LIMBO: DYNAMIC MAP EVALUATION UNDER AMBIGUOUS EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic object maps must decide whether a detection belongs to a moved object, a replacement, a new arrival, or an object that has disappeared. Benchmarks usually score these decisions against one annotated history, although the same observations may admit several physical histories. We introduce LIMBO, an evaluation protocol built around sensor-matched compatible histories. It distinguishes an observed realization from a persistent entity and credits only claims shared by every history admitted under a declared evidence contract. In 355 controlled ProcTHOR groups, HOTA changes its preferred identity decision between hidden relocation and replacement even though the contracted mapper input is unchanged; 149 groups are also byte-identical in RGB-D. LIMBO preserves their common location and leaves lineage unresolved. In released 3RScan data, 5.92% of persistent targets in the sealed test split and 5.77% in a separate validation cohort have multiple annotation-compatible owners. The primary category-blind utility analysis finds no universal policy winner, whereas a separate audit of two released mappers exposes a coverage–precision trade-off. In ten fresh held-out simulated scenes, an online lifecycle mapper obtains 60.0% relocation retention, removal F1, and addition F1 from anonymous oracle regions. Simple RGB-D connected components reduce these scores to 20.0%, 10.7%, and 10.0%, showing strong sensitivity to proposal quality. LIMBO is intended for settings where unsupported certainty is costly. Realized-world and probabilistic scores remain appropriate for other decision objectives.

1 INTRODUCTION

Robots use object maps to search, manipulate, and reason about changing indoor scenes. The map must preserve an object's identity when it moves, add a new entity when an object appears, and remove an entity when it leaves. These operations are usually evaluated against an annotated sequence that supplies one correspondence between observations. The resulting score mixes two questions: did the system locate the visible object, and did the available measurements identify its physical history?

The second question is sometimes undecidable from the input. Suppose a mug is hidden, and a mug-shaped object appears at a new location. The same endpoint depth and mask can arise from moving the original mug or from removing it and placing a visually identical mug. A mapper that receives only those endpoints must produce the same output in both cases. A benchmark that selects one history as truth therefore penalizes a defensible answer and may reward a confident identity claim unsupported by the input. Unlike detector noise or tracker failure, this ambiguity cannot be removed by a better inference rule unless the observation interface also changes.

We address this evaluation problem with LIMBO. An *evidence contract* declares what a mapper receives. A finite construction grammar enumerates the physical histories consistent with that contract. The evaluator keeps only typed claims true in every compatible history. A location claim can therefore be correct while an entity claim is unresolved. Stronger evidence, such as a continuous transport observation or a verified robot action, can shrink the compatible set and make additional claims decidable (Figure 1).

We ask whether this distinction changes what dynamic-map benchmarks measure, whether it appears in natural repeated scans, and whether the interface remains useful with online RGB-D observations. The paper contributes the following.

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.

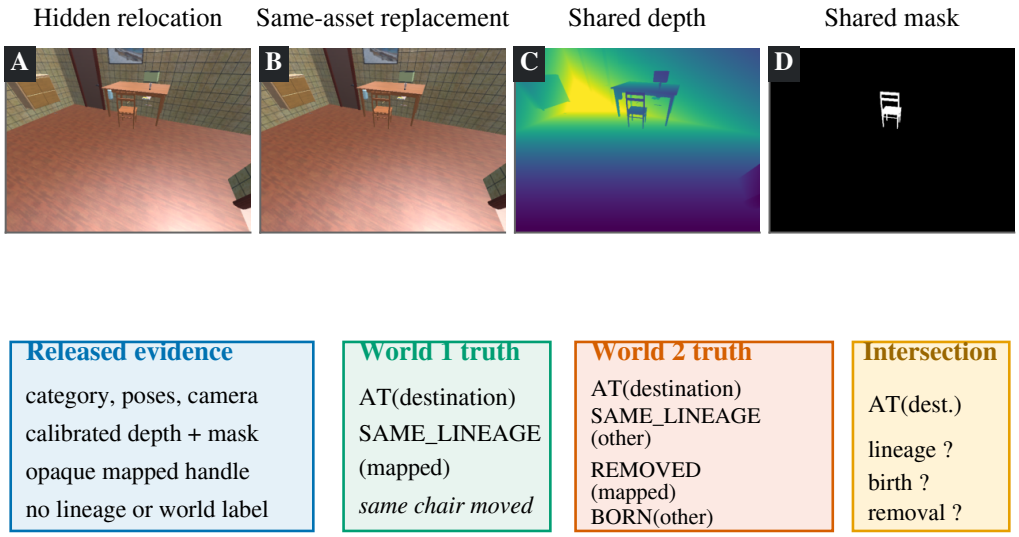

Figure 1: **One input, two incompatible physical histories.** A Chair pair from the held-out benchmark renders hidden relocation (A) and same-asset replacement (B). Their diagnostic RGB images are byte-identical, although RGB is not part of this evidence contract. The released depth (C) and mask (D) are also byte-identical. Public fields expose geometry and opaque handles; only the scorer sees the distinct world truths. Their intersection supports the destination realization, but not lineage, birth, or removal.

- A typed evaluation protocol that separates spatial realizations from persistent entities and tests each claim across a finite compatibility group.
- A leakage-audited ProcTHOR benchmark and evidence ladder that separate non-identifiability, missing observations, and inference error.
- Natural-data and released-system audits, followed by an online RGB-D experiment that measures sensitivity to non-oracle region proposals.

LIMBO is a measurement protocol rather than a new tracker. Its guarantees are relative to a versioned finite grammar, which defines the test universe but does not enumerate every physically possible history.

2 RELATED WORK

Dynamic object maps and uncertain association. Multiple-hypothesis semantic mapping delays uncertain data association (Bernreiter et al., 2019); Poisson multi-Bernoulli mixture and trajectory formulations preserve distributions over births, deaths, and associations (García-Fernández et al., 2018; Granström et al., 2018). Khronos represents metric-semantic change over time (Schmid et al., 2024). GLASSES maintains active and missing object libraries for later open-vocabulary search (Bogenberger et al., 2026). ConceptGraphs builds open-vocabulary 3D scene graphs for perception and planning (Gu et al., 2024), while SuperMap represents temporal instance relations for visual-language navigation (Zhao et al., 2026). Object state estimators and scene-graph memory address partial observability in household tasks (Du et al., 2022; Kurenkov et al., 2023). These methods estimate uncertain world state. LIMBO instead asks which semantic lifecycle claims their released outputs support under a specified observation interface.

Changing-scene benchmarks. 3RScan provides repeated RGB-D scans and instance relations for re-localization studies (Wald et al., 2019a); RIO evaluates instance re-localization across changed indoor scans (Wald et al., 2019b), and Living Scenes evaluates multi-object relocalization and reconstruction (Zhu et al., 2024). DynamicTHOR supplies human-centric object motion (Wang et al.,

Table 1: The uncertainty represented by related evaluation frameworks. Representative sources are cited in the surrounding text (Luiten et al., 2021; Xia et al., 2025a; Geifman & El-Yaniv, 2019; Libkin, 2011).

Evaluator / method	Uncertainty source	Reference object	Scored output	Guarantee
HOTA / IDF1	none in reference	one trajectory	hard tracks	label agreement
pGOSPA / pTGOSPA	predictor	one trajectory/set	distribution	proper realized-world score
SelectiveNet	predictor confidence	one label	answer/reject	risk–coverage
Multi-reference reduction	reference ambiguity	history set	hard tracks	mean/best/worst
Certain answers	incomplete reference	possible databases	query answer	true in every database
LIMBO	evidence + grammar	physical histories	typed map claims	true in every history

2026); ProcTHOR and AI2-THOR provide the procedural houses and interactive simulator used here (Deitke et al., 2022; Kolve et al., 2017). These resources annotate or generate a physical history. LIMBO's distinct evaluation unit is a set of histories deliberately matched under an evidence contract.

Metrics under uncertainty and incomplete information. CLEAR MOT, IDF1, and HOTA compare detections and associations with labelled trajectories (Bernardin & Stiefelhagen, 2008; Ristani et al., 2016; Luiten et al., 2021); GOSPA and trajectory-set metrics expose localization and cardinality error (Rahmathullah et al., 2017; Garcia-Fernandez et al., 2020). Probabilistic extensions score uncertain multi-object estimates (Xia et al., 2025a;b), and ambiguity in tracker evaluation has long been recognized (Drummond & Fridling, 1992). Possible-world semantics and certain answers formalize database queries under incomplete information (Lipski, 1981; Imieliński & Lipski, 1984; Libkin, 2011); selective prediction exposes risk–coverage trade-offs (Geifman & El-Yaniv, 2017; 2019).

Table 1 places LIMBO beside these approaches. PMBM and trajectory PMBM retain association hypotheses, pGOSPA and pTGOSPA score a distribution against one realized reference, and selective prediction decides when to answer a label. A set of references still needs a mean, best, or worst reduction, each of which chooses a different policy toward ambiguity. Certain answers use an intersection over incomplete databases. LIMBO instantiates this semantics for dynamic maps by defining sensor-matched histories, typed lifecycle claims, and a public mapper interface. A location atom may be supported even when all lineage atoms are unresolved. A probabilistic or hypothesis-set mapper can be adapted by emitting an atom only when every retained hypothesis entails it. This certain-answer objective is useful when unsupported commitments are costly. It complements, rather than replaces, proper probabilistic scores for calibrated predictive distributions.

3 EVIDENCE-RELATIVE DYNAMIC MAPS

3.1 ENTITIES, REALIZATIONS, AND CLAIMS

An *entity* denotes a persistent physical lineage. A *realization* is an observation-scoped spatial token or mapped footprint. This factorization lets a map report where an object is without asserting which physical lineage produced that occurrence. We use the finite claim language

$$\mathcal{C} = \{\textsc{At}(r, p), \textsc{Absent_At}(r, p), \textsc{Same_Lineage}(r, e), \textsc{Born}(e), \textsc{Removed}(e)\}. \tag{1}$$

The claims state, respectively, that realization r occurs at pose p, the mapped footprint r is unoccupied at p, r belongs to entity e, e entered the map domain, or e left it. Thus ABSENT_AT concerns a realization, not permanent entity nonexistence. Poses are quantized to 10 cm translation and 10° yaw for symbolic matching; continuous localization error is scored separately.

Fix a versioned finite benchmark grammar G. An evidence contract E induces a nonempty set $\mathcal{W}_G(E)$ containing every manifest-listed history whose public projection matches E and survives the declared attrition rules. This set is complete only relative to G. A *compatibility group* is the set of histories constructed for one public endpoint record under one evidence contract. Attrition rules are the prespecified rendering and integrity checks that determine whether a candidate history enters the scored group. Each group declares a finite ground-atom universe $\mathcal{C}_g$; $T(w) \subseteq \mathcal{C}_g$ is its complete positive truth set, with unlisted atoms false. Every current realization has exactly one explicit SAME_LINEAGE owner in every world. The mapper receives an opaque handle for the mapped entity,

but never the scorer's raw physical lineage or counterfactual entity symbol. Benchmark-supported claims are

$$\Gamma_G(E) = \bigcap_{w \in \mathcal{W}_G(E)} T(w). \tag{2}$$

Tracker-minted IDs and scorer-only IDs occupy different namespaces. Let ρ_g fix separate roles for public handles such as the mapped and explicitly imported entities and canonically rename reference-side anonymous entities. For prediction P, the scorer chooses the injective partial map π from prediction-side anonymous entities to reference-side anonymous roles that maximizes $|\pi(P) \cap \rho_g(\Gamma_G(E))|$; unmatched predictions receive distinct fresh roles. This is identifier alignment, analogous to matching unlabelled track IDs; it cannot alter a predicate or public anchor. The mapping preserves equality and distinction across SAME_LINEAGE, BORN, and REMOVED, and leaves spatial claims unchanged. A partial $\textsc{Born}(x)$ may match one of several reference-side anonymous entities, but one x cannot match two entities or the public imported handle. Alignment is local to a compatibility group and does not impose tracker-ID continuity across groups from the same house.

