ROOMER: LEARNING WHAT TO SEE AND WHEN TO STOP FOR ROOM RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern room re-identification pipelines extract increasingly rich global, local, fine-grained, and geometric evidence, but still process every query through a rigid cascade. This uniform execution wastes computation on already-resolved queries and can even overturn correct decisions when later evidence is noisy. We present ROOMER, a multi-turn vision-language agent that adaptively allocates test-time evidence, deciding at each turn whether to commit, verify candidates geometrically, or inspect them semantically. To equip the agent with a reliable semantic eye, we introduce **Sight-OPSD**, which uses training-only regional and geometric teaching views to teach a 4B raw-image Inspector *what to see*: which cross-view evidence supports or refutes shared room identity. The resulting Inspector requires no privileged inputs at inference and also serves as ROOMER's policy backbone. We train the Router with cost-efficient multi-turn group-relative credit assignment. Beyond trajectory-level Outcome Credit for final correctness, localized Commitment Credit supervises when to stop, while Probe Credit uses evidence displacement to assign credit to semantic inspections that actually improve the evidence state. This localized supervision mitigates an *afraid-to-commit* collapse in which the policy keeps acquiring evidence instead of answering, while discouraging uninformative use of the expensive semantic Inspector. ROOMER improves both recognition and evidence efficiency: Sight-OPSD raises the 4B Inspector's query-weighted balanced accuracy from 87.80% to 91.12%, exceeding the 397B Qwen3.5 model by 1.46 points, while the full agent reaches 88.79% Recall@1 on AirRoom-Contested, 10.27 points above the fixed AirRoom Full pipeline. On the natural AirRoom distribution, ROOMER further reaches 96.46% Recall@1, a 2.19-point improvement over AirRoom Full. Compared with correctness-only routing, localized credit further improves contested Recall@1 by 1.59 points while using 38.8% less tool budget—a Pareto gain showing that ROOMER can perform better by looking less. The code will be made publicly available upon acceptance.

1 INTRODUCTION

Room re-identification (ReID) asks whether a newly observed image depicts a previously seen physical room despite substantial changes in viewpoint, illumination, occlusion, and movable content. The challenge is not merely to recognize a room category or retrieve a visually similar scene: different rooms within the same building may share nearly identical furniture and layouts, while two views of the same room may overlap only in a few stable local structures. Consequently, physical identity can hinge on small cross-view correspondences that are easily overwhelmed by whole-image appearance similarity.

To expose this challenging regime, we construct **AirRoom-Contested**, a diagnostic subset containing queries with strong retrieval ambiguity or disagreement between complementary evidence sources. Conventional representations degrade sharply in this regime: even a strong DINOv2 global descriptor (Oquab et al., 2024) falls below 35% Recall@1. More revealingly, a 397B-scale vision–language model still struggles to verify room identity from the raw query–candidate pair, yet improves markedly when explicit local correspondences, discrepancies, or geometric consistency are exposed. This gap suggests that the bottleneck is not model capacity alone, but discovering *which cross-view evidence is diagnostic*. The model must learn not merely how to compare two images, but *what to see*.

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.

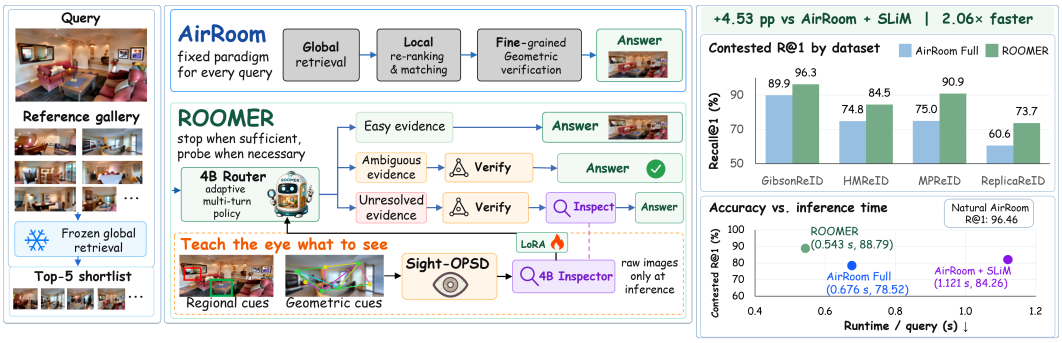


Figure 1: **ROOMER learns to allocate evidence adaptively.** Starting from the same retrieved shortlist, fixed pipelines process every query through the same stages, whereas ROOMER commits directly on easy queries and selectively invokes geometric verification or semantic inspection on ambiguous cases. The right panels summarize the resulting accuracy and inference-efficiency gains.

ROOMER first equips a compact **Inspector** with a reliable semantic eye. We introduce **Sight-OPSD** (Sight-Guided On-Policy Self-Distillation), which teaches a 4B raw-image student *what to see*. An EMA teacher receives training-only regional cues that expose identity-supporting correspondences or discrepancies and geometric cues that reveal cross-view structural consistency, while the student sees only the original image pair. After distillation, both the teacher and these teaching views are discarded, leaving a lightweight Inspector that internalizes their guidance without privileged inputs at inference.

A strong Inspector, however, should not be invoked uniformly. Existing room-ReID systems such as AirRoom (Yao et al., 2025) acquire increasingly expensive evidence through a fixed coarse-to-fine cascade, even when the current evidence is already sufficient; noisy later evidence may also overturn a correct hypothesis. ROOMER instead formulates room re-identification as a multi-turn decision process in which a Router either commits to a candidate, invokes geometric verification, or semantically inspects a selected candidate, allocating computation according to the ambiguity of each query.

Terminal correctness (Shao et al., 2024) alone is too coarse for learning such multi-turn behavior: successful trajectories may contain unnecessary tool calls, while failed trajectories may still contain locally useful decisions. This ambiguity is particularly costly in our setting because inference efficiency is governed by two local decisions: whether the current evidence is already sufficient to stop, and, if not, which candidate is worth spending an expensive semantic inspection on. We therefore retain final correctness as trajectory-level *Outcome Credit*, but complement it with two localized signals. *Commitment Credit* supervises the stop-or-continue decision, rewarding a timely commitment while assigning negative credit specifically to unnecessary continuation beyond a correct stopping opportunity. *Probe Credit* evaluates each semantic inspection by the evidence progress it produces for the ground-truth candidate relative to its competitors, allowing an informative probe to receive positive credit even if a later decision causes the overall trajectory to fail, and a misleading probe to be penalized even within an eventually successful trajectory. Together, they teach the Router not merely to reach the correct answer, but to stop at the right time and probe the right candidate.

Our experiments support both stages of this formulation. The distilled 4B Inspector substantially improves verification on AirRoom-Contested, surpassing the 397B-scale model and reaching the level of GPT-4o. When embedded into the full agent, these gains translate to strong room re-identification performance on both AirRoom-Contested and the natural AirRoom distribution. When embedded into the full agent, the Inspector is no longer treated as a mandatory final stage: ROOMER learns to reserve semantic inspection for only a small fraction of genuinely ambiguous queries. ROOMER also improves runtime over strong fixed pipelines by selectively escalating only unresolved queries, showing that room ReID can benefit from *looking less, but looking where it matters*.

Our contributions are fourfold:

- We introduce ROOMER, an adaptive multi-turn room re-identification agent that replaces fixed evidence processing with learned decisions of when to commit and when to acquire additional evidence.

- We introduce **AirRoom-Contested**, a diagnostic evaluation subset that concentrates on challenging room-ReID queries with strong retrieval ambiguity or conflicting evidence, enabling focused evaluation of physical-room discrimination beyond conventional full-distribution testing.
- We develop **Sight-OPSD**, which teaches a 4B raw-image Inspector what to see by distilling complementary regional and geometric cues from training-only teacher views.
- We introduce multi-turn group-relative credit assignment with trajectory-level **Outcome Credit** and localized **Commitment** and **Probe Credits** to learn when to stop and where to inspect, improving recognition while reducing unnecessary evidence acquisition.

2 Related Work

Room Re-identification and Visual Place Recognition. Visual place recognition (VPR) matches the same place across viewpoint and appearance changes using global descriptors (Arandjelović et al., 2016; Qiu et al., 2024; Lu et al., 2024b; Izquierdo & Civera, 2024b), relation-aware representations (Lu et al., 2024a), and hard-sample mining (Izquierdo & Civera, 2024a; Chen et al., 2026). Indoor localization and room ReID further exploit learned visual descriptors (Sahdev & Tsotsos, 2016; Cabrera et al., 2024), scene text (Tao et al., 2025), and increasingly capable local feature matching, from SuperGlue and LightGlue (Sarlin et al., 2020; Lindenberger et al., 2023) to more recent matchers such as SLiM (Choo & Li, 2026), to resolve fine-grained geometric ambiguity. AirLoc (Li et al., 2023) combines object appearance with inter-object geometry, while AirRoom (Yao et al., 2025) integrates global, object-level, and keypoint evidence in a fixed coarse-to-fine pipeline for room identity determination. ROOMER is complementary to advances in individual matchers: rather than uniformly applying increasingly strong evidence modules, it learns a multi-turn policy that decides when the current evidence is sufficient and when additional geometric or semantic evidence should be acquired.

