 143 | 49 | 34.3 | 31.5 |
| CUE | 143 | 4 | 2.8 | 2.8 |
| CUE+POS | 143 | 4 | 2.8 | 2.8 |
| TIER hedge | 143 | 48 | 33.6 | 23.1 |
| TIER component | 143 | 62 | 43.4 | 30.1 |
| MASA-R50 component | 143 | 77 | 53.8 | 51.0 |
| R50 Cosine–H | 143 | 41 | 28.7 | 23.8 |

Table 38 lists all eight nearest-other entities across six pairs. Each has another annotated container member closer than its designated predecessor. This geometry defines an exploratory subset of natural changes. The dataset does not specify a two-object exchange intervention. Distances use annotated centroids in the shared reference frame, separately from the released rigid-motion displacement above. The case archive retains all 15 policies' decisions. Table 35 gives their complete panel-level annotation partition.

Figures 8 and 9 retain all eight cases. For each query, designated predecessor and nearest other member, frame selection maximizes depth-consistent projected-point support, then normalized box area, with frame-ID tie breaking. Main-case selection maximizes the smallest normalized box area across its three chosen observations, with case-letter tie breaking. This rule selects C before loading policy outcomes. Existing boxes are checked against the annotated 3D points and saved depth. Crops include 12% padding, use a uniform clockwise $90^\circ$ rotation and preserve aspect ratio. The source manifest records every candidate frame, selected crop, transformation and all 15 policies' outputs. The displayed frames illustrate each entity. Policies use their recorded multi-frame representations.