Before claim or query scoring, the interface rejects malformed predicate arguments and jointly inconsistent outputs. The interface rejects five classes of malformed output: multiple owners or endpoint poses; coincident AT and ABSENT_AT claims; birth of an initial entity; assignment of a current realization to a removed entity; and unequal current poses for one entity. An ordinary track ID compiles to SAME_LINEAGE, whereas physical BORN and REMOVED require explicit lifecycle output; a new track ID is not silently treated as physical birth. For $\bar{P} = \pi(P)$ and $\bar{\Gamma} = \rho_g(\Gamma_G(E))$, supported claims are $\bar{P} \cap \bar{\Gamma}$, overreach is $\bar{P} \setminus \bar{\Gamma}$, and omitted support is $\bar{\Gamma} \setminus \bar{P}$. We define evidential soundness and completeness as

$$S(P, E) = \frac{|\bar{P} \cap \bar{\Gamma}|}{|\bar{P}|}, \qquad C(P, E) = \frac{|\bar{P} \cap \bar{\Gamma}|}{|\bar{\Gamma}|}, \tag{3}$$

and group-level evidential F1 as their harmonic mean. We set soundness to 1 when P is empty and completeness to 0 when $\Gamma_G(E)$ is empty. Primary scores average group-level quantities within each house and then average houses. Consequently, displayed macro soundness, completeness, and F1 are separate averages; F1 need not equal the harmonic mean of the other two displayed columns. We also report the alternative harmonic mean computed after aggregation. Houses are the independent experimental units. Evidential F1 is a score within this fixed claim schema; changing predicate granularity or weights defines a different benchmark version.

3.2 THE HARD-DECISION CEILING

Let w_m be hidden relocation and w_r identical replacement. Both are admitted by G and expose the same E, but $\textsc{Same_Lineage}(r, e) \in T(w_m)$ whereas $\textsc{Same_Lineage}(r, n), \textsc{Born}(n), \textsc{Removed}(e) \in T(w_r)$. An evidence-only deterministic method has the same output in both worlds. If a benchmark samples one balanced latent world, either forced decision obtains expected accuracy at most $1/2$. Randomization cannot improve this bound: conditioning on each internal random seed yields a deterministic rule with accuracy at most $1/2$, so averaging over seeds does too. Eq.(2) instead marks the disputed lifecycle claims unresolved while retaining $\textsc{At}(r, p)$. More generally, if valid evidence refines the compatible set so $\emptyset \neq \mathcal{W}_G(E') \subseteq \mathcal{W}_G(E)$, then $\Gamma_G(E) \subseteq \Gamma_G(E')$. Stronger evidence can resolve claims but cannot invalidate a previously entailed claim under a fixed language. Formal statements and boundary conditions appear in Appendix A.

3.3 AN EVIDENCE LADDER FOR LIFECYCLE CLAIMS

The benchmark implements five contracts. *Endpoint geometry* releases camera calibration, category, pose, depth, and a binary mask. Paired hidden relocation and replacement then entail location but not lineage. *Endpoint plus free space* projects the old surface into the new depth image; a clear ray establishes local absence, never permanent removal. *Passive continuous transport* releases an ordered 8–24 observation tube whose grammar excludes an unobserved swap, supporting lineage. *Own action* releases a verified pickup–transport–place record bound to an opaque handle, also supporting lineage. *Explicit import/export* releases trusted lifecycle events supporting birth and removal. This contract provides an upper bound on what ordinary perception can establish.

Table 2: Benchmark cohorts. Endpoint groups also have a corresponding explicit-event group, whose repeated count is omitted.

Split	Houses	Groups	Classes	Endpoint	Rays	Action	Transport
Discovery	48	755	17	173	84	156	169
Selection	20	321	16	71	42	60	77
Held-out	32	498	14	111	62	102	112

This ladder distinguishes three failure sources that a single identity label conflates. The evidence may be physically insufficient, an informative view may be missing, or the mapper may fail to use an available observation. Under the declared contract, LIMBO counts only the third case as inference error.

4 BENCHMARK AND EVALUATION PROTOCOL

4.1 PHYSICAL CONSTRUCTION AND LEAKAGE CONTROLS

We generate fresh suites from 100 mutually disjoint ProcTHOR source records. Each source key contains the dataset name, immutable revision, official split, and source index. Targets follow a fixed hash order independent of method scores. Each suite attempts stationary, hidden relocation, same-asset replacement, removal, moved-out, occlusion, birth, reappearance, passive transport, and own-action histories. When no native visual twin exists, a distinct simulator lineage using the same asset is inserted before rendering; twins never coexist at the same collision pose.

Only identical public projections may enter an exact compatibility group. Depth equality is bytewise and binary masks require exact array equality. Raw RGB is excluded from the equivalence relation and audited separately. Public records contain opaque handles and content-addressed arrays, while scorer records contain lineages and truth. Prediction runs without scorer paths. Schema checks reject simulator IDs, asset IDs, source indices, world labels, lineage tables, and colour maps. Replays are invariant to input order and private-label permutations. Fixed integrity tests reject forged events, wrong poses, circular witnesses, missing worlds, duplicate handles, and hash mismatches.

The resulting benchmark contains 1574 groups from 100 source-disjoint houses (Table 2). Discovery, selection, and held-out contain 173, 71, and 111 byte-exact endpoint pairs. Frozen source-shard records retain 1,154 attempted targets and per-intervention failure reasons; all 24 merge/recovery links resolve uniquely in the included hash-chain audit. The paper names map to the artifact's train-plus-development, validation, and test partitions, respectively. Held-out construction and scoring occurred only after the adapter-pair selection criterion was fixed.

4.2 OUTPUTS, BASELINES, AND METRICS

We evaluate 16 public-only output configurations. Controls include hard reuse, hard replacement, abstention, spatial-only matching, and spatial matching with a time-to-live rule. Three mechanism-level adapters represent active/missing libraries (GLASSES-style), projected-depth and temporal-instance behavior (SuperMap-style), and frame-to-frame assignment (track continuity). They are behavioral adapters under one interface and do not reproduce the accuracy of the published systems. A compact track-oriented MHT branches over existing, new, and false associations, enforces global exclusion, and supports both hard output and the intersection over retained hypotheses (Bernreiter et al., 2019).

For every stateful backend, an *evidence-scoped reporting* variant leaves association and spatial state unchanged but restricts emitted lifecycle claims to those admitted by the contract. This reporting ablation tests whether withholding unsupported claims sacrifices current-location utility. Contract metadata is intentionally public because the evaluand is claim support under that contract. Soundness and completeness are reported jointly, so always abstaining cannot win by learning only the contract.

Alongside Eqs. (3), we report one post-commit fixed-projection and world-balanced IDF1/HOTA, state-set GOSPA ($p = 2, c = 1\,\text{m}, \alpha = 2$), lifecycle errors, and 1,000 deterministic single-world projections. The fixed projection uses a salt generated after every method output is frozen and hash-

committed in the recorded experiment pipeline. This ordering is author-attested rather than externally timestamped, so the value is an offline sensitivity diagnostic, not a prediction-time public leaderboard target. Revealing the salt in advance lets a projection oracle recover every selected endpoint label and obtain HOTA 1.0. World-balanced and per-world values do not depend on that salt. GOSPA minimizes the capped p-power matching objective before decomposing localization, misses, and false detections. Our HOTA code is protocol-frozen for these forced category/location matches and is not presented as a general HOTA replacement. Six subject-specific three-valued query types ask for current realization and mapped-entity location, whether a realization belongs to or is a moved occurrence of a specified mapped entity, whether that realization's owner (or the explicit imported entity) entered the domain, and whether the mapped entity left it. A reference answer is true or false only when all compatible worlds agree; otherwise it is unresolved. The prediction compiler uses the same subjects and never transfers a birth or removal claim from an unrelated entity. Boolean location answerability and continuous pose error are reported as separate tasks; a wrong pose is not silently converted into an unsupported Boolean answer. An unsupported answer is a non-unresolved prediction that disagrees with this reference. Appendix B.1 gives the complete truth conditions.

4.3 Natural repeated scans and official systems

The natural study uses 3RScan's released reference/rescan instance relations and grouped ambiguity annotations as *measurement-compatible associations*, not as exhaustive physical counterfactual worlds (Wald et al., 2019a;b). A UUID-hash split of all 385 train environments fixes 64/64/96/161 environments for fit/development/calibration/sealed test before train meshes are acquired. The sealed test split is held out from the 3RScan train release. A single 14-dimensional XYZ-only logistic scorer consumes oracle instance segments, known poses, and scene transforms. Semantic category is excluded from this scorer. Target-pooled finite-sample nonconformity quantiles produce candidate edges, and maximum matchings define possible and certain owners. The 90/95% labels name target-level calibration quantiles; because targets are clustered within environments, we claim no scene-equal population coverage guarantee. The scorer is a controlled measurement instrument, not a deployable association method.

The category-blind interface defines the primary natural utility analysis. An additional sensitivity analysis applies a category-equality gate after the sealed labels had been inspected. It evaluates six policies, including always unresolved, local singleton, and development-tuned confidence rejection. Outputs are partial injections, and risk extrema use one coherent official maximum matching per pair. Truth-independent predicted maps move asserted owners to current centroids and otherwise retain named stale reference locations. Strict floor budgets query unresolved targets. We therefore treat the gated result as descriptive; Appendix K specifies costs, map construction, and controls.

Separately, we run the authors' released GLASSES and ConceptGraphs (Bogenberger et al., 2026; Gu et al., 2024) posed-RGB-D components on the identical 47-environment, 110-pair validation cohort, with their released models, vocabularies, thresholds, and native frame strides. Labelled meshes and change metadata remain scorer-only. A frozen endpoint matcher measures single-owner correctness, membership in the official allowed owner set, and lineage supported in every maximum matching of the official allowed-owner graph. Empty-output environments remain in all scene-equal aggregates; low native coverage cannot be repaired after seeing the annotations. Appendix K gives the complete protocols and system boundaries.

All reported intervals use 10,000 environment-bootstrap draws and quantify finite-corpus resampling, not population generalization. For ProcTHOR, discovery combines 32 train houses with 16 development houses; the 20-house selection cohort chose the reported pair before the 32-house held-out cohort was opened. Discovery and selection are not independent. The final Limbo framing and several semantic corrections were made after an earlier prespecified benchmark gate failed. Appendix L identifies the affected quantities and separates the original sealed analyses from later descriptive reanalyses.

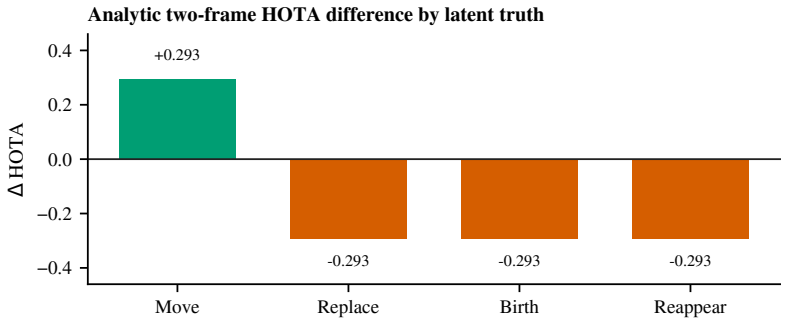

Figure 2: **Single-world HOTA depends on the latent history.** Bars are analytic values for the exact two-frame, one-object construction, repeated by all 111 held-out endpoint groups rather than sampled group estimates. Per-world HOTA favors reuse for hidden relocation and a new ID for replacement, birth, and reappearance; endpoint LIMBO F1 ties in all four cases because the contracted endpoint evidence is unchanged. Positive values favor reuse.

5 RESULTS

5.1 SINGLE-WORLD SCORES DEPEND ON THE LATENT HISTORY

The controlled benchmark gives the central diagnostic. Each of the 355 endpoint groups contains multiple semantic histories with matched public evidence. In the two-frame, one-object case, HOTA gives reuse minus new ID exactly $1-1/\sqrt{2} = +0.293$ under hidden relocation and -0.293 under the alternative history. Both hard outputs have identical endpoint LIMBO soundness (0.5), completeness (1.0), and F1 (0.667), because both predict the supported location and one unsupported identity claim. The metric difference comes entirely from the latent label.

Reducing the compatible references to a mean, best, or worst label changes the decision rule but does not recover claim-level support. The mean introduces a world prior, best accepts one compatible answer, and worst penalizes every hard answer. A 48-configuration sweep over one to eight frames and zero to four distractors produces 24 effects spanning $[-0.293, +0.293]$. LIMBO instead retains the supported location and rejects both unsupported identity claims; full reductions appear in Appendix F.