On-Policy Distillation. On-policy distillation (OPD) supervises student-generated trajectories, reducing the mismatch introduced by fixed teacher trajectories (Agarwal et al., 2024; Gu et al., 2024). OPSD (Zhao et al., 2026) further constructs teacher advantage from privileged reasoning contexts, allowing the student to internalize contextual advantages without relying solely on model-scale differences. Recent work extends this principle to multi-turn and step-aware supervision (Zhou et al., 2026; Zhang et al., 2026b) and to multimodal tasks, where privileged visual contexts improve fine-grained perception and grounding (Yuan et al., 2026; Zhang et al., 2026a; Liu et al., 2026; Yoon et al., 2026). **Sight-OPSD** instead targets cross-view identity discrimination: regional and geometric teaching views guide the student toward the correspondences, discrepancies, and structural consistency that determine whether two observations share the same physical room.

Agentic Reinforcement Learning. Group-relative reinforcement learning and its variants have become widely used for large-model reasoning and agent training (Shao et al., 2024; Yu et al., 2026; Liu et al., 2025), including multi-turn search, long-horizon interaction, and visual tool use (Jin et al., 2025; Wang et al., 2025; Zheng et al., 2026b). Because terminal rewards provide weak attribution over long trajectories, recent methods introduce step-level rewards, local relative advantages, information-gain objectives, or other localized signals for intermediate decisions (Yu et al., 2025; Feng et al., 2026; Wang et al., 2026; Tian et al., 2026; Zheng et al., 2026a). ROOMER specializes this problem to cost-adaptive room ReID: trajectory-level Outcome Credit evaluates final correctness, while localized Commitment and Probe Credits supervise the two decisions that directly govern evidence cost—when to stop and which candidate to inspect.

3 ROOMER: Adaptive Evidence Reasoning

3.1 Overview

Room re-identification asks which room of a gallery a query image q was captured in, where each room is represented by a handful of reference views. A frozen global retriever produces a shortlist $\mathcal{C}(q) = \{c_1, \ldots, c_K\}$ ordered by cosine similarity, after which ROOMER reasons over candidate identities using two stages (Fig. 2). Stage I learns a correspondence-aware *Inspector*: training-only regional and geometric evidence is distilled into backbone parameters θ, so the deployed verifier

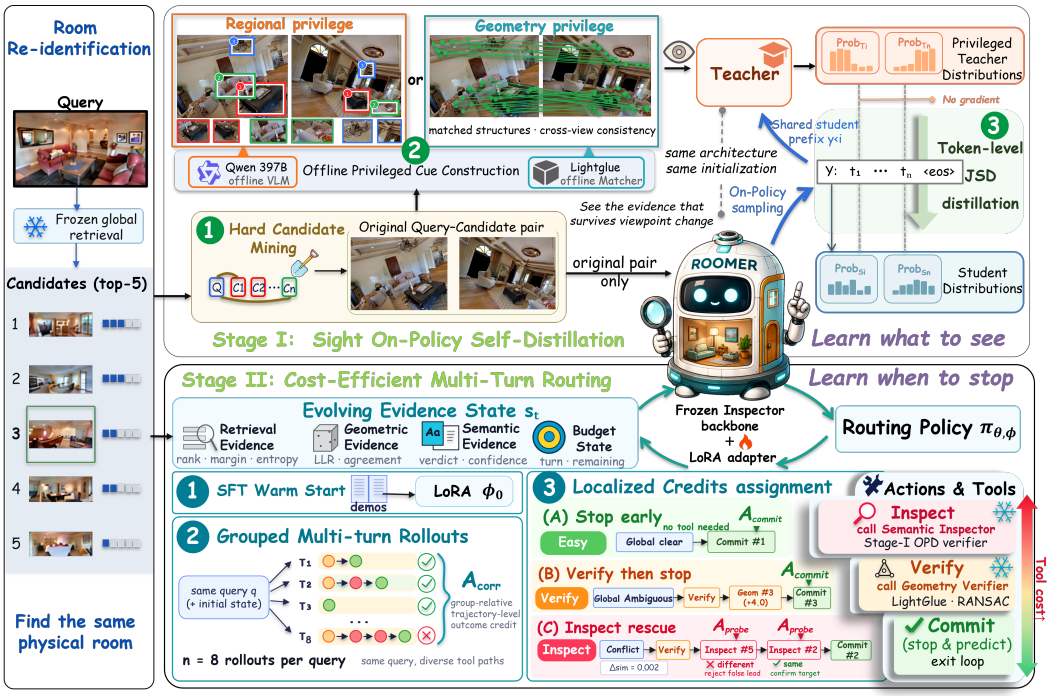


Figure 2: **ROOMER overview.** Stage I distills training-only regional and geometric evidence into a raw-image 4B Inspector using a same-architecture EMA teacher. Stage II freezes this backbone and trains only a LoRA routing adapter over an evolving retrieval, geometric, and semantic evidence state. The Router selectively commits, verifies, or inspects, with multi-turn training combining trajectory-level Outcome Credit and localized Commitment and Probe Credits.

operates directly on a raw query–candidate pair. Stage II freezes θ and trains a routing adapter ϕ over an evolving evidence state, allowing $\pi_{\theta,\phi}$ to decide whether to commit, invoke geometric verification, or inspect a candidate semantically. The two stages therefore share the same visual backbone: with the adapter attached the model routes evidence acquisition, while without it the frozen backbone serves as the Inspector.

3.2 Learning What to See

Stage I learns a binary cross-view *Inspector* that maps a raw query–candidate pair $x = (I_q, I_c)$ to `same` or `different`. Its goal is not simply to perceive more detail, but to develop the right *sight* for physical identity: recognizing which local correspondences remain stable across views and which discrepancies rule out a shared room. We train this semantic eye with **Sight-OPSD**, using regional and geometric teaching views only during training while keeping the student's input unchanged.

Training Pair Construction. Each training query contributes one positive pair with a reference view from its ground-truth room $r^\star$ and one hard negative pair. The negative is the strongest non-ground-truth impostor exposed by global retrieval, local matching, or geometric verification, prioritizing candidates that outrank $r^\star$ in any channel; when none does, we use the most similar non-ground-truth candidate under global retrieval. This yields balanced supervision over challenging same-room variation and visually plausible different-room alternatives.

Teaching Views. Sight-OPSD provides the teacher with two complementary visual views of each hard pair. The *regional view* highlights identity-supporting correspondences for positive pairs and decisive local discrepancies for negatives, teaching *where* the comparison should be grounded. The *geometric view* exposes verified keypoint correspondences and their spatial consistency, teaching whether the observed structures can coexist under a shared physical scene. We denote an available training-only view by $\varphi \in \mathcal{V}(x, y)$, with $\mathcal{V}(x, \textsf{same}) = \{\varphi^{+}_{\text{reg}}(x), \varphi_{\text{geo}}(x)\}$ and $\mathcal{V}(x, \textsf{different}) = \{\varphi^{-}_{\text{reg}}(x), \varphi_{\text{geo}}(x)\}$. The student never observes these views.

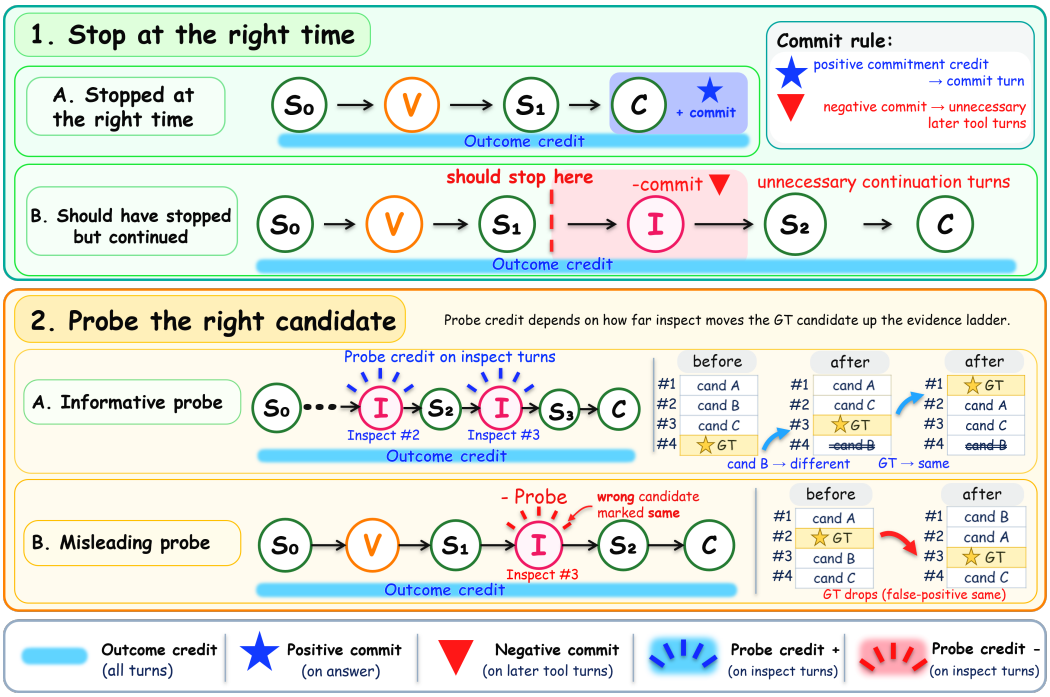


Figure 3: **Credit placement over grouped multi-turn trajectories.** Outcome Credit is broadcast over the completed trajectory. Commitment Credit lands on the commit turn when the accumulated evidence first places the ground-truth room in the decision position, or on tool calls that continue beyond that point. Probe Credit is attached only to inspection turns according to the resulting movement of the ground-truth candidate on the evidence ladder.