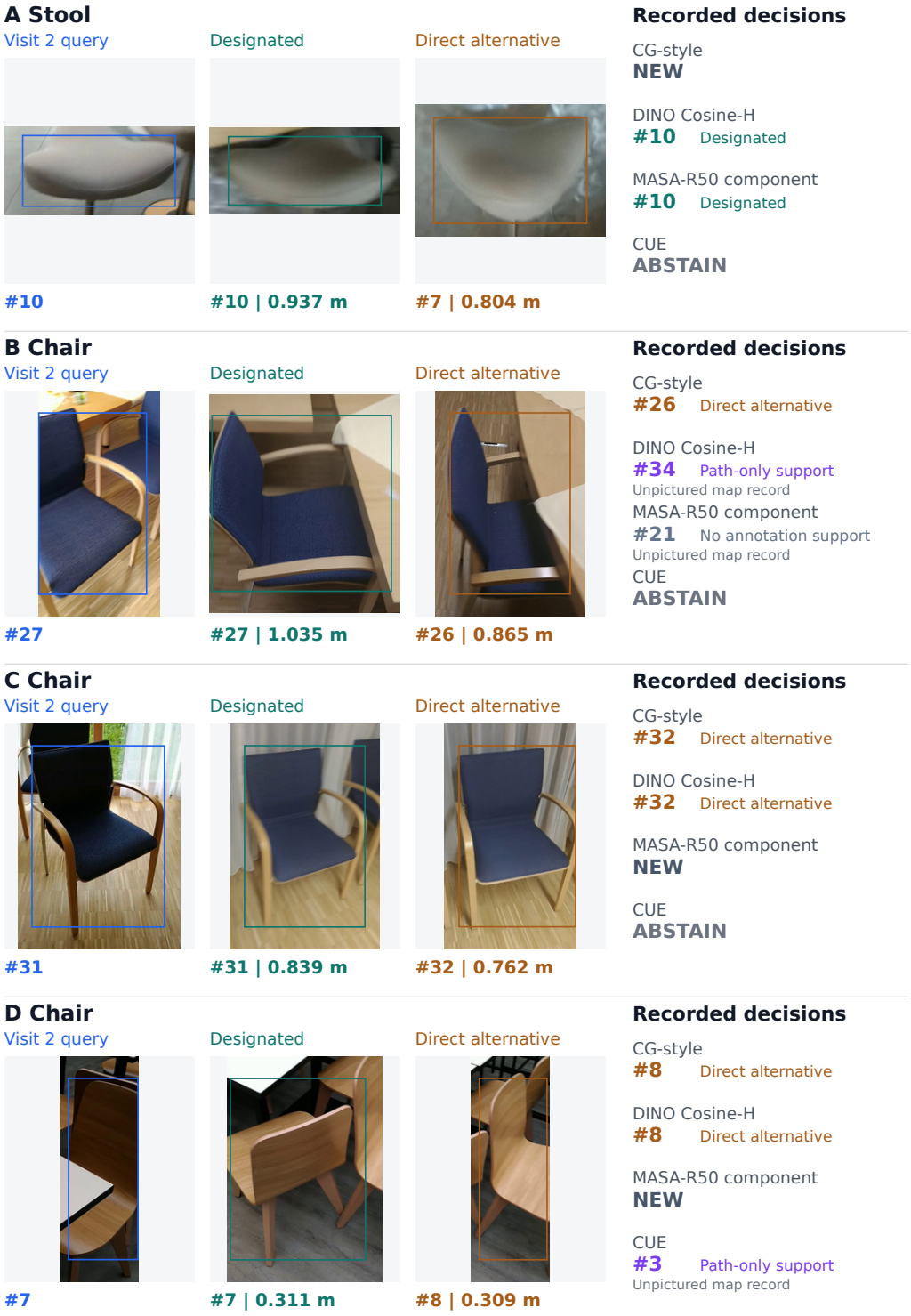


Figure 8: **Natural nearest-other cases A–D.** Columns show the revisit query, its annotated predecessor and the nearest other group member. Boxes mark verified annotation bounds. Distances are query-to-predecessor centroid distances in the shared frame. The saved decisions retain selected object IDs, NEW and ABSTAIN. An unpictured member or candidate is distinct from the other object shown. All 15 policies remain in the case archive and the population-level table above.