An aggregate across unequal contract supports is unsuitable for method ranking. Appendix E reports this descriptive audit: the apparent aggregate LIMBO difference comes from free-space groups where HOTA is undefined, while the HOTA difference comes from endpoint groups where LIMBO ties. On the contracts shared by both metrics, LIMBO differences are zero.

5.2 ADDITIONAL EVIDENCE CAN RESOLVE LINEAGE

Although RGB was excluded from construction, a byte audit finds 149 endpoint groups across 77 houses that also match exactly in RGB (109 inserted and 40 native alternatives); all already match in depth, mask, category, calibration, and camera pose. Any deterministic frozen RGB-D feature extractor therefore receives identical endpoint input on this subset. For the same objects, 66 held-out passive-transport and 102 own-action records contract four admitted histories to one and add lineage support, changing the conforming compiler's F1 from 0.667 to 1.000. Adding an unobserved swap or nonbinding-action history removes that support in every pair. These are grammar-conformance tests, not measurements of real action-binding reliability (Appendix F). Across 325 held-out groups, adding appearance-compatible replacement or delayed observation lowers compiler soundness from 100% to 81.1% and 87.8%, respectively. This sensitivity is why support must always be interpreted relative to the declared grammar.

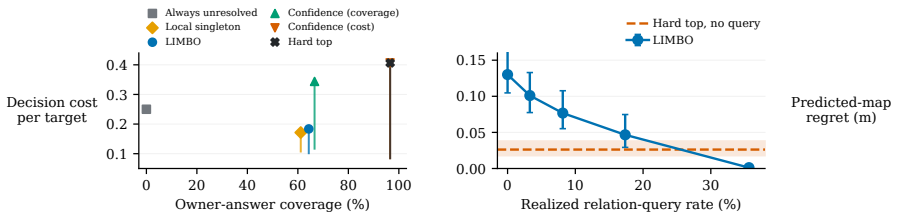

Figure 3: **Natural-data utility under the retrospective category gate.** The category-blind interface is primary; scene-equal gated values span 161 sealed environments. Left, segments connect coherent optimistic and worst-case costs. Right, LIMBO map regret under strict query budgets; bars and the hard-top band are 95% environment-bootstrap intervals. These are frontier projections, not a universal-dominance claim.

5.3 Online RGB-D is sensitive to proposal quality

The online extension tests whether the interface remains useful when proposals come from RGB-D rather than scorer masks. On ten fresh held-out simulated scenes, anonymous oracle regions give 60.0% relocation retention, 60.0% removal F1, and 60.0% addition F1 at 4.83 ms per frame. Deterministic RGB-D connected components have 63.0% proposal recall and reduce the three lifecycle scores to 20.0%, 10.7%, and 10.0%, while latency rises to 94.64 ms. Thus naive connected components introduce a severe additional failure source. The experiment uses known camera poses and does not include learned detection or embodied control. The runs used AI2-THOR 5.0.0 CloudRendering on two RTX 3090 GPUs; the mapper environment used Python 3.10 with the pinned packages in `requirements-rgbd.txt` (NumPy 2.3.5, OpenCV 4.11.0.86, and SciPy 1.15.3). Latency measures mapper processing only and excludes rendering. Separate attrition, grammar, and hypothesis-set audits are reported in Appendix K.3.

5.4 Natural association ambiguity recurs across cohorts

In the sealed test split, held out from the 3RScan train release, 527 of 8,896 persistent targets (5.92%) have multiple officially allowed owners, and 122 of 328 pairs admit multiple official maximum matchings. The separately evaluated validation cohort contains 161 ambiguous targets among 2,789 (5.77%) and 48 of 110 pairs. These rates concern released annotation-compatible association, not the number of physically possible histories.

The geometry-only 95% conformal graph covers 94.85% of owners scene-equally [95% environment-bootstrap interval 93.49, 96.00] but returns 4.69 candidates on average. Its much larger 88.17% local candidate ambiguity reflects information missing from this deliberately restricted interface, not annotation ambiguity. Coverage falls to 39.70% for the 471 targets displaced by at least 1 m. The geometry-only result therefore describes annotation compatibility rather than a solution to long-range re-identification.

5.5 Natural-data utility shows no universal winner

The primary category-blind utility analysis does not support a general policy advantage. It costs 0.268–0.292 at zero query, above the 0.250 always-unresolved control. A retrospective category gate yields a non-dominated worst-case frontier for LIMBO (0.184 upper cost, 64.32% coverage, 0.130 m regret) but remains non-confirmatory; all budgets and open-set unresolved cases are in Appendix K.

5.6 Released systems trade coverage for precision

GLASSES yields 141 evaluable associations and ConceptGraphs yields 3 under their released configurations (Table 3). Sparse ConceptGraphs has perfect compatible precision, while coverage-aware evidential F1 favors GLASSES by 0.0805 [0.0579, 0.1053]; this low-coverage ranking is descriptive. Evidence-scoped reporting withholds 31 unsupported GLASSES lineage answers while retaining all

Table 3: Official-system outputs on the same 2,789-target natural task. "Eval." is matched cross-endpoint association count; "Single," "Compat.," and "Certain" count primary-owner correctness, membership in the allowed owner set, and global all-matching lineage support. "Location" is retained persistent current locations. Scene F_1 averages environments, including empty-output environments.

System	Eval.	Single	Compat.	Certain	Location	Scene F_1
GLASSES	141	114	115	110	237	0.082
ConceptGraphs	3	3	3	3	3	0.001

237 matched current locations, so ambiguity does not excuse predictions outside the allowed owner set.

6 Discussion and Limitations

The controlled collisions isolate a basic measurement problem: location and lineage need not be equally identifiable from the same observation. Scoring a complete trajectory against one selected history hides this distinction. Limbo makes it explicit by applying certain-answer semantics to typed map claims. This framing is most useful when a false lifecycle commitment carries a higher cost than leaving lineage unresolved. A calibrated probabilistic score serves a different purpose when decisions depend on expected loss.

The distinction also changes how mapping failures are diagnosed. Missing transport evidence, incorrect association, and loss of a current location have different consequences, although a single identity label may merge them. The typed interface preserves these differences for hard trackers, hypothesis-set trackers, and systems that abstain. The 3RScan analysis shows that analogous ambiguity occurs in released annotations. Its low-coverage system comparison, however, cannot establish how often this ambiguity changes rankings in deployed mappers.

Every guarantee is relative to the finite grammar used to construct a group. The extension experiments show that adding plausible histories can withdraw lineage support while leaving location support intact. This dependence is a property to report, not evidence that the grammar is physically complete. The natural study also relies on oracle segments and known poses, and the ten-scene RGB-D experiment uses anonymous oracle masks in its primary condition. A stronger system-level test would combine learned proposals, pose estimation, and a prospectively frozen cohort, then measure whether claim errors alter search or navigation decisions.

7 Conclusion

Limbo makes an implicit benchmark assumption explicit. It separates spatial realization from physical entity, constructs sensor-matched compatible histories, and scores only typed claims that survive their intersection. Controlled experiments expose the error caused by selecting one hidden history, and natural scans reveal the corresponding ambiguity in released annotations. The protocol evaluates which association claims follow from the available evidence; association itself remains the mapper's task.

AI use statement

Generative AI tools were used only to polish the language and organization of the manuscript. The authors designed the methods, conducted and analyzed the experiments, prepared the figures, verified the citations and results, and take responsibility for the final content.

Ethics statement

The study uses simulated household scenes and the existing 3RScan train and validation releases; it collects no new personal data and involves no human subjects. Persistent semantic maps can reveal sensitive household routines. Limbo makes provenance and unresolved identity explicit, but it is

not a privacy mechanism. Deployed systems still require appropriate consent, access control, and data-retention policies.

Reproducibility statement

Section 4 and Appendix H specify the protocols, hashes, and build. The supplement contains schemas, source, manifests, immutable summaries, generators, and tests. Full reruns require external ProcTHOR/Unity assets and provider-authorized 3RScan data.

References

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. ISSN 1687-5281. doi: 10.1155/2008/246309.

Lukas Bernreiter, Abel Roman Gawel, Hannes Sommer, Juan Nieto, Roland Siegwart, and Cesar Cadena Lerma. Multiple hypothesis semantic mapping for robust data association. *IEEE Robotics and Automation Letters*, pp. 1–1, 2019. ISSN 2377-3774. doi: 10.1109/lra.2019.2925756.

Benjamin Bogenberger, Oliver Harrison, Orrin Dahanaggamaarachchi, Lukas Brunke, Jingxing Qian, Siqi Zhou, and Angela P. Schoellig. Where did i leave my Glasses? open-vocabulary semantic exploration in real-world semi-static environments. *IEEE Robotics and Automation Letters*, 11(3): 3342–3349, March 2026. ISSN 2377-3774. doi: 10.1109/lra.2026.3656790.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, 2022.

Oliver E. Drummond and Barry E. Fridling. Ambiguities in evaluating performance of multiple target tracking algorithms. In Oliver E. Drummond (ed.), *Signal and Data Processing of Small Targets 1992*, volume 1698, pp. 326–337, August 1992. doi: 10.1117/12.139399.

Yilun Du, Tomas Lozano-Perez, and Leslie Pack Kaelbling. Learning object-based state estimators for household robots. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12558–12565, October 2022. doi: 10.1109/iros47612.2022.9981287.

Ángel F. García-Fernández, Jason L. Williams, Karl Granström, and Lennart Svensson. Poisson multi-bernoulli mixture filter: Direct derivation and implementation. *IEEE Transactions on Aerospace and Electronic Systems*, 54(4):1883–1901, August 2018. doi: 10.1109/TAES.2018.2805153.

Angel F. Garcia-Fernandez, Abu Sajana Rahmathullah, and Lennart Svensson. A metric on the space of finite sets of trajectories for evaluation of multi-target tracking algorithms. *IEEE Transactions on Signal Processing*, 68:3917–3928, 2020. ISSN 1941-0476. doi: 10.1109/tsp.2020.3005309.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Karl Granström, Lennart Svensson, Yuxuan Xia, Jason Williams, and Angel F. Garcia-Fernandez. Poisson multi-bernoulli mixture trackers: Continuity through random finite sets of trajectories. In *2018 21st International Conference on Information Fusion (FUSION)*, pp. 1–5, July 2018. doi: 10.23919/ICIF.2018.8455849.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024.

Tomasz Imieliński and Witold Lipski. Incomplete information in relational databases. *Journal of the ACM*, 31(4):761–791, Sept. 1984. ISSN 1557-735X. doi: 10.1145/1634.1886.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI, 2017. URL `https://arxiv.org/abs/1712.05474`.

Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martín-Martín. Modeling dynamic environments with Scene Graph Memory. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17976–17993, 23–29 Jul 2023.

Leonid Libkin. Incomplete information and certain answers in general data models. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, SIGMOD/PODS '11, pp. 59–70, June 2011. doi: 10.1145/1989284.1989294.

Witold Lipski. On databases with incomplete information. *Journal of the ACM*, 28(1):41–70, January 1981. ISSN 1557-735X. doi: 10.1145/322234.322239.

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, February 2021. ISSN 1573-1405. doi: 10.1007/s11263-020-01375-2.

Abu Sajana Rahmathullah, Angel F. Garcia-Fernandez, and Lennart Svensson. Generalized optimal sub-pattern assignment metric. In *2017 20th International Conference on Information Fusion (Fusion)*, pp. 1–8, July 2017. doi: 10.23919/icif.2017.8009645.

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*, pp. 17–35, 2016. ISBN 9783319488813. doi: 10.1007/978-3-319-48881-3_2.

Lukas Schmid, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments. In *Robotics: Science and Systems XX*, RSS2024, July 2024. doi: 10.15607/rss.2024.xx.081.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. 3RScan: A rich RGB-D instance re-localization dataset using scene alignment. In *2019 International Conference on 3D Vision (3DV)*, 2019a.

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)*, pp. 7657–7666, Oct. 2019b. doi: 10.1109/iccv.2019.00775.

Chenxu Wang, Dunzheng Wang, Dong Wang, Xinghang Li, Yongchang Li, and Huaping Liu. DynamicTHOR: A scalable dataset of human-centric dynamic scenes for embodied AI. *Scientific Data*, 13(1), May 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-07201-7.