Sight-OPSD. Rather than asking the student to reproduce an offline teacher response, Sight-OPSD lets the raw-image student expose its own current mistakes and asks a better-informed version of the same model how those responses should change when the decisive visual cues are made explicit. Given a raw pair x, the student samples $u_{1:N} \sim p_\theta(\cdot \mid x)$. A same-architecture EMA teacher with parameters $\bar{\theta}$ observes a teaching view $\varphi \in \mathcal{V}(x, y)$, and both models score the same student-generated prefix $u_{<t}$,

$$p^t = p_\theta(\cdot \mid x, u_{<t}), \qquad \bar{p}^t = p_{\bar{\theta}}(\cdot \mid \varphi(x), u_{<t}). \tag{1}$$

Because architecture and response history are shared, their disagreement isolates how access to the relevant visual cues changes the model's interpretation. With $m^t = \frac{1}{2}(p^t + \bar{p}^t)$, we minimize the clipped token-level Jensen–Shannon divergence

$$\mathcal{L}_{\text{Sight}} = \frac{1}{N} \sum_{t=1}^{N} \min\left(\frac{1}{2}\mathrm{KL}(p^t \,\|\, m^t) + \frac{1}{2}\mathrm{KL}(\bar{p}^t \,\|\, m^t),\ \epsilon \right). \tag{2}$$

The teacher follows the student by exponential moving average and receives no gradient. Training therefore distills not an additional inference module, but a better visual habit: after Sight-OPSD, the teacher and teaching views are removed, and the Inspector must recover the relevant identity cues directly from the raw pair.

3.3 Learning When to Stop

Cost-Efficient Multi-Turn Routing. With the Stage-I Inspector frozen, we train only a LoRA routing adapter to allocate test-time computation. At turn t, the Router observes an evolving state s_t containing retrieval statistics, available geometric and semantic evidence, and the remaining budget, and chooses $a_t \in \{\texttt{commit}(k), \texttt{verify}, \texttt{inspect}(k)\}$. $\texttt{commit}(k)$ selects candidate k and terminates; $\texttt{verify}$ runs the frozen Geometry Verifier over the shortlist and returns a calibrated geometric LLR for each candidate; and $\texttt{inspect}(k)$ applies the frozen Stage-I Inspector to one

query–candidate pair and returns a semantic same-room verdict. Tool outputs are appended to s_t for subsequent decisions. We warm-start the adapter with state-grounded demonstrations and then optimize grouped multi-turn rollouts sampled from the same query. Final correctness alone cannot distinguish whether a successful trajectory stopped efficiently or whether an intermediate probe was locally useful. We therefore retain correctness as trajectory-level *Outcome Credit* and add two localized behavioral signals: *Commitment Credit* for the stop-or-continue decision and *Probe Credit* for candidate-level semantic inspection. Figure 3 illustrates this distinction between trajectory-level Outcome Credit and localized Commitment and Probe Credits.

Outcome Credit. We retain final correctness as the trajectory-level group-relative signal. For a group $G(q)$ of rollouts sampled from the same query,

$$A_i^{\text{corr}} = \text{corr}_i - \frac{1}{|G(q)|} \sum_{j \in G(q)} \text{corr}_j, \tag{3}$$

where $\text{corr}_i \in \{0, 1\}$ denotes final room correctness. We use the group mean without standard-deviation normalization. Because Outcome Credit evaluates the completed trajectory, it is broadcast over all decision turns; it indicates whether the episode succeeds, but not which local decisions used computation well.

Commitment Credit: Stop at the Right Time. Outcome Credit does not reveal whether a correct trajectory stopped as soon as further computation became unnecessary. We therefore introduce *Commitment Credit* to supervise this local stop-or-continue decision. Let $\hat{c}_{i,t}$ denote the current leader under the accumulated evidence ordering and let $c^\star$ be the ground-truth room. We define

$$t_i^\star = \min\{t : \hat{c}_{i,t} = c^\star\}, \tag{4}$$

the earliest decision point at which the evidence already acquired places the correct room in the decision position. The binary rollout attribute commit_i records whether rollout i commits at this correct stopping opportunity rather than escalating further, yielding

$$A_i^{\text{commit}} = \beta \left(\text{commit}_i - \frac{1}{|G(q)|} \sum_{j \in G(q)} \text{commit}_j \right). \tag{5}$$

Ground truth is used only to construct this training-side signal: it is never included in the Router state, and both the rationale and action are generated solely from the observable evidence s_t.

Although A_i^{commit} is computed from a rollout-level attribute, its learning signal is localized. Positive Commitment Credit is assigned only to the terminating `commit` turn. If a rollout continues after $t_i^\star$, negative credit is assigned only to the subsequent tool calls:

$$\mathcal{K}_i = \begin{cases} \{T_i\}, & A_i^{\text{commit}} > 0, \\ \{t \mid t \geq t_i^\star,\ t < T_i,\ a_{i,t} \in \{\texttt{verify}, \texttt{inspect}(k)\}\}, & A_i^{\text{commit}} < 0. \end{cases} \tag{6}$$

Thus Commitment Credit is computed from a trajectory property but delivered only to the decisions responsible for timely stopping or unnecessary continuation.

Probe Credit: Probe the Right Candidate. When additional evidence is necessary, semantic inspection is the most expensive action, so the Router must learn which candidate is worth probing. We evaluate each inspection by the evidence progress it produces rather than by the eventual trajectory outcome. For the frozen candidate set $\mathcal{C}(q)$, candidates are ordered lexicographically by

$$\kappa_t(c) = \Big(\text{tier}_t(c), \mathbf{1}[\text{geometry available}]\, \text{llr}_t(c), \cos(q, c) \Big), \tag{7}$$

where a positive semantic verdict raises a candidate to the highest semantic tier, while a negative verdict does not hard-eliminate it. Let $\rho_{i,t}^-$ and $\rho_{i,t}^+$ be the ground-truth rank immediately before and after an `inspect`(k) action. Its evidence displacement is

$$\Delta_{i,t} = \big(\rho_{i,t}^- - \rho_{i,t}^+\big) + \big(\text{tier}_{i,t}^+(c^\star) - \text{tier}_{i,t}^-(c^\star)\big), \tag{8}$$

where the tier term also rewards new positive confirmation when the ground-truth candidate is already highly ranked. We then compute

$$A_{i,t}^{\text{probe}} = \eta\big(\Delta_{i,t} - \overline{\Delta}_{q,\text{inspect}}\big), \tag{9}$$

where $\overline{\Delta}_{q,\text{inspect}}$ is the mean displacement over valid inspection turns from the same query group. Probe Credit is attached only to the corresponding inspection turn: probes that advance the ground-truth candidate relative to sibling inspections receive positive local credit, whereas misleading probes receive negative credit. Ground-truth identity is used only to construct this training-side signal and is never exposed to the Router; rollouts in which $c^\star \notin \mathcal{C}(q)$ are excluded from this channel.

Combining the trajectory-level outcome with the two localized behavioral credits, the learning advantage for token u at turn t is

$$\boxed{A_{i,u} = A_i^{\text{corr}} + \mathbf{1}[t \in \mathcal{K}_i] A_i^{\text{commit}} + \mathbf{1}[a_{i,t} = \texttt{inspect}(k)] A_{i,t}^{\text{probe}}} \tag{10}$$

A small mechanical action-validity penalty is applied to illegal or contradictory actions. Importantly, no explicit tool-cost term is added to the reward: efficiency emerges from localizing supervision to unnecessary continuation and to the quality of expensive semantic probes.