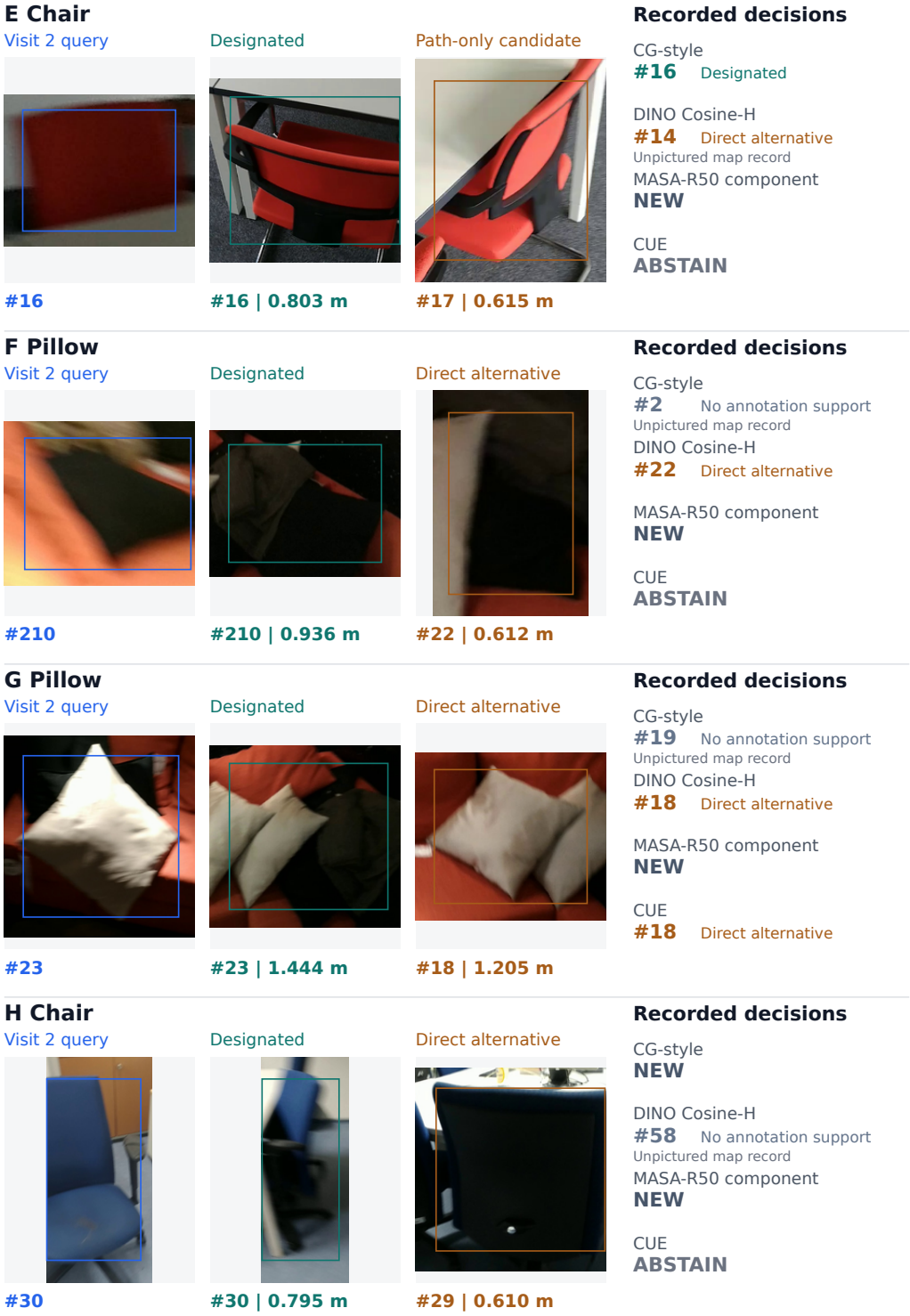


Figure 9: **Natural nearest-other cases E–H.** Same selection and display protocol as Figure 8. The series retains designated and alternative links, other selections, NEW and unresolved decisions. Original blur and partial visibility remain visible. Case selection supplies qualitative context. Population-level rates use the complete evaluation sets.

Table 38: All eight exploratory nearest-other cases. Reference labels are UUID prefixes. Query and Other identify annotated objects in the rescan and reference respectively. $d_{\mathrm{des}}$ is distance to the designated predecessor and $d_{\mathrm{other}}$ to the nearest other container member. Every listed case satisfies $d_{\mathrm{other}} < d_{\mathrm{des}}$. Full UUIDs, correspondences and distances are retained in the generated case CSV.

| Case | Reference | Query id | $K$ | Other id | $d_{\mathrm{des}}$ (m) | $d_{\mathrm{other}}$ (m) |
|---|---|---|---|---|---|---|
| A | 1d234004 | 10 | 4 | 7 | 0.937 | 0.804 |
| B | 2e4a3964 | 27 | 10 | 26 | 1.035 | 0.865 |
| C | 2e4a3964 | 31 | 10 | 32 | 0.839 | 0.762 |
| D | 355465d0 | 7 | 7 | 8 | 0.311 | 0.309 |
| E | 569d8f13 | 16 | 7 | 17 | 0.803 | 0.615 |
| F | 75c25973 | 210 | 2 | 22 | 0.936 | 0.612 |
| G | 75c25973 | 23 | 2 | 18 | 1.444 | 1.205 |
| H | d7d40d62 | 30 | 8 | 29 | 0.795 | 0.610 |

### C.8 Coordinate and information-path controls

**Implementation audit.** The CUE implementation exactly reproduces all 584 saved cache episodes before intervention. Changing only the second visit's coordinate frame changes 50,944 of 55,189 posteriors and 384 actions (Table 39). Candidate sets and raw cues remain unchanged to the stated tolerance. The NEW prior directly compares query and map centroids when computing local density, so the complete CUE implementation requires comparable cross-visit coordinates. A common rigid transform applied to both visits changes no posterior or action at this tolerance. Descriptor-only similarity transforms pass with a maximum numerical error of $1.58 \times 10^{-12}$ on the valid cached geometry. Randomized and cleared evaluation metadata, and opaque entity renaming with order shuffling, produce no substantive change for these fixed inputs. This isolation check begins after perception; it does not certify the generator or visibility-filtering stages. All changed actions in the coordinate intervention occur outside the baseline top-score tie tolerance.