Yuxuan Xia, Ángel F. García-Fernández, Johan Karlsson, Kuo-Chu Chang, Ting Yuan, and Lennart Svensson. Probabilistic GOSPA: A metric for performance evaluation of multiobject filters with uncertainties. *IEEE Transactions on Aerospace and Electronic Systems*, 61(5):15113–15121, October 2025a. ISSN 2371-9877. doi: 10.1109/taes.2025.3580523.

Yuxuan Xia, Ángel F. García-Fernández, Johan Karlsson, Yu Ge, Lennart Svensson, and Ting Yuan. Probabilistic trajectory GOSPA: A metric for uncertainty-aware multi-object tracking performance evaluation. In *2025 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 1–7, Sept. 2025b. doi: 10.1109/mfi67357.2025.11259230.

Shibo Zhao, Guofei Chen, Honghao Zhu, Zhiheng Li, Changwei Yao, Nader Zantout, Seungchan Kim, Wenshan Wang, Ji Zhang, and Sebastian Scherer. SuperMap: A spatio-temporal SLAM system for visual-language navigation, 2026. URL https://arxiv.org/abs/2608.22896.

Liyuan Zhu, Shengyu Huang, Konrad Schindler, and Iro Armeni. Living scenes: Multi-object relocalization and reconstruction in changing 3D environments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28014–28024, June 2024. doi: 10.1109/cvpr52733.2024.02646.

A Formal Semantics and Boundary Conditions

Proposition 1 (hard-decision ceiling). Let c_0 be true only in $w_0 \in \mathcal{W}_G(E)$ and c_1 be true only in $w_1 \in \mathcal{W}_G(E)$. Suppose a balanced benchmark samples one of these worlds, and accuracy is the binary correctness of selecting exactly one of $\{c_0, c_1\}$. No deterministic or randomized evidence-only output exceeds expected accuracy $1/2$ on this decision.

Proof. The public input is E in both worlds, so determinism gives the same output $f(E)$. If it selects c_0, the output is correct only in w_0; if it selects c_1, it is correct only in w_1. Balanced sampling gives probability $1/2$. For a randomized rule, condition on its internal seed. Every conditioned rule is deterministic and has accuracy at most $1/2$; averaging over seeds preserves the bound. □

Proposition 2 (monotonic resolution). If valid additional evidence refines the compatible set such that $\emptyset \neq \mathcal{W}_G(E') \subseteq \mathcal{W}_G(E)$, then $\Gamma_G(E) \subseteq \Gamma_G(E')$.

Proof. Any claim in $\Gamma_G(E)$ belongs to $T(w)$ for every $w \in \mathcal{W}_G(E)$, and therefore for every $w \in \mathcal{W}_G(E')$. □

Proposition 3 (typed partial-claim separation). Consider an evidence contract with two compatible worlds sharing the atom $\textsc{At}(r, p)$ but disagreeing on the owner atoms $\textsc{Same_Lineage}(r, m)$ and $\textsc{Same_Lineage}(r, a)$. Any world-selection reduction that emits one complete hard trajectory must choose one owner and therefore cannot emit the supported location while remaining unresolved about lineage. A set-valued or probabilistic reduction can retain both owners, but it represents a different output type; it becomes the typed partial claim above only after an explicit intersection-and-compilation adapter.

Proof. The two owner atoms are mutually exclusive in each complete world, while the location atom is present in both. A complete hard trajectory contains one owner, so it asserts an atom absent from one compatible world. Retaining both owners avoids that assertion but does not by itself emit the location atom as an entailed lifecycle claim or distinguish its observation-scoped type. The intersection of the two complete truth sets contains the location atom and no owner atom, which is exactly the typed partial output. □

Boundary conditions. Monotonicity assumes a fixed claim language and fixed world truth. Changing pose quantization, retracting a trusted action, or changing the physical grammar creates a new benchmark version. If no claim is entailed, a nonempty output has soundness zero; empty output has soundness one and completeness zero. The group then remains useful as an overreach probe rather than an accuracy target.

Declared universe and relative completeness. Version 1 quantifies over the finite history grammar recorded in each manifest, not over every physically conceivable scene evolution. For a retained group, $\mathcal{W}_G(E)$ contains every successfully rendered grammar member admitted by the published attrition rules. Missing-world and allowlist audits verify this enumeration relative to G; they do not establish that G is physically complete. Consequently, "supported" throughout the paper means supported relative to the versioned grammar and evidence contract. Extending the grammar creates a new benchmark version and may withdraw formerly supported claims.

Table 4: Version-1 typed claims. A realization is observation-scoped; an entity is a persistent physical lineage. Birth and removal refer to the interval between the declared initial and endpoint times.

Claim	Arguments	Meaning
AT	realization, pose	occurrence is at pose
ABSENT_AT	realization, pose	occurrence is absent locally
SAME_LINEAGE	realization, entity	shared physical lineage
BORN	entity	entity entered map domain
REMOVED	entity	entity left map domain

B CLAIM SCHEMA AND EVIDENCE CONTRACTS

Every public JSONL handle belongs to the group allowlist and records the benchmark, split, house, group, realization, category, contract-specific evidence, content hashes, and grammar version. Hidden scorer JSONL records physical lineages, enabled intervals, per-world truth, compatibility membership, and corruption fixtures.

Table 5: Evidence ladder and entailed claim boundary.

Contract	Public evidence	Entailed boundary
Endpoint geometry	calibration, category, pose, depth, mask	AT; lineage unresolved for relocation/replacement twins
Endpoint + free space	endpoint plus projected old-surface rays	ABSENT_AT when clear; never REMOVED from rays alone
Passive transport	ordered 8–24 observation tube under bounded motion/no swap	AT and SAME_LINEAGE
Own action	successful pickup, transport, placement hash chain	AT and SAME_LINEAGE
Explicit events	trusted hash-chained import/export events	BORN and REMOVED; upper bound

B.1 SUBJECT-SPECIFIC QUERY SEMANTICS

Let A be a complete world truth set, P a mapper's partial claim set, m the public mapped entity, r the current realization, and i a public imported entity when the contract supplies one. Write $\text{owner}_A(r) = e$ when $\text{SAME_LINEAGE}(r, e) \in A$ and $\text{pose}_A(e) = p$ when both $\text{SAME_LINEAGE}(r', e)$ and $\text{AT}(r', p)$ occur in A. The schema rejects multiple owners of one realization. The four history queries have the following world-level Boolean truth conditions:

$$q_{\text{same}(r,m)}(A) = [\text{owner}_A(r) = m], \tag{4}$$

$$q_{\text{moved}(r,m)}(A) = [\text{owner}_A(r) = m] \wedge [d(r) > 0.10\,\text{m}], \tag{5}$$

$$q_{\text{born}(r)}(A) = [\text{BORN}(\text{owner}_A(r)) \in A], \tag{6}$$

$$q_{\text{removed}(m)}(A) = [\text{REMOVED}(m) \in A]. \tag{7}$$

Here $d(r)$ is the contract-defined initial-to-current displacement. In an evaluation record, WAS_MOVED therefore stores r as its subject and m as its mapped-entity subject. In an explicit-import group without a current realization, the birth query is $q_{\text{born}(i)}(A) = [\text{BORN}(i) \in A]$. Thus a birth claim about an unrelated entity cannot answer a query about r or i.

For a Boolean query q, the group reference is true when $q(T(w))$ is true in every compatible world, false when it is false in every world, and unresolved otherwise. $\text{WHERE_NOW_REALIZATION}(r)$ returns true with pose p only when every world contains $\text{AT}(r, p)$ for the same quantized p; otherwise it is unresolved. $\text{WHERE_NOW_ENTITY}(m)$ returns true with p when every world gives $\text{pose}_{T(w)}(m) = p$, false when every world contains $\text{REMOVED}(m)$, and unresolved otherwise. A realization location can therefore be supported without persistent identity, and ABSENT_AT never implies REMOVED.

Prediction sets are intentionally partial. Their compiler returns true for a queried atom or witness, false only from an explicit exclusive alternative, and unresolved when neither is present. In particular, an owner different from m makes IS_SAME and WAS_MOVED false; owner m makes WAS_REMOVED false; and $\text{WAS_BORN}(r)$ is true only when P contains $\text{BORN}(\text{owner}_P(r))$. If $\text{owner}_P(r) = m$ and

no such birth exists it is false, whereas an alternative owner without its own birth claim is unresolved. For the explicit-import form, only $\textsc{Born}(i)$ answers true.

Admissibility is checked before claim and query scoring. Spatial predicates take only a realization and pose; SAME_LINEAGE takes only a realization and entity; and BORN/REMOVED take only an entity. A partial set is rejected if it gives one realization multiple owners or multiple current poses; contains both $\textsc{At}(r, p)$ and $\textsc{Absent_At}(r, p)$; asserts BORN for an initially mapped entity; assigns a current realization to a removed entity; or locates current realizations of one entity at unequal poses. Local absence at one pose may coexist with presence elsewhere, and birth plus later removal of an otherwise unobserved transient entity is not ruled out. Thus every accepted output has an endpoint model, and no precedence rule can turn an inconsistent set into separate query answers. These rules precede the anchored, co-reference-preserving normalization used for claim scoring.

B.2 WORKED ENDPOINT EXAMPLE

Let r be the endpoint realization at pose p, m the public opaque handle for the previously mapped entity, and a_1 the first anonymous entity role. Consider the two-world restriction $\mathcal{W}_G(E) = \{w_m, w_r\}$, where w_m is hidden relocation and w_r is replacement. After entity-role normalization, the atom universe is $\{\textsc{At}(r, p), \textsc{Same_Lineage}(r, m), \textsc{Same_Lineage}(r, a_1), \textsc{Born}(a_1), \textsc{Removed}(m)\}$, and the complete positive truth sets are:

Table 6: A complete truth table for one endpoint example. Unlisted atoms in a row are false. Each current realization has exactly one explicit owner.

World	$\textsc{At}(r, p)$	$\textsc{Same_Lineage}(r, m)$	$\textsc{Same_Lineage}(r, a_1)$	$\textsc{Born}(a_1)$	$\textsc{Removed}(m)$
Hidden relocation w_m	1	1	0	0	0
Replacement w_r	1	0	1	1	1
Intersection $\Gamma_G(E)$	1	0	0	0	0

The public mapper input is identical in both rows and exposes r, p, m, but not the scorer's raw physical lineages, counterfactual entity ID, or world label. A reuse output compiles to $P_m = \{\textsc{At}(r, p), \textsc{Same_Lineage}(r, m)\}$; any tracker-minted fresh ID $\hat{n} \neq m$ is alpha-normalized to $P_{\text{anon1}} = \{\textsc{At}(r, p), \textsc{Same_Lineage}(r, a_1)\}$. A second distinct anonymous entity would receive a_2, and repeated occurrences of either role remain co-referential across SAME_LINEAGE, BORN, and REMOVED. For either output, one of two predictions is supported and the sole entailed claim is recovered, giving soundness $1/2$, completeness 1, and F1 $2/3$. In the released four-history group, birth and reappearance add different lifecycle truth while sharing the replacement render and alternative lineage; the intersection and these scores are unchanged. For contrast, $\textsc{Absent_At}(r, p_0)$ would mean only that the mapped footprint represented by r is clear at old pose p_0; it would not state that entity e has permanently left the map domain.

C CONSTRUCTION AND METHOD REGISTRY

Each qualified house contributes fixed-order targets with at least 1 m destination displacement. Ten histories are attempted: stationary, hidden relocation, same-asset replacement, removed, moved-out, occluded, birth, reappearance, passive transport, and own-action transport. Failed interventions remain in attrition manifests. Exact endpoint groups require byte-equal depth and equal binary masks. RGB is held out for a separate distinguishability audit.

The constructor audit in Table 7 is indexed by the final retained public houses and all public group handles. The audit contains 508, 247, 321, and 498 public groups for train, development, validation, and test, respectively; train plus development are the combined G1 cohort. The passed endpoint rows are the retained endpoint-group counts used in the benchmark cohort table. It reports attempt-level outcomes, not physical-history prevalence. "Unavailable" means that a downstream contract was not attempted because its prerequisite group did not exist. The conditional pass rate is passed divided by passed plus failed. Attrition is substantial and varies across contracts and splits, so we make no independence assumption and do not interpret the retained cohort as a representative sample of physical histories. Group construction is atomic: a failed member causes the whole contract group to be omitted rather than leaving a partially rendered world set in the scored corpus.