4 Experiments

4.1 AirRoom-Contested construction

Beyond the full benchmarks, we evaluate a diagnostic regime in which retrieval remains ambiguous or geometric evidence challenges the initial retrieval hypothesis. **AirRoom-Contested** is constructed from the test splits of MPReID, HMReID, GibsonReID, and ReplicaReID using two signals: *retrieval ambiguity*, measured by the top-1/top-2 room-similarity margin with dataset-specific thresholds calibrated only on training-split retrieval errors, and *appearance–geometry disagreement*, identified by a fixed LightGlue-based matcher when geometric verification changes or fails to support the retrieval hypothesis. Test membership depends only on frozen retrieval and matching outputs; it uses neither test labels nor predictions from ROOMER or the full AirRoom pipeline. Full definitions and matcher settings are provided in the appendix.

4.2 Experimental Setup

Evaluation Protocols. We evaluate ROOMER under two complementary protocols derived from the same contested regime. *AirRoom-Contested Verification* isolates the Inspector's ability to distinguish physical-room identity from visually plausible resemblance at the image-pair level, whereas *AirRoom-Contested Re-identification* evaluates the complete retrieval-and-decision problem over the room gallery. For verification, we use 4,718 contested queries, each paired with one reference from its ground-truth room and one hard negative, yielding 9,436 balanced pairs. All models receive exactly the same two raw images at inference, without access to any training-only teaching views. We report balanced accuracy (BA) for each dataset and a query-weighted aggregate across the four datasets (All). For re-identification, we use scene-disjoint training and test partitions, with queries retrieving only from galleries in the corresponding partition so that all test-room identities remain unseen during training. All methods operate on identical test queries and galleries. We report room-level Recall@1 (R@1) on both AirRoom-Contested and the natural AirRoom distribution.

Baselines. For pair verification, we compare the Sight-OPSD Inspector with unadapted Qwen3.5 models from 4B to 397B and representative closed-source MLLMs. For room ReID, we compare against DINOv2, Patch-NetVLAD, AnyLoc, SAGE, and successive AirRoom stages (Global, Global+Local, and Full). We additionally replace AirRoom's original matcher with SLiM to form a stronger fixed-pipeline baseline, separating the benefit of adaptive routing from that of simply upgrading the matcher.

Implementation Details. The Inspector is initialized from Qwen3.5-4B and trained with full-parameter Sight-OPSD. We then freeze the backbone and train a rank-32 LoRA routing adapter by supervised warm-start followed by multi-turn group-relative optimization. Policy training uses eight rollouts per query, 120 optimization steps, and a learning rate of 10^{-5}; The retriever, Geometry Verifier, and Inspector remain frozen throughout Router training. Further details are provided in the appendix.

Table 1: Room re-identification Recall@1 (%) on AirRoom-Contested and the natural AirRoom distribution under the same scene-disjoint protocol. All methods use identical queries and galleries within each setting; All is query-weighted across datasets.

Method	AirRoom-Contested					Natural AirRoom Distribution				
	MP	HM	Gibson	Replica	**All**	MP	HM	Gibson	Replica	**All**
DINOv2 (TMLR'24)	35.18	25.09	42.28	39.81	34.17	60.72	53.46	62.35	76.44	59.42
Patch-NetVLAD (CVPR'21)	43.66	33.28	50.92	41.36	41.99	65.58	64.13	63.24	62.41	64.45
AnyLoc (RA-L'23)	70.70	58.10	86.50	69.59	71.08	92.34	89.69	85.85	88.57	90.04
AirRoom Global (CVPR'25)	70.72	58.09	86.50	69.59	71.08	95.84	88.72	93.45	88.67	92.40
AirRoom Local (CVPR'25)	73.59	66.51	88.93	69.83	75.62	96.18	90.81	94.63	88.67	93.49
AirRoom Full (CVPR'25)	75.02	74.83	89.92	60.58	78.52	96.45	93.02	95.11	85.57	94.27
SAGE (ICLR'26)	82.52	76.60	90.14	71.53	81.85	96.85	93.34	95.71	87.40	94.77
AirRoom + SLiM (CVPR'26)	85.04	79.35	92.16	**73.72**	84.26	97.06	93.52	95.85	87.72	94.96
ROOMER (Ours)	**90.87**	**84.46**	**96.34**	**73.72**	**88.79**	**98.37**	**95.12**	**97.42**	**89.54**	**96.46**

Table 2: Cross-view verification on AirRoom-Contested (4,718 queries; 9,436 balanced pairs). We report dataset-wise balanced accuracy (%) and the query-weighted aggregate (All). Every model receives the same two raw images.

Model	MP	HM	Gibson	Replica	All ↑
Closed-source MLLMs					
Gemini 3.7 Flash	96.06	91.51	95.22	90.02	93.59
GPT-5.6 Sol	95.10	90.65	94.24	88.84	92.64
GPT-4o	93.18	88.03	92.28	86.70	90.43
Open-weight MLLMs					
Qwen3.5-4B	90.87	84.46	91.62	79.93	87.80
Qwen3.5-9B	91.26	86.13	92.24	79.22	88.64
Qwen3.5-27B	91.83	86.88	92.60	79.57	89.19
Qwen3.5-397B-A17B	92.22	87.46	92.96	80.17	89.66
Sight-OPSD 4B (Ours)	**94.91**	**88.06**	**93.49**	**85.64**	**91.12**

4.3 MAIN RESULTS

Room re-identification on AirRoom-Contested. Table 1 shows that ROOMER reaches **88.79%** R@1 on AirRoom-Contested, improving over AirRoom Full by **10.27** points and AirRoom+SLiM by **4.53** points. Upgrading the matcher alone already raises the fixed AirRoom pipeline from 78.52% to 84.26%, but a further gap remains to ROOMER, showing that adaptive evidence acquisition provides gains beyond stronger matching. ROOMER also reaches **96.46%** R@1 on the natural AirRoom distribution, exceeding AirRoom Full and AirRoom+SLiM by **2.19** and **1.50** points, respectively. The larger improvement on AirRoom-Contested indicates that adaptive routing is especially beneficial when the initial retrieval remains unresolved, while preserving strong performance on the conventional distribution.

Cross-view identity verification on AirRoom-Contested. As shown in Table 2, the 4B Sight-OPSD Inspector reaches **91.12%** query-weighted balanced accuracy using only raw image pairs, improving over the unadapted 4B backbone by **3.32** points and the 397B Qwen3.5 model by **1.46** points, while reaching a level comparable to GPT-4o. These results show that teaching a compact model what cross-view evidence matters can provide substantial gains beyond model scaling alone.

4.4 ABLATION STUDIES

What teaches the Inspector to see? Table 3 separates hard-pair supervision from the two Sight-OPSD teaching views. Pair SFT provides only a modest gain over the frozen 4B backbone, whereas adding

Table 3: Ablation of Stage-I **Sight-OPSD**. All variants use the same hard training pairs and receive only raw image pairs at inference. "Reg." and "Geo." denote the regional and geometric teaching views available only during training.

Training	MP	HM	Gibson	Replica	All ↑
Frozen Qwen3.5-4B	90.87	84.46	91.62	79.93	87.80
Pair SFT	91.65	85.30	92.10	81.10	88.54
Sight-OPSD (Reg. only)	93.10	87.20	92.80	84.15	90.05
Sight-OPSD (Geo. only)	93.55	86.62	92.93	83.10	89.89
Sight-OPSD (Reg.+Geo.)	**94.91**	**88.06**	**93.49**	**85.64**	**91.12**

Table 4: Ablation of localized routing credits on AirRoom-Contested. Outcome is trajectory-level correctness; Commit and Probe localize supervision to stopping and semantic inspection, respectively.

Training objective	Outcome	Commit	Probe	R@1 ↑	Budget ↓	Inspect ↓
Correctness-only GRPO	✓			87.20	4.10	42.0%
+ Commitment Credit	✓	✓		87.56	2.75	21.5%
+ Probe Credit	✓		✓	87.43	3.70	34.0%
Full ROOMER	✓	✓	✓	**88.79**	**2.51**	**18.8%**

either regional or geometric guidance yields a substantially stronger verifier. Their combination performs best: regional cues teach where identity-supporting correspondences or discrepancies lie, while geometric cues teach whether those structures remain physically consistent across views.

Effect of localized multi-turn credit. Table 4 shows complementary effects from the two localized credits. Commitment Credit sharply reduces unnecessary escalation, while Probe Credit improves where semantic inspection is allocated. Combining both raises R@1 from 87.20% to **88.79%** while reducing the average tool budget from 4.10 to **2.51** (**38.8%**) and the inspection rate from 42.0% to **18.8%**. This supports localizing supervision to both stopping and probing decisions rather than relying on trajectory correctness alone.