| Cache | Episodes | Queries | Δ posterior | Δ action |
|---|---|---|---|---|
| Simulation: clean and tint | 196 | 19,409 | 17,933 | 101 |
| Simulation: cached 0.05 m | 118 | 11,651 | 10,756 | 81 |
| Simulation: cached 0.15 m | 118 | 11,668 | 10,741 | 96 |
| Simulation: cached 0.50 m | 118 | 11,670 | 10,755 | 93 |
| 3RScan: all cached pairs | 34 | 791 | 759 | 13 |
| Execution total | 584 | 55,189 | 50,944 | 384 |

Table 39: Fixed-input coordinate-frame audit. A seeded proper rotation and 10 m translation are applied to visit 2. Embeddings, entity membership and calibration remain fixed. Counts report changes at tolerance $10^{-9}$. Real-data counts include calibration pairs; repeated cache variants are execution coverage, not independent replicates.

### C.9 Independent representation and component controls

The MASA-R50 control changes the representation on the same cached detector boxes and physical-query populations. Table 40 separates the simulated specified-partner event from the real other-member event. R50 cosine–Hungarian retains the original multi-prototype assignment rule. The source entity-batch component uses its single-exemplar scoring and assignment protocol from Appendix B.6. It produces no confirmed correct merge on the simulated panel, with one unknown candidate accounting for the correctness upper bound. In 3RScan it makes 10 designated matches and 55 NEW decisions on the 88 existing-object queries.

Table 40: Independent representation on the original diagnostic panels. Brackets bound unknown candidate identities. The four event columns retain their distinct definitions and populations. Complete action partitions and all original policies remain in the source tables.

| | Simulation ($N = 106$) | | 3RScan ($N = 88$) | |
|---|---|---|---|---|
| Policy | Correct | Specific partner | Designated | Other member |
| CG-style | [0.019,0.038] | [0.906,0.925] | 0.852 | 0.057 |
| DINO Cosine–H | [0.019,0.047] | [0.811,0.840] | 0.330 | 0.534 |
| MASA-R50 entity-batch component | [0.000,0.009] | [0.472,0.481] | 0.114 | 0.125 |
| R50 Cosine–H | 0.000 | 0.802 | 0.170 | 0.159 |
| CUE | 0.000 | 0.226 | 0.330 | 0.216 |
| CUE+pos | 0.009 | 0.774 | 0.761 | 0.068 |

Ranking and calibration transfer vary by dataset (Table 41). The source component makes 4 confirmed wrong merges and 1 unknown-candidate merge among 207 simulated commitments, versus 77 merges disagreeing with the designated label among 143 real commitments. After real-data calibration, it retains three of 715 queries at cost 0.256. R50 cosine–Hungarian retains nine at 0.247. Thresholded CG-style and the CUE+POS expected-cost rule have lower real-data costs. Source tables include paired intervals, common-coverage partial AURC and held-out full-grid results.

Table 41: Independent representation with component/proxy decisions at their original thresholds and calibration-fitted rejection, at costs $(5, 1, 0.25)$. Simulation uses the 1036-query ambiguity slice; 3RScan uses 715 held-out queries. Coverage is measured on those fixed populations. Brackets bound unknown candidate costs.

| | Simulation slice | | 3RScan full test | |
|---|---|---|---|---|
| Policy | Coverage | Cost | Coverage | Cost |
| CG-style + threshold | 0.426 | [0.192,0.197] | 0.450 | 0.158 |
| DINO Cosine–H + threshold | 0.270 | 0.293 | 0.825 | 1.718 |
| MASA-R50 entity-batch component | 1.000 | [0.986,1.005] | 1.000 | 1.597 |
| MASA-R50 entity-batch component + threshold | 0.104 | 0.238 | 0.004 | 0.256 |
| R50 Cosine–H | 1.000 | [0.506,0.520] | 1.000 | 1.032 |
| R50 Cosine–H + threshold | 0.181 | [0.248,0.253] | 0.013 | 0.247 |
| CUE | 0.207 | [0.329,0.338] | 0.733 | 0.695 |
| CUE expected cost | 0.031 | 0.242 | 0.509 | 0.324 |
| CUE+pos expected cost | 0.501 | [0.224,0.234] | 0.811 | 0.128 |
| All abstain | 0.000 | 0.250 | 0.000 | 0.250 |