Table 7: Constructor attrition by final split and evidence contract. Counts are attempt-level and include passed, failed, and unavailable statuses.

Split	Contract	Attempted	Passed	Failed	Unavail.	Conditional pass
Train	Endpoint	207	116	91	0	56.0%
Train	Action destination	207	123	84	0	59.4%
Train	Free-space	116	45	71	0	38.8%
Train	Own-action	105	105	0	0	100.0%
Train	Stationary	116	45	0	71	100.0%
Train	Occlusion	116	13	32	71	28.9%
Train	Passive transport	116	68	48	0	58.6%
Development	Endpoint	98	57	41	0	58.2%
Development	Action destination	98	57	41	0	58.2%
Development	Free-space	57	21	36	0	36.8%
Development	Own-action	51	51	0	0	100.0%
Development	Stationary	57	21	0	36	100.0%
Development	Occlusion	57	5	16	36	23.8%
Development	Passive transport	57	35	22	0	61.4%
Validation	Endpoint	126	71	55	0	56.3%
Validation	Action destination	126	74	52	0	58.7%
Validation	Free-space	71	31	40	0	43.7%
Validation	Own-action	60	60	0	0	100.0%
Validation	Stationary	71	31	0	40	100.0%
Validation	Occlusion	71	11	20	40	35.5%
Validation	Passive transport	71	46	25	0	64.8%
Test	Endpoint	208	111	97	0	53.4%
Test	Action destination	208	124	84	0	59.6%
Test	Free-space	111	46	65	0	41.4%
Test	Own-action	102	102	0	0	100.0%
Test	Stationary	111	46	0	65	100.0%
Test	Occlusion	111	16	30	65	34.8%
Test	Passive transport	111	66	45	0	59.5%

The evaluator contains five method tiers: hard and spatial controls; mechanism-level GLASSES-style, SuperMap-style, and track-continuity adapters; executable compact MHT outputs; evidence-scoped reporting policies; and trusted explicit events as an upper bound. The compact MHT uses association sigma 0.75 m, gate 4.0 m, detection/miss/birth/false probabilities 0.90/0.10/0.10/0.02, three tolerated misses, at most 64 global hypotheses and 64 branches per parent, log-weight pruning at -12, and N-scan depth 3. Its certain output emits a claim only when every retained nonzero-support global hypothesis contains it.

Table 8: Cross-contract descriptive registry of all 16 held-out output configurations. The rows span different tasks and should not be read as a system leaderboard. Unsupported is mean history-query error and Location is realization availability. Scoped rows are contract-conformance reporting ablations and do not change backend state. The GLASSES-style, SuperMap-style, and track-continuity rows are mechanism adapters, not published-system accuracy measurements. The abstention HOTA value follows the protocol's forced category/location convention and is not general HOTA.

Output	Sound.	Compl.	Macro F_1	HOTA	Unsupported	Location
Abstain all	1.000	0.000	0.000	0.367	0.000	0.000
GLASSES mechanism adapter	0.882	0.974	0.895	0.864	0.280	1.000
GLASSES adapter + scope	1.000	0.974	0.974	0.977	0.000	1.000
Hard replace	0.674	0.682	0.603	0.977	0.381	1.000
Hard reuse	0.882	0.890	0.811	0.864	0.280	1.000
Spatial only	1.000	0.765	0.835	0.652	0.000	1.000
Spatial TTL	0.777	0.785	0.706	0.977	0.269	1.000
Spatial TTL + scope	1.000	0.785	0.820	0.977	0.000	1.000
SuperMap mechanism adapter	0.840	0.974	0.867	0.977	0.169	1.000
SuperMap adapter + scope	1.000	0.974	0.974	0.977	0.000	1.000
TOMHT certain	0.909	0.786	0.760	0.652	0.116	1.000
TOMHT certain + scope	1.000	0.786	0.821	0.652	0.000	1.000
TOMHT hard	0.882	0.890	0.811	0.968	0.149	1.000
TOMHT hard + scope	1.000	0.890	0.890	0.977	0.000	1.000
Track-continuity adapter	0.882	0.890	0.811	0.977	0.140	1.000
Track adapter + scope	1.000	0.890	0.890	0.977	0.000	1.000

D Sensor Stress and Composition

Stress seeds are derived from the published corruption version, condition, group, and observation handle. Atomic conditions use miss probabilities 0.10/0.25, false-positive probabilities 0.02/0.05, 3/7-pixel mask erosion and dilation, 10% boundary dropout, 5/15% depth zeroing, 0.005/0.02 m Gaussian depth noise, and 0.01 m+1° / 0.05 m+5° pose noise. Low and high coupled conditions combine their respective settings.

Table 9: Coupled-stress GLASSES-style results. Location is current-realization availability. All occlusion false-withdrawal rates are zero.

Split	Stress	Ray F1	TP	FN	Location
Discovery	Low	0.9440	59	7	0.9177
Discovery	High	0.8214	46	20	0.7831
Selection	Low	0.9667	29	2	0.8942
Selection	High	0.7347	18	13	0.7788
Held-out	Low	0.8916	37	9	0.9015
Held-out	High	0.9048	38	8	0.7754

The original compound gate is conjunctive, so the high value cannot replace the failed prespecified low-coupled held-out row. The failure comprises nine missed free-space withdrawals rather than unsafe withdrawal under occlusion.

The repeated-corruption audit fixes a new seed namespace and applies 256 corruptions per cohort and coupled condition. It changes no method, threshold, corruption magnitude, or original fixed-draw result. Figure 4 and Table 10 therefore describe seed variability rather than overwrite the fixed test draw.

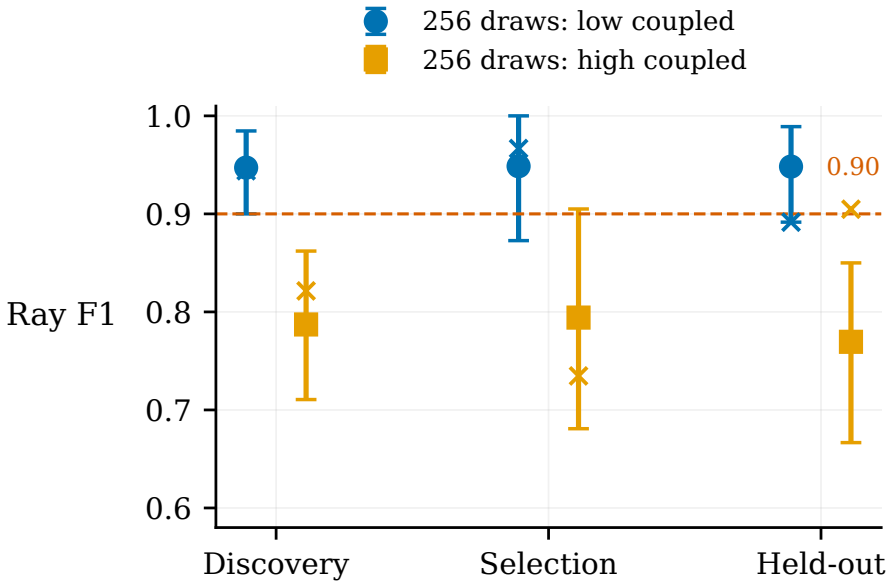

Figure 4: Ray F1 across 256 corruptions. Points and bars are means and 2.5–97.5 percentiles; crosses are the original fixed draws.

Category-equal GLASSES-style minus track-continuity differences are 0.0569/0.0719/0.0631 on discovery/selection/held-out. Every leave-one-category and leave-one-house value remains positive. Endpoint scores agree for both native and inserted twin strata. In a finite-corpus planning diagnostic, 2,000 deterministic subsets are positive in 0.9985 of four-house draws and every draw from eight houses upward; this is not independent generalization evidence.

Table 10: Repeated coupled stress. Intervals are empirical corruption-draw quantiles, not across-environment uncertainty intervals.

Split	Low mean [95%]	Pass prob.	High mean [95%]	Fixed low
Discovery	0.947 [0.900, 0.985]	0.996	0.787 [0.711, 0.862]	0.944
Selection	0.949 [0.873, 1.000]	0.938	0.794 [0.681, 0.905]	0.967
Held-out	0.948 [0.892, 0.989]	0.965	0.769 [0.667, 0.850]	0.892

Table 11: Composition sensitivity. Category-equal is the unweighted mean of per-category mean F1 differences. Each leave-category and leave-house value is instead the group-weighted mean after deleting one category or house; the reported ranges span those deletions and are not expected to contain the category-equal estimand. Native/inserted gives endpoint-group counts.

Split	Category equal	Leave-category range	Leave-house range	Native/inserted
Discovery	0.057	[0.080, 0.092]	[0.085, 0.089]	31/142
Selection	0.072	[0.082, 0.100]	[0.091, 0.100]	20/51
Held-out	0.063	[0.088, 0.096]	[0.089, 0.095]	22/89

E Contract-Aligned and Aggregation Audits

The original aggregate method comparison spans different task supports. Table 12 reports paired effects separately for every evidence contract. Fixed HOTA is defined only for endpoint geometry and passive transport. On those common-support groups, Limbo F1 differences are exactly zero in all cohorts; house-macro HOTA differences are -0.071 $[-0.094, -0.047]$, -0.077 $[-0.125, -0.021]$, and -0.112 $[-0.139, -0.088]$ for discovery, selection, and held-out. The earlier aggregate sign contrast therefore summarizes different tasks rather than a within-contract reversal.

Table 12: Contract-stratified GLASSES-style minus track-continuity effects. A dash means that the conventional trajectory metric is undefined for that contract. Values are house-macro means.

Split	Contract	Groups	Δ macro F1	Δ fixed HOTA
Discovery	Endpoint geometry	173	+0.000	-0.136
	Endpoint + free space	84	+0.892	–
	Passive transport	169	+0.000	+0.000
	Own action	156	+0.000	–
	Explicit events	173	+0.000	–
Selection	Endpoint geometry	71	+0.000	-0.146
	Endpoint + free space	42	+0.813	–
	Passive transport	77	+0.000	+0.000
	Own action	60	+0.000	–
	Explicit events	71	+0.000	–
Held-out	Endpoint geometry	111	+0.000	-0.203
	Endpoint + free space	62	+0.809	–
	Passive transport	112	+0.000	+0.000
	Own action	102	+0.000	–
	Explicit events	111	+0.000	–

Primary F1 first takes the harmonic mean per group, then averages groups within houses and houses. Table 13 recomputes the comparison by taking the harmonic mean after macro-aggregation and by weighting contracts equally. The magnitude changes, so the aggregation order is part of the metric definition; the sign and house-bootstrap intervals remain positive under both sensitivity choices.

F Semantic, Reference, and Refinement Audits

The benchmark declares an explicit finite atom universe for every group and exactly one Same_Lineage owner for every current realization in every world. Automated audits cover all 1,574 groups: every atom universe equals the union of its world truths, with zero missing or multiple current owners. The 355 endpoint groups each contain four semantic histories and exactly two unique endpoint render states. The mapped and explicitly imported public handles are distinct fixed an-

Table 13: F1 aggregation sensitivity. Intervals contain 95% of 10,000 house-bootstrap draws.

Split	Group-macro ΔF_1	Post-aggregate ΔF_1	Contract-equal ΔF_1
Discovery	+0.073 [+0.056, +0.090]	+0.034 [+0.025, +0.043]	+0.178 [+0.166, +0.190]
Selection	+0.085 [+0.064, +0.105]	+0.040 [+0.029, +0.052]	+0.163 [+0.143, +0.182]
Held-out	+0.084 [+0.063, +0.104]	+0.040 [+0.029, +0.050]	+0.162 [+0.145, +0.178]

chors; all private entities receive injective structural roles that preserve equality and distinction across SAME_LINEAGE, BORN, and REMOVED. Prediction-side anonymous entities are then matched to these roles by an exact maximum-support injective assignment; unmatched prediction entities remain distinct overreach. New tracker IDs compile only to alternative lineage claims and do not silently create physical BORN claims. Unit tests verify public-anchor separation, partial outputs, multiple births, multi-entity renaming invariance, cross-predicate co-reference, subject-specific birth queries, endpoint overreach, malformed arguments, and jointly inconsistent outputs. This alpha-equivalence is intentionally group-local: persistent IDs are compared within each history group, not constrained across separate groups in a house.