4.5 Efficiency and Behavioral Analysis

Table 5 shows that ROOMER requires only 0.543 s per query, reducing latency by 19.7% relative to AirRoom Full (0.676 s) and running 2.06× faster than AirRoom+SLiM (1.121 s). On the natural AirRoom distribution, 61.6% of queries commit directly and only 5.0% require semantic inspection; on AirRoom-Contested, inspection rises to 18.8%, showing that the learned Router concentrates expensive evidence acquisition on unresolved cases.

Table 5: Efficiency and test-time routing behavior. Left: accuracy and average runtime per query under identical hardware. Right: fractions of queries resolved at different evidence levels.

Accuracy–Efficiency

Method	Cont. R@1	Nat. R@1	Time ↓
AirRoom Full	78.52	94.27	0.676 s
AirRoom + SLiM	84.26	94.96	1.121 s
ROOMER	**88.79**	**96.46**	**0.543 s**

Routing Behavior

Distribution	Direct	Geo.	Inspect
Natural	61.6%	33.4%	5.0%
Contested	12.0%	69.1%	18.8%

5 Conclusion

We presented ROOMER, an adaptive multi-turn approach to room re-identification that learns what to see and when to acquire additional evidence. Sight-OPSD distills training-only regional and geometric cues into a deployable 4B Inspector, while localized Outcome, Commitment, and Probe Credits train the Router to allocate evidence efficiently. Experiments on AirRoom-Contested and the natural distribution show improved recognition and inference efficiency over strong fixed pipelines.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, language and style editing, translation, feedback on research methodology and experimental design, interpretation and presentation of experimental results, and refinement of scientific figures. All AI-assisted outputs were reviewed and, where applicable, independently verified by the authors. The authors take full responsibility for the final content, including all technical claims, experimental results, analyses, and references.

ETHICS STATEMENT

This work uses existing indoor room-reidentification datasets and does not involve human-subject intervention. Indoor imagery may nevertheless contain privacy-sensitive visual information. Applications of room recognition should therefore follow the licenses and privacy requirements of the underlying data, minimize unnecessary retention of visual records, and avoid identifying occupants or inferring sensitive personal attributes. Deployment in surveillance or other high-impact settings requires appropriate authorization and governance beyond the scope of this work.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.3 specify the model components, action space, Sight-OPSD objective, frozen tools, and localized Router credits. Section 4 defines the pair-verification and room-reidentification protocols. The appendix provides dataset and pair statistics, AirRoom-Contested construction details, training hyperparameters, tool configurations, interaction budgets, checkpoint selection, hardware, prompts, and cost accounting. All train/test partitions are scene-disjoint, with test-room identities absent from training and development.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomáš Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Juan José Cabrera, Vicente Román, Arturo Gil, Oscar Reinoso, and Luis Payá. An experimental evaluation of siamese neural networks for robot localization using omnidirectional imaging in indoor environments: Jj cabrera et al. *Artificial Intelligence Review*, 57(8):198, 2024.

Shunpeng Chen, Changwei Wang, Rongtao Xu, Jinzhou Lin, Wenhao Xu, Li Guo, Shibiao Xu, et al. Sage: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *International conference on learning representations*, volume 2026, pp. 28660–28683, 2026.

Sin Wai Choo and Bo Li. Scalable feature matching via state space modeling and sparse correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6685–6694, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In *European Conference on Computer Vision*, pp. 240–257. Springer, 2024a.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17658–17668. IEEE, 2024b.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Bowen Li, Sebastian Scherer, Yun-Jou Lin, Chen Wang, et al. Airloc: Object-based indoor relocalization. *arXiv preprint arXiv:2304.00954*, 2023.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782. IEEE, 2024a.

Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. *Advances in Neural Information Processing Systems*, 37:5789–5816, 2024b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Qibo Qiu, Shun Zhang, Haiming Gao, Honghui Yang, Haochao Ying, Wenxiao Wang, and Xiaofei He. Emvp: Embracing visual foundation model for visual place recognition with centroid-free probing. *Advances in Neural Information Processing Systems*, 37:120928–120950, 2024.

Raghavender Sahdev and John K Tsotsos. Indoor place recognition system for localization of mobile robots. In *2016 13th Conference on computer and robot vision (CRV)*, pp. 53–60. IEEE, 2016.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Huaqi Tao, Bingxi Liu, Calvin Chen, Tingjun Huang, He Li, Jinqiang Cui, and Hong Zhang. Textinplace: Indoor visual place recognition in repetitive structures with scene text spotting and verification. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11081–11088. IEEE, 2025.

Yang Tian, Rui Wang, Xumeng Wen, Junjie Li, Shizhao Sun, Lei Song, Jiang Bian, and Bo Zhao. Pbsd: Privileged bayesian self-distillation for long-horizon credit assignment. *arXiv preprint arXiv:2606.09348*, 2026.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, et al. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, volume 2026, pp. 104502–104529, 2026.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Runmao Yao, Yi Du, Zhuoqun Chen, Haoze Zheng, and Chen Wang. AirRoom: Objects matter in room reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Yuanqing Yu, Zhefan Wang, Weizhi Ma, Shuai Wang, Chuhan Wu, Zhiqiang Guo, and Min Zhang. Steptool: Enhancing multi-step tool usage in llms via step-grained reinforcement learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3952–3962, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Yan Zhang, Daiqing Wu, Huawen Shen, Can Ma, and Yu Zhou. Learn where to click from yourself: On-policy self-distillation for gui grounding. *arXiv preprint arXiv:2605.00642*, 2026a.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning. *arXiv preprint arXiv:2605.27140*, 2026b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yuyao Zheng, Haipeng Sun, Junwei Bao, Lemao Liu, Hongfei Jiang, Yang Song, and Dejing Dou. Pgpo: Potential-guided policy optimization for multi-turn agentic tasks. *arXiv preprint arXiv:2609.02236*, 2026a.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026b.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Sage-opd: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026.

A Data Construction and Evaluation Protocols

A.1 Room identities, shortlists, and evaluation units

The prediction target is a physical room rather than a room category. A query and a gallery image are a positive pair only when they depict the same room; sharing a category, furniture style, or building is insufficient. The retrieval shortlist contains distinct room hypotheses, each associated with its reference image. We retain the initial retrieval handles throughout an episode: `#1` always denotes the initially highest-ranked candidate, even when geometry or inspection favors another room. Tool calls add evidence to this fixed shortlist rather than retrieving new candidates.

We distinguish pair verification from room re-identification. In verification, each selected query contributes a positive pair and a hard negative pair, and the Inspector predicts `same` or `different` independently for each pair. In re-identification, the Router must select one room from the entire shortlist. A good pair classifier does not automatically solve the latter problem: the correct room may be absent, several candidates may receive positive semantic verdicts, or the Router may commit to the wrong candidate despite having acquired useful evidence.

Scene partitions are fixed before training and evaluation, and a query retrieves only from the gallery assigned to its partition. Development data are used for model selection, while test-room identities remain outside training and development. Pair manifests and query manifests are tracked separately so that filtering a teacher form or a malformed pair cannot silently change the room-level evaluation population.

A.2 Contested selection and hard-pair construction

Retrieval ambiguity. For a query q from dataset d, let $s_1(q) \geq s_2(q)$ denote the global-descriptor similarities of the two highest-scoring *distinct rooms*, and define the retrieval margin as

$$m(q) = s_1(q) - s_2(q). \tag{11}$$

We calibrate a dataset-specific ambiguity threshold using only training queries on which global retrieval makes a top-1 error:

$$\tau_d = Q_{0.90}\big(\{m(q) : q \in \mathcal{Q}^d_{\text{train}}, \hat{r}_1(q) \neq r^\star(q)\}\big), \tag{12}$$

where Q_p is the empirical p-quantile, $\hat{r}_1(q)$ is the global-retrieval top-1 room, and $r^\star(q)$ is the ground-truth room. After calibration, τ_d is frozen. A test query enters the retrieval-ambiguity arm when

$$\mathcal{M}_d = \big\{q \in \mathcal{Q}^d_{\text{test}} : m(q) < \tau_d\big\}. \tag{13}$$

Thus, ground truth is used only to calibrate τ_d on the training split; membership of a test query in $\mathcal{M}_d$ depends only on its frozen global-retrieval scores.

Appearance–geometry disagreement. Independently of the ambiguity arm, we apply a fixed LightGlue-based construction matcher to the global top-K shortlist, with $K = 5$ in all experiments. Let $\hat{r}_g(q)$ denote the candidate preferred by this frozen geometric criterion, and let $\text{valid}_{\text{geo}}(q)$ indicate that at least one shortlisted candidate passes the matcher's fixed geometric acceptance criterion. We define

$$\mathcal{D}_d = \big\{q \in \mathcal{Q}^d_{\text{test}} : \hat{r}_g(q) \neq \hat{r}_1(q) \ \vee \ \neg\text{valid}_{\text{geo}}(q)\big\}. \tag{14}$$

The first condition captures disagreement between appearance-based retrieval and geometric evidence, while the second captures cases in which geometry fails to support any retrieved hypothesis. The construction matcher is frozen and is used only to define the evaluation subset; it is distinct from the calibrated online Geometry Verifier used by the Router, whose LLR is described in Section C.

Contested subset. The final AirRoom-Contested split is the union

$$\mathcal{Q}^d_{\text{cont}} = \mathcal{M}_d \cup \mathcal{D}_d. \tag{15}$$

For diagnostic analysis, this union can be decomposed into the three disjoint groups $\mathcal{M}_d \setminus \mathcal{D}_d$, $\mathcal{D}_d \setminus \mathcal{M}_d$, and $\mathcal{M}_d \cap \mathcal{D}_d$. Test-set membership therefore depends only on frozen retrieval and matching outputs: neither test labels, learned ROOMER components, nor predictions from the full AirRoom pipeline participate in query selection.

Hard-pair construction. Pair construction for Inspector training is separate from contested test-set selection. For each training query q, we form one positive pair using a reference view from its ground-truth room $r^\star$. To construct the hard negative, consider all non-ground-truth candidates that outrank $r^\star$ under at least one of the three frozen evidence channels used for mining: global retrieval, local refinement, or geometric verification. Among these candidates, we select the impostor attaining the best rank in any channel; if no non-ground-truth candidate outranks $r^\star$, we instead choose the most similar non-ground-truth candidate under global retrieval. This produces one positive and one hard negative per selected query. The mining ranks and associated difficulty tags are used only for training-pair construction and analysis and are never exposed as additional test-time selection criteria or model inputs.

B SIGHT-OPSD TRAINING DETAILS

B.1 PRIVILEGED VIEWS AND TEACHER-FORM FILTERING

The student always receives the two original RGB images and the fixed raw-image verification prompt. The teacher receives one of the forms in Table 6. Regional evidence identifies corresponding physical parts for positives and discriminative discrepancies for negatives. Geometric evidence exposes spatially consistent correspondences over the image pair. These are training-only visual contexts: the geometric teacher view is not an online call to the Router's Geometry Verifier.

Before training, the base verifier generates a greedy answer under each candidate privileged form. A form survives only when its images are complete, its verdict is parseable, and its verdict agrees with the pair label. A pair can retain either privileged form or both; training samples one surviving form for a visit. Pairs assigned to the raw-image anchor receive no additional visual privilege. Offline annotation models assist asset construction, whereas the model that supplies token-level supervision during optimization is the same-architecture EMA teacher.

B.2 STUDENT ROLLOUTS, EMA UPDATES, AND OUTPUT CONSTRAINTS

For each sampled pair, the student generates one response from its raw-image context. The student and teacher then score exactly this response under their respective visual contexts, using the same response prefix at every token position. The teacher does not provide a separate target transcript. Its parameters are updated only after a successful student optimizer step,

$$\bar{\theta} \leftarrow (1-\mu)\bar{\theta} + \mu\theta, \tag{16}$$

with the teacher detached from backpropagation. In the recorded training recipe, $\mu = 0.05$, so the latest student contributes five percent to each EMA update. Resuming training restores the EMA state as well as the student and optimizer states.

The verification response contains a short explanation followed by the two parseable fields

```
verdict: same | different
confidence: <integer 0-100>
```

The vertical bar denotes the two allowed alternatives, rather than literal output text. A frozen scaffold lexicon constrains student rollout generation and masks teacher-distribution tokens associated with annotation scaffolds, such as boxes, crops, and matching lines. This discourages references to privileged visual annotations that are absent from the student's input. Parsing and scaffold checks are separate from semantic correctness.

B.3 MEMORY-EFFICIENT DISTRIBUTION MATCHING

At response position t, let $\mathcal{S}_t$ contain the student's top-k_{vocab} vocabulary items. Both distributions are evaluated on this same support; remaining probability mass is collected into one tail event:

$$\begin{aligned} \widetilde{p}^{\,t} &= \big((p^t(v))_{v\in\mathcal{S}_t},\ 1-\sum_{v\in\mathcal{S}_t} p^t(v)\big), \\ \widetilde{\bar{p}}^{\,t} &= \big((\bar{p}^t(v))_{v\in\mathcal{S}_t},\ 1-\sum_{v\in\mathcal{S}_t} \bar{p}^t(v)\big). \end{aligned} \tag{17}$$

Table 6: Teacher contexts in Stage I. Internal form names identify the corresponding assets; all student contexts contain only the raw image pair.

Teacher form	Visual input	Supervision role
Regional (`crop8`)	Two boxed full images and three paired regional crops (eight images in total).	Local correspondence or discrepancy.
Geometric (`D1`)	Two boxed full images and a side-by-side RANSAC-inlier visualization.	Cross-view structural consistency.
Raw-image EMA anchor	The same two unannotated images seen by the student.	Temporal smoothing without extra visual privilege.

The recipe uses $k_{\text{vocab}} = 256$. Teacher probabilities use a full-vocabulary normalization evaluated in chunks, rather than renormalizing only the selected entries. The tail therefore preserves omitted probability mass, although its internal distribution is not resolved by this compressed representation.

With w_t indicating an eligible response position, the implemented masked objective is

$$\mathcal{L}_{\text{Sight}} = \frac{\sum_t w_t \min\{\text{JSD}(\widetilde{p}^t, \widetilde{\widetilde{p}}^t), \epsilon\}}{\max(1, \sum_t w_t)}. \tag{18}$$

The JSD mixing weight is one half, as in Eq. 2. The active Stage-I objective is distribution matching on student-generated responses, not a GRPO outcome loss or cross-entropy against an offline teacher transcript. Answer filtering uses pair labels to select reliable contexts; it does not turn those labels into the token-level training target.

B.4 Coverage and optimization

One coverage pass visits every unique training pair, choosing one of its available teacher forms. Deterministic weighted repeats fill the final batch, and a fixed seed controls the shuffle. Dataset weights proportional to $\sqrt{N_d}$ and balanced within-dataset positive/negative targets govern weighted sampling decisions. The coverage constraint is applied first: these weights do not imply that mandatory visits themselves have exactly balanced empirical counts, nor that every teacher form is visited once.

Stage I updates all student parameters with AdamW. The recorded recipe uses learning rate 10^{-6}, weight decay 0.01, cosine decay, warmup ratio 0.03, gradient clipping at 1.0, global batch size 64, and per-token JSD clipping at 0.05. Training uses bf16, fully sharded data parallelism, gradient checkpointing, and four A100 GPUs. Student rollout decoding uses temperature 1.0, top-$p = 0.95$, one response per pair, and at most $1{,}024$ new tokens, with thinking disabled.

C Router Observations, Actions, and Frozen Tools

C.1 Observation construction and anonymity

Each decision receives the query and five candidate thumbnails, together with a textual evidence state. Retrieval fields include each candidate's cosine score, the leading score margin, a score ratio, an entropy summary, and adjacent rank gaps. Geometry fields include correspondence inliers, coverage, spatial-cluster counts, calibrated LLRs, the geometric leader, and whether it agrees with the retrieval leader. Inspection fields include the inspected handle, the parsed verdict, a calibrated confidence percentile, and a short note. The state also exposes remaining budget, the current turn, and any protocol error from the preceding action.

Tool fields remain unavailable until the corresponding action has been executed. Candidate handles retain their original meaning; neither the ground-truth rank nor the training-side evidence displacement is inserted into the observation. Dataset names, room identifiers, file paths, and other private identifying strings are redacted from rendered text and tool notes. The confidence percentile returned by the Inspector is a calibrated score percentile, not a posterior probability that the selected room is correct.

C.2 Action serialization and episode termination

The paper's actions map to the executable protocol as follows:

$$\texttt{commit}(k) \leftrightarrow \texttt{answer \#k}, \quad \texttt{verify} \leftrightarrow \texttt{geometry}, \quad \texttt{inspect}(k) \leftrightarrow \texttt{inspect \#k}. \tag{19}$$

One response contains one non-empty rationale and one action, serialized as

```
<why>brief evidence-grounded rationale</why><act>action</act>
```

The rationale is limited to 64 tokens. It is part of the generated decision turn, but its readability is not evidence of causal faithfulness.

The selected handle must belong to the frozen shortlist. Geometry can be called at most once, and each candidate can be inspected at most once. These constraints do not impose a mandatory geometry-before-inspection cascade. Once the interaction budget or turn limit is exhausted, only an answer is legal. An illegal action is recorded and returned as an error in the next observation; repeated illegal behavior or exhaustion triggers deterministic termination.

C.3 Geometric and semantic evidence

The recorded Geometry Verifier uses SuperPoint features with at most 2,048 keypoints, LightGlue matching, and fundamental-matrix RANSAC with a three-pixel threshold and confidence 0.99. A 3×3 query-image grid summarizes the spatial distribution of inliers. One call processes every candidate, rather than just the currently preferred room. Correspondence statistics are calibrated using pooled training data to produce an LLR clipped to $[-4, 4]$. A positive LLR favors same-room evidence but is not a guarantee of physical identity. The raw matching configuration and the fitted calibration are frozen during Router optimization.