Table 14 applies standard reductions to the same endpoint predictions. Mean over four histories is sensitive to the fact that replacement, birth, and reappearance share one trajectory. Mean after deduplicating the two render states, best case, and worst case all tie reuse and new ID, but for different reasons: best case accepts either choice, whereas worst case penalizes both. None expresses the atom-level result that location is supported while either lineage is not. LIMBO does so directly.

Table 14: HOTA reductions for the two fixed endpoint outputs. Δ is reuse minus new ID. The single rows are analytic two-frame, one-object cases.

Reference reduction	Reuse	New ID	Δ
Single: relocation	1.000	0.707	0.293
Single: alternative	0.707	1.000	-0.293
Mean: four histories	0.780	0.927	-0.146
Mean: two renders	0.854	0.854	0.000
Worst case	0.707	0.707	0.000
Best case	1.000	1.000	0.000
LIMBO F_1	0.667	0.667	0.000

To test that this conclusion is not fixed by the two-frame construction, we vary pre-change frames, post-change frames, and distractors over $\{1, 2, 4, 8\} \times \{1, 2, 4, 8\} \times \{0, 1, 4\}$. The 48 configurations yield 24 distinct reuse-minus-new-ID HOTA effects over $[-0.292893, +0.292893]$. This is a deterministic sequence-length and dilution calculation, not 48 independent sampled datasets or a crowded association test. Detections are exact, the target has one category, and stationary distractors have distinct categories; frame matching is therefore forced and the protocol-frozen implementation reduces to the standard HOTA association formula in this audit.

The retrospective RGB audit compares source bytes rather than embeddings. It finds 75/173, 37/71, and 37/111 RGB-identical endpoint groups in discovery, selection, and held-out, spanning 37, 18, and 22 houses. Since depth, masks, categories, calibration, and poses already match, these 149 groups are exact RGB-D endpoint collisions. The audit covers observed endpoints only; intermediate streams may differ.

Finally, endpoint and stronger-evidence records are paired by source house, mapped handle, and endpoint pose. Held-out supplies 66 passive-transport and 102 own-action pairs; selection supplies 46 and 60; discovery supplies 103 and 156. In every pair, the declared grammar contracts four worlds to one, entailed claims grow from one to two, and contract-conforming track-continuity F1 changes from 0.667 to 1.000. This numerical change is a deterministic compiler check. For the relaxed-grammar sensitivity, the stronger public record is held fixed while its scorer world set gains one symbolic, non-rendered counterfactual with an unobserved mid-transport swap or an action log that does not bind physical identity. The resulting two-world intersection supports lineage in zero pairs. The test isolates the formal trust assumption; the physical frequency and detectability of these counterfactuals remain unmeasured.

Table 15: Held-out refinements paired on the same object and endpoint.

Added evidence	Pairs	Worlds	Entailed	Report F_1
Continuous transport	66	4→1	1→2	0.667→1.000
Own-action log	102	4→1	1→2	0.667→1.000

The same symbolic extension can be summarized across all paired cohorts. Table 16 reports the declared entailed-claim count and the count after admitting one non-rendered identity-swap or identity-unbound world per pair. It removes only the lineage claim, while retaining the current-location claim. This retrospective sensitivity examines the formal grammar boundary. It neither estimates physical-history prevalence nor shows that the added world would render. Rendering attrition is separately retained at the constructor-attempt level; retained groups are written atomically, so a failed member does not leave an incomplete group in the scored corpus.

Table 16: Retrospective symbolic grammar sensitivity for paired refinements. The extension holds the public record fixed and adds one non-rendered identity counterfactual.

Cohort	Contract	Pairs	Declared claims	Extended claims	Lineage support
Discovery	Continuous transport	103	206	103	100%→0%
Discovery	Own-action log	156	312	156	100%→0%
Selection	Continuous transport	46	92	46	100%→0%
Selection	Own-action log	60	120	60	100%→0%
Held-out	Continuous transport	66	132	66	100%→0%
Held-out	Own-action log	102	204	102	100%→0%

G Query-Association Sanity Check

The query audit excludes every evidence-scoped variant; abstention was not part of the source eligible-method rows. It treats houses as bootstrap clusters and uses 10,000 draws with seed 2401. Thus the result cannot be driven by methods whose unsupported-answer rate is zero by construction. Table 17 reports every interval rather than comparing significance thresholds. Across leave-one-method-out fits, soundness ρ ranges are $[-0.964, -0.552]$, $[-0.941, -0.550]$, and $[-0.958, -0.483]$ for discovery, selection, and held-out. These checks address clustering and method-family sensitivity, but the target still shares the benchmark reference and remains an internal consistency measurement.

Table 17: Spearman association with unsupported-answer rate after excluding scoped variants. Bootstrap intervals resample houses.

Split	Rows/houses	Soundness ρ [95% interval]	IDF1 ρ [95% interval]	HOTA ρ [95% interval]
Discovery	240/48	-0.718 [-0.772, -0.654]	0.024 [-0.050, 0.107]	0.034 [-0.038, 0.119]
Selection	100/20	-0.724 [-0.810, -0.622]	0.037 [-0.091, 0.190]	0.067 [-0.048, 0.199]
Held-out	160/32	-0.654 [-0.714, -0.587]	-0.041 [-0.098, 0.012]	-0.029 [-0.086, 0.025]

H Runtime, Integrity, and Reproducibility

Every MHT size uses five warmups and 30 timed reconstructions. Setup and import are excluded. Positions use a 1 m grid with 0.01 m offsets, which bounds candidate degree. Dense all-to-all ambiguity lies outside this scaling test.

Prediction operates against a public-only root with no scorer path. Audits require exact public/scorer allowlists, content hashes, source-house uniqueness, reverse-order replay, absence of private label keys, shuffled-label invariance, and rejection of fixed corruptions. Discovery, selection, and held-out pass. The joint provenance key is (ProcTHOR, immutable revision, official source split, source index): all 1,574 records resolve to 100 unique source keys and house handles, with zero cross-cohort overlap or split mismatch. Predictions are hash-committed before the conventional projection salt is supplied in the recorded pipeline. The archive has no external timestamp and therefore attests, rather than independently proves, this chronology. A deliberate leak oracle reaches HOTA 1.0 on

Table 18: Compact MHT scaling on a Ryzen 9 3900X (24 logical CPUs). Candidate degree is deliberately bounded; five warmups and 30 reconstructions exclude setup/import. State is recursive deep size, not process RSS.

N	Active p95 (ms)	Missing p95 (ms)	Active state (MB)	Missing state (MB)
1	0.113	0.111	0.007	0.009
10	3.935	3.327	0.042	0.049
25	9.372	9.857	0.094	0.106
50	22.295	19.944	0.146	0.196
100	45.587	40.446	0.252	0.364
250	140.337	105.640	0.540	0.820
500	364.629	248.836	1.053	1.611
1000	1052.657	618.255	2.028	3.145

all 173/71/111 endpoint groups when given that salt early, so fixed projection is retained only as an offline diagnostic. The included source-shard, recovery, and qualification records retain all attempt-level outcomes; their audit resolves all 24 hashes referenced by the three cohort merges and two recovery records without a missing or duplicate target. The repository test report and deterministic nine-clause paper-readiness audit are included as generated artifacts.

Table 19: Reproduction boundary. "Included" means the anonymous archive is sufficient; external rows require the named upstream resource.

Target	Required material	Status
Paper tables, figures, PDF	included JSON, source, TeX	included; deterministic
Source tests and scorer audits	included fixtures, Python 3.10/3.11	included
Full ProcTHOR rerender	ProcTHOR revision + Unity build	external public assets
Official GLASSES rerun	released code/models + GPU	external public assets
Raw 3RScan rerun	provider-authorized scans	access-controlled

All paper tables and figures derive from immutable JSON summaries and can be regenerated without simulator or mapper access. The archive distinguishes this artifact-only build from full experiment execution: the latter requires the external ProcTHOR/Unity assets and the released GLASSES environment; raw-data 3RScan reruns additionally require provider-authorized scans. In the anonymous ICLR package, the build command is:

```
make
```

The source template is official ICLR 2027 commit `46ed6f4c6cef5b175dde23639e77d44c3463b230`; the unmodified style file has the recorded integrity digest `797deef41724e93761426ac0cbcca46279a91cc650dd1f0ce76a4f08d2098ea6`. Bibliography records were retrieved from DOI, arXiv, or official PMLR metadata; their provenance is stored with the artifact.

I Complete 3RScan Annotation-Availability Case Study

The annotation audit protocol was fixed before downloading any newly authorized validation scan. It includes 47 references, 110 rescans, and 157 unique scans. The official layout supplies 471 files totaling 7,362,075,762 bytes. A deterministic manifest stores every URL, byte count, and content digest. Full ZIP CRC passes, no partial file remains, all 38,659 RGB/depth/pose index sets agree, and all 7.60 million PLY vertices parse with instance IDs matching semantic inventories.

One sequence has a single RGB-D-pose triplet and remains included. Two semantic OBB centroids near 10^{290}–10^{305} make aligned distance unavailable; their instances remain in lifecycle and PLY counts. Metadata contains 390 rigid rows with seven duplicates, leaving 383 unique category-consistent relations. It also contains 22 removal annotations conflicting with scan membership and 43 shared-ID label conflicts; none receives support.

There are 1,092 reference-only IDs, of which 74 have valid removal evidence, and 224 rescan-only IDs with no explicit birth evidence. Hard set difference therefore adds 1,242 unsupported lifecycle claims in 107/110 pairs and 46/47 environments. Scene-equal lifecycle soundness is 0.0805 with 95% bootstrap interval [0.0373, 0.1344]; unsupported-query error is 0.9195 [0.8656, 0.9627].

Table 20: Complete labelled 3RScan validation cohort. Hard declares scan-set differences to be physical removal or birth.

Audit item	Micro/count	Scene/coverage
Reference-only	74/1092 supported	1018 unknown
Rescan-only	0/224 supported	224 unknown
Pairs with overreach	46/47 env.	107/110 pairs
Hard lifecycle soundness	0.056	0.080 [0.037, 0.134]
Hard unsupported-query rate	0.944	0.920 [0.866, 0.963]
Conservative sound./recall	1.000/1.000	0 unsupported

A post-result sensitivity was specified before subgroup counts. Excluding exactly wall, floor, and ceiling leaves 1,100 unsupported cases across 102 pairs and 43 environments. At least 100 labelled PLY vertices support 84.4% of unsupported reference-only and 87.9% of unsupported rescan-only instances, so the effect is not confined to structural surfaces or zero-point fragments.

J Official GLASSES Posed-RGB-D Annotation-Availability Case Study

The mapper audit fixes the GLASSES authors' released commit, Pixi lock, YOLOv8l-World, MobileSAM, OpenCLIP ViT-H-14, 198-class vocabulary, thresholds, and all annotation-audit validation sequences before inference. Each environment maps its reference scan once; every rescan restores that endpoint at time -1. The released 3RScan loader supplies RGB-D frames and calibrated poses, with official scene alignment applied to rescans. Labelled meshes, instance IDs, and lifecycle metadata remain scorer-only. Two shards complete 57/57 and 53/53 pairs.

The prespecified prediction requires evidence-scoped reporting to remove unsupported physical lifecycle interpretations in at least 10 environments while changing no conventional GLASSES state. It passes in 38/47 environments. Table 21 reports micro and scene-equal results. Empty-claim scenes receive lifecycle soundness one and query error zero under the stated convention. Nine of 47 environments have no hard lifecycle output. Excluding them, the 38-environment scene-equal lifecycle soundness is 0.021, while pooled soundness over the same 202 claims is 0.035.

Table 21: Released GLASSES posed-RGB-D outputs on all 110 validation pairs. Scene values use 10,000 environment-bootstrap draws; both decoders consume bit-identical maps. Empty-output scenes have soundness 1 and unsupported error 0 by convention and are separately counted in text.