Semantic inspection runs the Stage-I model without the routing adapter on the raw query and one selected candidate. The recorded decoding configuration uses greedy generation, temperature zero, at most 512 new tokens, two images per request, and thinking disabled. A parser extracts the final verdict, and the calibrated confidence and redacted note are appended to the state. The Inspector does not receive regional crops, matching-line visualizations, ground-truth identities, or the Router's training credits. Its weights remain fixed while the routing adapter learns which pairs to inspect.

D Router Training and Controlled Comparisons

D.1 State-grounded supervised warm start

Supervised demonstrations are generated by three behavior policies: direct stopping from retrieval, geometry followed by a decision, and sequential inspection after geometry. The inspection policy visits candidates using observed evidence, stopping after a positive verdict or when the budget requires an answer. These behavior policies are label-blind during action generation. Ground truth is used only afterward to retain successful demonstrations and attach audit tags.

Each demonstration rationale refers to evidence available in that turn, including remaining budget when it motivates stopping. Neither a future tool result nor an unseen label can justify the recorded action. The supervised stage teaches the response protocol, evidence reading, and budget-aware termination. Reinforcement learning starts from this routing adapter while the distilled backbone and both tools remain frozen.

D.2 Behavior strata and grouped policy optimization

Training uses the five strata in Table 7. Their labels control sampling frequency and do not enter the observation or reward computation. Fixed weights are converted to integer batch quotas by largest-remainder allocation with carry across batches. This preserves the intended long-run proportions without requiring every batch to represent fractional quotas exactly. Development and test data retain their evaluation distributions rather than adopting the training mixture.

Table 7: Behavior strata for Router training. The categories describe outcomes of frozen evidence sources and identify complementary routing skills.

Stratum	Defining situation	Target behavior
EASY	Non-contested query; retrieval and geometry are correct.	Commit directly when evidence suffices.
CONFLICT_G	Retrieval is correct in a contested or over-processing-prone case.	Preserve a useful retrieval decision.
GEO_FIX	Retrieval is wrong and geometry is correct.	Acquire geometry and stop appropriately.
INSPECT_TRAP	Semantic inspection can favor a confusable impostor.	Avoid being misled by an extra probe.
INSPECT_FIX	Retrieval and geometry are wrong, but semantic evidence can resolve the query.	Escalate and inspect useful candidates.

The routing adapter has LoRA rank 32 and scaling parameter 64. The recorded RL recipe samples eight rollouts for each query, uses 32 query groups per training batch, learning rate 10^{-5}, entropy coefficient 10^{-3}, and an actor micro-batch size of one. It uses four A100 GPUs and an SGLang rollout backend, with checkpoints evaluated on held-out development queries. Outcome and commitment baselines are computed within the same query group without standard-deviation normalization.

E BUDGET

For a legal trajectory with n_i^{turn} decision turns, n_i^{geom} all-candidate geometry calls, and n_i^{inspect} pair inspections, abstract cost is

$$C_i = c_{\mathrm{turn}} n_i^{\mathrm{turn}} + c_{\mathrm{geom}} n_i^{\mathrm{geom}} + c_{\mathrm{inspect}} n_i^{\mathrm{inspect}}. \tag{20}$$

The recorded accounting uses $c_{\mathrm{turn}} = 0.3$, $c_{\mathrm{geom}} = 1.0$, and $c_{\mathrm{inspect}} = 2.0$, with a per-query budget of 8.5 and at most six turns. The geometry charge already includes the entire shortlist; it is not charged five times again. Thus a direct answer costs 0.3, geometry followed by an answer costs 1.6, and geometry followed by one inspection and an answer costs 3.9. These are arithmetic examples under the accounting convention, not runtime measurements.

The budget constrains legal interaction and is visible to the policy, but no term $-\lambda C_i$ is added to the reward. Inspection rate is the fraction of queries with at least one inspection; mean inspection count includes every inspection and can exceed that rate. Direct-answer rate, geometry-only stopping rate, and inspection-using rate describe different trajectory paths. Forced fallback and illegal-action frequency are recorded separately so that failure to produce a legal decision is not mistaken for deliberate early stopping.

F PROMPTS USED FOR ANNOTATION, DISTILLATION, AND ROUTING

The system uses five model-facing prompt families. P1 and P2 generate training-only regional annotations for positive and hard-negative pairs, respectively. P3 is shared by the raw-image student, the deployed Inspector, and the pair-verification evaluation. P4 is used by the EMA teacher for both regional and correspondence-map privileged views. The latter was internally called D1: D0 contained only the two boxed full images, whereas D1 added a third visualization of cross-image RANSAC inlier correspondences. The index is an internal teacher-form identifier rather than the name of a geometric algorithm, and this training-only view is distinct from the Router's online Geometry Verifier. P5 consists of the Router system instruction and the dynamically rendered observation. Tables 8–12 reproduce the prompts verbatim; braces in the final observation schema mark runtime substitutions rather than literal prompt tokens.

Table 8: P1: offline positive-pair region-annotation prompt (verbatim).

```
You are given a QUERY room photo and a CANDIDATE room photo. They ARE the SAME
physical room, seen from different viewpoints and possibly different lighting;
the visible overlap between the two views may be large or small.

Task: output EXACTLY 3 region−PAIRS that most convincingly identify this as the
SAME room. Fill the 3 slots in this priority order:
1. Room−specific objects/fixtures visible in BOTH images (distinctive furniture,
   appliances, artwork, built−ins) — the same physical thing, not just the same type.

2. Structural signatures visible in BOTH (layout corners, window/door placement,
   ceiling features, molding junctions, stairs).
3. Shared permanent surfaces with a DISTINCTIVE material or pattern.
4. Only when slots would otherwise stay empty: plain shared surfaces (wall color,

   floor material). Then (a) anchor each box at a junction or fixture — floor−wall
   transition, baseboard, corner, outlet, vent — never a featureless center patch;
   (b) prefer COMPLEMENTARY surfaces (e.g., wall + floor + trim) whose combination
   is discriminative, over three patches of the same surface;

   (c) mark kind "surface_plain" and confidence "low".
Avoid generic items (plain wall, generic TV, standard door) at priorities 1−3
unless they carry a unique identifying detail (damage, sticker, unusual placement).

For each pair: query_bbox [x1,y1,x2,y2] and the matching candidate_bbox
[x1,y1,x2,y2], each normalized 0−1000 in its OWN image. Boxes must be tight
around the feature (never near−full−image), and the two boxes of a pair must
show the same physical thing.

If the two views share almost no visible area and you cannot honestly fill 3
pairs, output as many honest pairs as exist and set "no_covisible": true.

Output STRICT JSON only, no other text:
{"pairs":[{"query_bbox":[...],"candidate_bbox":[...],
  "confidence":"high|medium|low",
  "kind":"object_match|structure_match|surface_match|surface_plain",
  "reason":"<=15 words naming the specific shared feature"}],
 "no_covisible":false}
Do not state any room/building name.
```

Table 9: P2: offline hard-negative region-annotation prompt (verbatim).

```
You are given a QUERY room photo and a CANDIDATE room photo. They are DIFFERENT
physical rooms that look easily confusable (similar style, function, or furnishing).

Task: output EXACTLY 3 region-PAIRS that best EXPOSE they are different rooms.


Fill the 3 slots in this priority order:
1. FIXED structural differences: window/door count or placement, built-in
   fixtures, layout geometry, ceiling/floor architecture, permanent surfaces —
   features that could not differ between two photos of one room.


2. Trap exposure: a region pair where an apparent match is only a GENERIC or
   coincidental object — same type of item, visibly NOT the same physical thing
   (different model, size, wear, or position relative to structure).
3. Only if unavoidable: differences in movable furnishings — least preferred,


   since movable items can also differ between two photos of the SAME room.

For each pair: query_bbox and candidate_bbox (each normalized 0-1000 in its OWN
image), the two boxes covering the corresponding locations/objects being


contrasted. Boxes must be tight (never near-full-image).

Output STRICT JSON only, no other text:
{"pairs":[{"query_bbox":[...],"candidate_bbox":[...],

  "kind":"structural_difference|generic_coincidence|movable_difference",
  "reason":"<=15 words naming the specific difference"}]}
Do not state any room/building name.
```

Table 10: P3: raw-image pair-verification prompt used by the student and deployed Inspector (verbatim). The request carries the query and candidate RGB images in that order.

```
Task: room re-identification. Decide whether the CANDIDATE shows the same physical
room as the QUERY.
The two photos may differ in viewpoint, lighting, time, occlusion, and movable
objects. Judge only from visible evidence. Compare the room layout, permanent

surfaces, openings, furniture, objects, spatial relationships, and any distinctive
details that help decide whether the physical place is the same.
Write a concise explanation, then end with exactly two lines:
verdict: same | different

confidence: <integer 0-100>
```

Table 11: P4: privileged EMA-teacher prompt (verbatim). It is used with either the regional view or the correspondence-map teacher view internally denoted D1.