Official GLASSES audit	Micro/count	Scene/coverage
Hard lifecycle claims	7/202 supported	195 added unsupported
Affected environments	38/47	prespecified target 10
Lifecycle soundness	0.035	0.209 [0.102, 0.329]
Nonempty-scene soundness	0.035 pooled	0.021 (38/47 scenes)
Unsupported-query rate	0.965	0.791 [0.676, 0.897]
Primary matched objects	0.135 ref.	0.078 rescan
Class-agnostic sensitivity	0.492 ref.	0.361 rescan
Rigid UUID retention	17/383	0.044
Frame-step latency (s)	0.638 median	1.296 p95
Compute / peak VRAM	7.888 GPU-h	7.61 GB

Class-compatible matching changes little from 2.5 to 10 cm (13.42–13.49% reference, 7.69–7.76% rescan), whereas class-agnostic coverage reaches 49.19%/36.11%. Vocabulary and ontology alignment therefore dominate the selected distance tolerance. Strict UUID retention is 17/383 for rigid correspondences, 93/2,354 for stationary shared IDs, and 4/52 for non-rigid shared IDs. These are cross-dataset diagnostics rather than published-system accuracy reproduction.

All 38,659 frames contribute to timing. Mean/median/p95 mapper-step latency is 0.673/0.638/1.296 s. Feasibility plus full execution costs 7.888 GPU-hours on two RTX 3090s. Peak per-sequence allocated VRAM is 7.61 GB and peak process RSS is 13,128,004 KiB. Operational rescan state records 10,741 UUID creations, 2,050 first-missing events, and 10,386 dictionary deletions; none is promoted to a physical event. The runner omitted endpoint-serialization timing, which is disclosed rather than reconstructed.

K Natural Compatible Associations and Official-System Comparison

Table 22: Natural 3RScan association ambiguity in disjoint, separately evaluated train and validation cohorts. Targets are persistent current instances; multi-match pairs admit multiple official maximum matchings. Cohorts are not pooled.

Cohort	Env.	Pairs	Targets	Ambiguous	Multi-match pairs
Train sealed	161	328	8896	527 (5.92%)	122
Validation	47	110	2789	161 (5.77%)	48

K.1 Cohort, calibration, and leakage controls

The natural association protocol exhausts all 385 3RScan train environments and 793 reference/rescan pairs. A UUID hash assigned 64 environments to fitting, 64 to development, 96 to calibration, and 161 to sealed test. The split was fixed from metadata before acquiring any train PLY or semantic file. Acquisition then completed all 2,356 prescribed files for 1,178 scans (4,586,059,029 bytes) with no failure or residual partial file. The sealed cohort contains 328 pairs and 8,896 persistent current targets: 7,684 stationary, 1,029 rigid, and 183 non-rigid.

Each object is represented by a 14-dimensional geometry summary computed from XYZ points only. An initial non-sealed experiment was invalidated before sealed access because the labelled PLY RGB channels are uniform instance-visualization colors and can encode object IDs. The invalid artifact and its three-scorer selection are retained but excluded. The replacement `geometry_xyz_v2` addendum fitted exactly one logistic scorer over all reference instances; semantic category did not enter either its features or candidate gate. The cache fails closed on the old schema, and public feature records contain no raw object ID or scorer-only field. Target-pooled finite-sample nonconformity quantiles fixed the thresholds labelled 90% and 95% before sealed test access. Targets within an environment are dependent, so these labels do not assert a scene-equal population coverage guarantee; held-out scene-equal coverage is reported empirically.

Table 23: Sealed test candidate sets at the two fixed calibration thresholds. Coverage and ambiguity are environment-equal means with 10,000 environment-bootstrap draws; mean candidate size is micro-aggregated over all 8,896 targets.

Calibration target	Coverage [95% interval]	Ambiguous [95% interval]	Mean size
90%	90.95 [89.33, 92.41]	63.11 [60.51, 65.59]	2.51
95%	94.85 [93.49, 96.00]	86.00 [83.93, 87.96]	4.69

Table 23 shows the expected coverage–set-size trade-off. At 95%, globally certain edges have 90.57% precision but 10.79% true-owner recall. All 328 pairs admit more than one maximum matching in the model candidate graph. This model ambiguity must not be confused with the official ambiguity rate: the former depends on the restricted geometry interface and calibration threshold, whereas the latter counts released annotation alternatives and cannot tune the scorer.

Table 24: Descriptive sealed test performance by true-centroid displacement. These strata were computed only after the immutable primary result and are not additional confirmatory tests.

Displacement (m)	Targets	Coverage	Ambiguous	Top-1
$[0, 0.1)$	3559	99.49%	94.80%	94.41%
$[0.1, 0.5)$	4114	97.33%	88.31%	88.72%
$[0.5, 1.0)$	752	90.82%	78.72%	67.82%
$[1.0, \infty)$	471	39.70%	52.02%	26.33%

K.2 Corrected interface and utility analysis

After inspecting the sealed labels, we found three errors in the derived utility analysis. We first reproduced the original model outputs and sealed result, then corrected the downstream estimands

without refitting the scorer. Consequently, all results in this subsection are retrospective. Table 25 reports the actual category-blind interface and a separate category-equality gate. The gate removes 26,473 cross-category edges without changing the score or threshold. Both interfaces retain 94.59% target-micro true-owner coverage.

Table 25: Interface audit on the sealed cohort. "Certain recall" is true-owner recall among edges present in every maximum candidate matching. The category gate is explicitly retrospective.

Interface	Edges	Mean set	Ambiguous	Certain precision	Certain recall
XYZ, all references	41,698	4.687	88.17%	90.57%	10.79%
XYZ + category gate	15,225	1.711	34.64%	99.52%	63.22%

The analysis compares hard top, hard new, always unresolved, local singleton, LIMBO global certainty, and two confidence rejectors. The latter thresholds are selected only on development, either to minimize primary-annotation cost or to match LIMBO coverage. Every policy is converted to a partial injection. Queried edges have priority; ordinary collisions retain the largest frozen probability and then a pre-existing observation hash. If all permitted fields tie exactly, every colliding non-query edge becomes unresolved. This convention is conservative and invariant to target order and numeric ID renaming.

For one pair, risk extrema optimize the total asserted-owner error over one globally coherent maximum matching of the official allowed-owner graph. The original sum of targetwise possible or certain indicators can combine incompatible per-target choices and is retained only as marginal provenance. In the shared- owner counterexample where two targets each allow either of two owners but both predictions assert the same owner, every complete injection has exactly one error; the coherent bounds are therefore $(1, 1)$ rather than $(0, 2)$.

Table 26: Retrospective zero-query policy sensitivity under the category gate. The category-blind interface is primary in the paper. Cost ranges optimize one coherent official matching. Coverage and map regret are scene-equal means over 161 environments.

Policy	Coherent cost	Owner coverage	Map regret (m)
Hard top	0.081–0.406	96.56%	0.026
Always unresolved	0.250–0.250	0.00%	0.292
Local singleton	0.104–0.171	61.15%	0.137
LIMBO global certain	0.098–0.184	64.32%	0.130
Confidence, matched coverage	0.114–0.344	66.61%	0.200
Confidence, development cost	0.081–0.406	96.53%	0.028

The predicted persistent map is constructed before designated truth is read. Every owner starts at its reference centroid, an asserted partial-injection edge moves that owner to the corresponding current centroid, and every unassigned owner remains at an explicitly named stale reference location. Unowned current observations remain spatial realizations and do not update a persistent-owner query. Designated truth specifies which persistent owners and current centroids to score, not where the predicted map places them. The original statistic uses truth to invert a predicted owner into a current target; it is retained only as an *oracle-assisted target spatial surrogate*, never as predicted-map retrieval.

Table 26 shows no universal winner. Under optimistic risk, hard top dominates LIMBO. Under worst-case risk, LIMBO, hard top, and local singleton are all non-dominated on cost, owner coverage, and map regret. On the actual category-blind interface, however, LIMBO costs 0.268–0.292, versus 0.250 for always unresolved. The primary interface therefore provides no utility advantage in this cohort. At zero query, only 27 cost combinations are unique because query cost is inactive; the original count of 81 repeated each cell three times.

The 0/5/10/20/100% nominal budgets realize 0/3.29/8.15/17.36/35.63% query rates (Table 27). We also query the same targets from an always-unresolved policy as an acquisition-only control. Ten thousand deterministic environment-bootstrap draws provide intervals for means and paired differences. We treat the resulting cost grids and confidence curves as one descriptive sensitivity family.

Table 27: Strict per-environment floor budgets for category-gated LIMBO. Only initially unresolved targets can be queried. Coherent cost includes unit query cost; environments, not targets, are bootstrap units.

Nominal budget	Realized query	Owner coverage	Map regret (m)	Coherent cost
0%	0.00%	64.32%	0.130	0.098–0.184
5%	3.29%	67.61%	0.101	0.123–0.208
10%	8.15%	72.45%	0.077	0.158–0.240
20%	17.36%	81.66%	0.047	0.227–0.300
100%	35.63%	99.90%	0.001	0.363–0.425

Annotation set difference alone provides no physical lifecycle label. Among 621 annotation-unmatched current objects, 481 have at least one 95%-compatible prior. No such object is labelled a physical birth. Among 2,513 reference-only annotations, only 128 intersect a valid official removal relation; the other 2,385 remain unresolved. Fifty-seven shared-ID semantic conflicts remain explicit in the audit.

K.3 HYPOTHESIS-SET BASELINE

This analysis uses the same public benchmark records and scorer-only compatible histories as the controlled experiments; it introduces no new sealed cohort. The TOMHT likelihoods, gate, miss budget, birth and false-observation settings are fixed in `experiments/h31-hypothesis-set-baseline/protocol.md`. The bounded configuration reproduces the paper baseline (64 global hypotheses, 64 branches, -12 pruning, N-scan depth 3). The primary expanded configuration retains up to 4,096 hypotheses and branches, uses -24 pruning, and disables N-scan. All four outputs use the same typed endpoint claim interface: MAP, unscoped hypothesis intersection, public-contract-scoped intersection, and the LIMBO contract compiler. Tracker birth alternatives are not silently promoted to physical BORN claims.

Table 28 reports the expanded configuration on all cohorts. The held-out claim values are unchanged from the bounded run despite the much larger support. The extra hypotheses therefore do not recover the benchmark's omitted compatible histories. The false-certainty diagnostic is computed after prediction from the scorer-side entailed claim set; it is not used for model selection. Paired house-bootstrap intervals use 10,000 deterministic draws.

Table 28: Expanded-support hypothesis intersection across cohorts. Values are house-macro percentages; $|\bar{H}|$ is the mean retained global-hypothesis count.

Cohort	Sound.	Compl.	F1	False cert.	Ref. coverage	Mean $\lvert H\rvert$
Discovery	85.5	81.4	83.4	12.4	87.6	391.5
Selection	86.3	80.3	83.2	11.1	88.9	413.5
Held-out	83.1	80.9	82.0	14.2	85.8	379.1

Table 29: Held-out hypothesis-set audit (325 eligible groups). Values are house-macro percentages; "overreach" is pooled unsupported prediction rate. The contract compiler serves as the protocol reference rather than a tracker baseline.

Output	Sound. ↑	Compl. ↑	F1 ↑	Overreach ↓
MAP	81.8	100.0	90.0	17.1
Hypothesis intersection	83.1	80.9	82.0	17.7
Scoped intersection	100.0	80.9	89.4	0.0
LIMBO contract compiler	100.0	100.0	100.0	0.0

On held-out data, the expanded intersection minus MAP evidential-F1 difference is -0.075 with a 95% house-bootstrap interval $[-0.101, -0.046]$; expanded intersection minus contract-scoped intersection is -0.092 $[-0.103, -0.079]$. The scoped intersection minus compiler difference is -0.106 $[-0.118, -0.092]$. These intervals are finite-corpus descriptive uncertainty, not population guarantees. The expanded run has zero cap hits, but its 95th-percentile hypothesis count is 2,373 and latency is about 954 ms, compared with 9 and 14 ms for the bounded run. A deterministic replay reproduces

the public prediction digest and keeps posterior weight sums within 5.8×10^{-14} of one. Wall-clock latency fields naturally differ between replays; the prediction serialization and all claim digests are the audited reproducibility object.

K.4 Failure-stratified attrition

The audit deduplicates construction attempts by split, house, and object identity and joins the public attempt records to the qualification and scorer-only metadata. It reports category, displacement, endpoint-mask visibility, intervention, and contract strata without imputing missing displacement for failed attempts. Across all cohorts there are 685 deduplicated attempts. The held-out audit is shown in Table 30; the displacement and mask strata are shown below. The "own-action" rate is conditional on an action-compatible attempt and therefore does not estimate a rate over all targets. The complete machine-readable payload and hashes are in `experiments/h32-attrition-audit/results/h32_attrition.json`.

Table 30: Held-out construction attrition. Counts are attempt-level; pass rate is conditional on an attempted contract.

Contract	Attempted	Passed	Failed	Pass rate (%)
Endpoint	208	111	97	53.4
Free-space	111	46	65	41.4
Own-action	102	102	0	100.0

Table 31: Held-out endpoint attrition by displacement and initial mask size.

Stratum	Level	Attempted	Passed	Pass rate (%)
Displacement	1-2m	2	2	100.0
Displacement	>=2m	104	100	96.2
Endpoint mask	256-1023	98	94	95.9
Endpoint mask	>=1024	18	17	94.4

The dominant failure reasons are target non-pickupability, the absence of a native twin or qualified action destination, and non-target state changes in absence worlds. These are construction constraints, not failures of the scorer. Because the rates vary substantially by category and visibility, all headline inference remains conditional on the retained cohort.

K.5 Grammar-extension robustness

We fixed the expanded-support predictions before reading scorer-only history records. The three extensions are deterministic additions to the finite grammar: identity swap, delayed observation, and appearance-compatible replacement. We recompute the compatible-history intersection and contract compiler for every affected group; no extension is used for model selection. The identity-swap extension has zero affected held- out groups because those endpoint records already have empty lineage support. The appearance-compatible replacement extension affects 102 groups and turns the contract compiler's 100% baseline completeness into 81.1% pooled soundness, while completeness remains 100%. This supports the intended interpretation: the compiler is sound relative to the published grammar, not a physical-history oracle outside it. The full G1/G2/G3 JSON results and protocol are included under `experiments/h33-grammar-extension-robustness`.

Table 32: Held-out grammar-extension robustness (325 groups). Values are pooled percentages; "Comp." is the contract compiler and "Set" is the unscoped hypothesis intersection.

Extension	Affected	Comp. sound.	Comp. F1	Comp. overreach	Set F1	Set overreach
Appearance-compatible replacement	102	81.1	89.5	18.9	68.0	37.4
Delayed observation	66	87.8	93.5	12.2	86.1	17.7
Identity swap	0	100.0	100.0	0.0	80.7	17.7

K.6 Anonymous sensor-stream mapper

The mapper consumes an online stream of noisy centroids, semantic category, anonymous appearance descriptor, frame index, and sensing-region geometry. Simulator object IDs, event labels, event times, and scorer truth remain private. Four lifecycle policies are run with identical observations. A query asserts the mapper's current entity if available; otherwise it abstains. Costs are 0 for a correct assertion, 5 for an incorrect assertion, and 0.25 for an abstention. Development, validation, and the 20-scene held-out pilot use the same frozen configuration. The complete scene-level records are in `experiments/h34-end-to-end-lifecycle/results`.

Table 33: Held-out downstream query utility on 20 DynamicTHOR scenes. The input is an anonymous sensor stream, not raw RGB-D.

Policy	Query cost ↓	Coverage (%)	Correct (%)	False certainty (%)
BALM-confirmed	0.0212	96.9	96.6	0.3
GLASSES-inactive	0.0090	100.0	99.8	0.2
TTL-immediate	0.0299	100.0	99.4	0.6
Visibility-immediate	0.0090	100.0	99.8	0.2

K.7 Online RGB-D lifecycle experiment

The experiment evaluates explicit relocation, removal, and addition from an online RGB-D stream. Each scene deterministically selects one fixed reachable camera pose with at least three visible pickupable objects. One object is absent before mapping and later enabled, one is relocated to a distinct visible point on its receptacle, and one is disabled. The mapper runs after every observation and receives RGB, metric depth, anonymous regions, camera-frame centroids, a generic `object` category, and frame indices. It receives no simulator ID, semantic class, segmentation colour, intervention type, or event time. A runtime schema assertion rejects identity-bearing mapper fields. The scorer retains colour-to-object mappings separately and joins them only after tracking.

FloorPlan1–20 served as development and stress data; five of these scenes formed an initial runtime pilot, and two placement failures motivated a deterministic relocation fallback. We then fixed the implementation, action schedule, proposal settings, metrics, and bootstrap procedure before running FloorPlan21–30. All ten fresh held-out simulated scenes completed in both proposal conditions without scene replacement or subsequent tuning. This cohort is an internal held-out test, not an external-domain confirmation.

Table 34 reports scene-macro results. Under oracle per-frame masks, the evidence-confirmed lifecycle policy obtains 60.0% relocation retention, 60.0% removal F1, and 60.0% addition F1. Fixed scene-bootstrap 95% intervals are [30.0, 90.0], [30.0, 83.3], and [30.0, 90.0], respectively; false certainty is 0.24% at 94.2% assertion coverage. The other policies differ little, so the experiment provides no evidence of a substantial policy advantage.

The non-oracle sensitivity quantizes RGB and depth, extracts connected components, and caps each frame at 160 anonymous regions. Its proposal precision is 92.9% but recall is only 63.0%. Relocation retention, removal F1, and addition F1 fall to 20.0%, 10.7%, and 10.0%, and latency rises from 4.83 to 94.64 ms/frame. The lifecycle layer remains incomplete even with reliable regions, while naive RGB-D components introduce a much larger error. Neither condition includes learned detection, pose estimation, exploration, or embodied navigation.

K.8 Released ConceptGraphs execution

We pin the released ConceptGraphs-Detect implementation (Gu et al., 2024) and its released stride-5 overlap/sim-sum configuration. YOLO-World detections, MobileSAM masks, and ViT-H/14 embeddings are cached independently from mapping. A preregistered infrastructure addendum handles valid empty detections and clamps boxes exactly to the image domain without changing any model, threshold, vocabulary, stride, cohort, or scorer. Both detection shards cover all 157 scans, and two disjoint reference shards cover all 47 references and 110 pairs with zero failure.

The released endpoint filter retains only 25 reference and 61 rescan nodes (Table 35); the common class-compatible matcher accepts three at each endpoint and yields three cross-endpoint continuities.

Table 34: Online RGB-D results on ten fresh held-out FloorPlan scenes. All accuracy columns are scene-macro percentages; latency is online tracker time per frame. Oracle masks are anonymous within each frame, not identity tokens. "RGB-D CC" is the deterministic connected-component sensitivity.

Proposal	Policy	Prop. recall	Move ret.	Removal F1	Addition F1	Latency (ms)
Oracle masks	BALM-conf.	100.0	60.0	60.0	60.0	4.83
	GLASSES-inact.	100.0	60.0	60.0	60.0	4.82
	TTL-immed.	100.0	60.0	60.0	60.0	4.81
	Visibility-immed.	100.0	60.0	60.0	60.0	4.81
RGB-D CC	BALM-conf.	63.0	20.0	10.7	10.0	94.64
	GLASSES-inact.	63.0	20.0	10.7	10.0	94.59
	TTL-immed.	63.0	20.0	10.7	10.0	94.31
	Visibility-immed.	63.0	20.0	10.7	10.0	94.56

Table 35: Official ConceptGraphs endpoint output and common-matcher coverage. The low native coverage is retained without post-result threshold tuning. Counts are descriptive system outputs, not independent statistical replicates.

ConceptGraphs endpoint audit	Reference	Rescan
Native retained nodes	25	61
Primary matched nodes	3 (12.00%)	3 (4.92%)
Class-agnostic matched	11 (44.00%)	36 (59.02%)

All three are stationary-correct; no rigid continuity or new matched node is evaluable. ConceptGraphs exposes no missing/deletion state, so native-scoped and hard decoders are identical. The prespecified requirement of positive lifecycle coverage in at least 10 environments fails in all 47. We retain the released thresholds and draw no positive lifecycle-decoding conclusion for this configuration.

Detection covers 7,792 stride-5 frames, produces 35,506 raw detections, and contains 1,013 valid empty frames. It takes 1,341.866 GPU-seconds and peaks at 5.813 GB allocated GPU memory. Mapping takes 317.392 CPU-seconds, with 0.0087/0.1009 s median/p95 frame latency and 1,224,912 KiB peak process RSS.

K.9 Common-task scoring

Both official-system maps, all endpoint matches, and the tested ambiguity parser were fixed before counting validation ambiguity. The audit verifies all 314 endpoint file hashes and reconstructs allowed owners only when both IDs exist at the applicable endpoint with category-compatible labels. All 79 released ambiguity relations are valid. The 47 environments, rather than 110 pairs or 2,789 targets, are the independent bootstrap units. Empty-output environments remain in both system means and paired differences.

GLASSES yields 141 evaluable associations: 114 equal the primary owner, 115 belong to the allowed owner set, and 110 are globally certain. An exhaustive audit over all 2,789 targets finds zero differences between local singleton status and certainty in every official maximum matching. This equivalence is empirical for this cohort and need not hold for other matching graphs. ConceptGraphs yields 3/3/3. GLASSES covers seven ambiguous targets; five predictions are compatible and none supports one specific lineage in every allowed history. The scene-equal GLASSES-minus-ConceptGraphs hard evidential F1 difference is 0.0805 [0.05791, 0.10529]; association coverage differs by 0.05332 [0.03873, 0.06932]. Conversely, compatible accuracy differs by -0.10448 $[-0.17040, -0.04728]$, because ConceptGraphs answers only three associations. Reporting precision without coverage would therefore reverse the apparent comparison.

Evidence-scoped reporting converts 31 GLASSES hard lineage answers to unresolved and leaves ConceptGraphs unchanged. The operation retains exactly 237 and 3 matched persistent current locations, respectively; all endpoint hashes and runtime records remain unchanged. This scorer-side noninterference audit evaluates reporting only; it is neither an association algorithm nor an embodied navigation experiment.

L Evidence Status and Protocol Deviations

1. The controlled benchmark was constructed under a broader compound gate whose low-coupled ray F1 requirement failed (0.8916 versus 0.90). The final Limbo framing was selected after this result and should be read as the supported component of that study, rather than an independent preregistered confirmation.
2. Exact RGB-D equivalence applies to the 149 byte-audited endpoint groups. The remaining endpoint groups match only the fields in their declared evidence contract.
3. The 3RScan relations define measurement-compatible owner sets, not an exhaustive set of physical histories. The released-system runs use supplied poses and have low class-compatible coverage; they do not reproduce full SLAM, exploration, or robot operation.
4. GLASSES-style and SuperMap-style adapters measure mechanisms under the Limbo interface; they are not measurements of the published systems' reported accuracy. The corresponding natural-data experiment uses oracle segments, known poses, and scene transforms.
5. An early appearance scorer used instance-visualization colours and was excluded. Every reported natural association result uses the replacement XYZ-only scorer.
6. The primary natural utility interface is category-blind. The category-equality gate and coherent-matching utility analysis were specified after sealed-label inspection, so their results are descriptive. The primary category-blind result shows no advantage over always unresolved.
7. ConceptGraphs produces only three evaluable associations under its released configuration. We retain the released thresholds and treat the comparison as a coverage audit. Evidence-scoped reporting is likewise a scorer-side audit, not a deployable association method.
8. The RGB-D experiment uses anonymous oracle masks in its main condition and deterministic connected components as a sensitivity. Simulator IDs, classes, event labels, event times, and segmentation colours remain outside the mapper, while known poses and fixed cameras omit pose estimation, exploration, and robot execution.
9. Several semantic corrections followed held-out access. They enforce explicit realization ownership, separate tracker birth from physical birth, make entity-role normalization and partial-output matching injective, and bind each history query to its subject. Evidential F1, world-balanced HOTA/IDF1, method selection, and the hard-decision result are unchanged. Corrected query counts, unsupported-query rates, and correlations are retrospective.
10. A metadata migration corrected split labels in scorer rows and found no source or house overlap. The conventional fixed projection is retained only as an offline sensitivity after prediction commitment; the analytic per-world result is independent of this projection.
11. Scorer truth is released for reproducibility, so the benchmark is not a perpetual hidden leaderboard. A new confirmatory claim requires a new hidden, versioned cohort.