Task: room re−identification. Decide whether the CANDIDATE shows the same physical room as the QUERY.
The two photos may differ in viewpoint, lighting, time, occlusion, and movable objects. Judge only from visible evidence. Compare the room layout, permanent

surfaces, openings, furniture, objects, spatial relationships, and any distinctive details that help decide whether the physical place is the same.
The photos you receive carry quiet visual hints, and additional reference views may follow the two room photos. Together they point to regions worth close attention,

and to places that may show the same physical spot in both photos. Treat all of this as private guidance: absorb it silently, look at those places in the two room photos, and judge from what is actually visible there. Write your explanation and your two final lines as if you had seen only two plain room photos, describing each piece of

evidence by what it is in the room.
Write a concise explanation, then end with exactly two lines:
verdict: same | different
confidence: <integer 0−100>

Table 12: P5: Router system prompt and dynamically rendered observation schema. The system prompt is verbatim; bracketed fields in the schema are populated from the current episode state.

[SYSTEM]
Choose one legal room−retrieval action from answer #i, geometry, or inspect #i.
Read the current state before acting. In <why>, write two or three concise clauses
that identify the decisive global, geometry, and inspection evidence, explain any

conflict, and justify why the chosen action should stop or continue. Then output
exactly one <act>. Use exactly: <why>...</why><act>...</act>.

[DYNAMIC USER OBSERVATION]

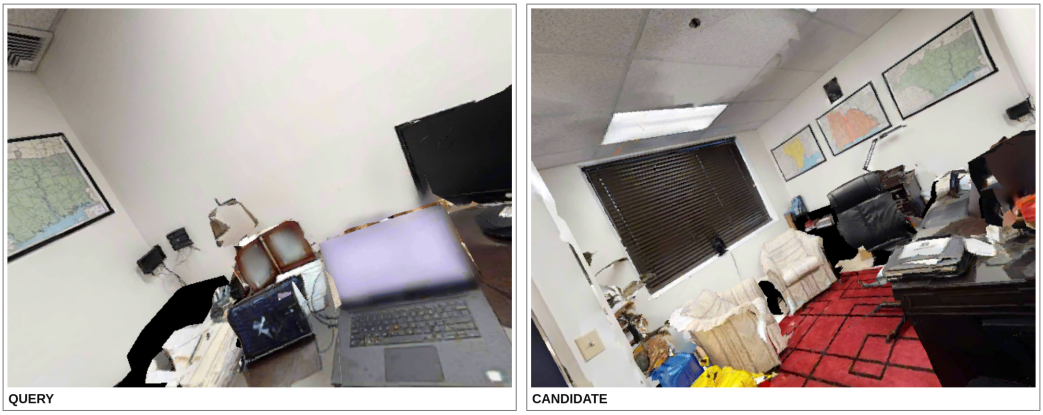
[RETRIEVAL STATE]
budget_remaining: {budget} turn: {turn}/{max_turns}
candidates (by retrieval rank):
 #1 cos={cosine} [inl={n} cov={coverage} llr={llr} clus={n}]

    [verdict={same|different} margin={margin} (pct={percentile}) note="{note}"]
 ...
 #5 cos={cosine} [...]
margin(1,2)={margin} ratio(1/2)={ratio} softmax_entropy@5={entropy}

gap_profile: [{gap_1}, {gap_2}, {gap_3}, {gap_4}]
geometry: not run | geom_top1=#{i} agrees_with_global=YES|NO
inspections: none | #{i}, ...
[IMAGES] query + candidates #1–#5 (224px thumbnails, in rank order)

<image> <image> <image> <image> <image> <image>

Table 13: **Complete pair-verification comparison.** Sight-OPSD identifies stable cross-view correspondences and returns the correct `same` verdict. Gemini 3.7 Flash and GPT-5.6 Sol interpret changes in visible wall coverage and furniture as structural contradictions and return `different`. This pair was selected qualitatively after model outputs had been frozen; it is not used to estimate aggregate accuracy.



Sight-OPSD **same; confidence 95; correct**
Same physical office. The large framed regional map on the query's left wall matches the rightmost framed map above the desk in the candidate. The black wall-mounted box remains immediately below it, while the articulated task lamp and cluttered workstation occupy the same adjacent area. The drop-ceiling grid, fluorescent panel, and desk-wall geometry are consistent under the large viewpoint rotation.

Gemini 3.7 Flash **different; confidence 95; wrong**
The QUERY and CANDIDATE images show different rooms. The QUERY room has a plain white wall with a single map on the left, while the CANDIDATE room has multiple maps on the wall and a red carpet. The furniture and layout are also distinct, with the QUERY room featuring a simple desk setup and the CANDIDATE room having additional chairs, a larger desk, and more clutter. These differences indicate they are not the same physical room.

GPT-5.6 Sol **different; confidence 94; wrong**
Both images show a cluttered office, but the permanent layout and furnishings differ. The query has a large TV on the wall/right side and a conference-style table with laptops, while the candidate has a large window with dark blinds, multiple framed maps, a red patterned carpet, and executive desks and chairs. These structural and spatial differences indicate different rooms.

G Complete Inference Trace Examples

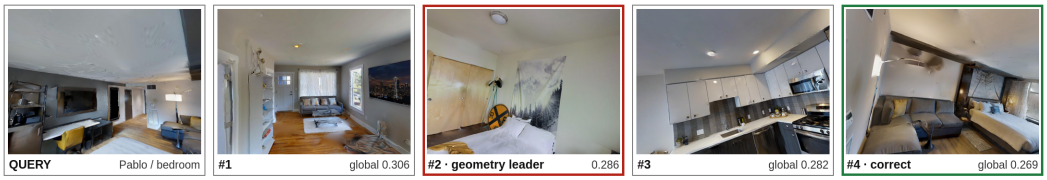
G.1 Pair verification against strong closed models

Table 13 shows a positive pair from the frozen 1,000-pair comparison subset. Sight-OPSD correctly recognizes the same office under a large viewpoint rotation, whereas Gemini 3.7 Flash and GPT-5.6 Sol both predict `different`. The reliable correspondence is the large framed regional map: it appears on the query's left wall and as the rightmost map above the desk in the reference view. The black wall-mounted device remains directly below it, and the adjacent articulated task lamp, workstation, ceiling grid, and fluorescent panel provide compatible context.

G.2 A complete routed episode

Table 14 presents the complete recorded trajectory for `query_2df35c2cf97ec8bf034d`. The true room begins at retrieval rank four. The initial margin is only 0.020, so the Router calls geometry. Geometry disagrees with the retrieval leader and selects candidate `#2`, although every candidate has a negative LLR. The Router therefore treats the geometric preference as weak and inspects `#2`. The Inspector rejects it by identifying an incompatible tapestry, closet, and bed configuration. The Router then inspects the next geometric candidate, `#4`. This view repeats the query's gray L-shaped sofa

Table 14: **Complete multi-turn Router trajectory.** Red marks the incorrect geometric leader and green marks the true room. After `geometry`, the Router uses one negative and one positive semantic inspection to resolve the conflict, then commits to candidate `#4`. Ground-truth highlighting and room identifiers are analysis annotations and were not visible to the Router.



Turn	Observed evidence and reasoning	Action
1. Initial	**Retrieval:** #1 0.306, #2 0.286, #3 0.282, #4 0.269. The top-two margin is only 0.020 (ratio 1.070; entropy 1.566), so the global leader is uncertain. **Budget:** 8.5. No spatial check has been made.	`geometry`
2. Geometry	**Geometric LLRs:** #1 -4.0, #2 -2.8, #3 -4.0, #4 -3.3. Candidate #2 becomes the geometric leader, disagreeing with global #1; every LLR remains negative, so the preference is weak. **Budget:** 7.2. Inspect the geometric leader before committing.	`inspect #2`
3. Inspect #2	**Inspector: different**, margin 5.6 (percentile 60.1241). The query shows the sofa–bed–TV arrangement, whereas #2 shows a separate bedroom with a wall tapestry and wooden closet doors. The layouts are incompatible. **Budget:** 4.9. Reject #2 and test the next geometric candidate.	`inspect #4`
4. Inspect #4	**Inspector: same**, margin 9.8 (percentile 5.64135). Both views contain the gray L-shaped sofa with a yellow cushion beside the bed, together with compatible wood flooring and the TV/console region. **Budget:** 2.6. The positive semantic evidence resolves the conflict in favor of #4.	`answer #4`

with a yellow cushion, the adjacent bed, wood flooring, and the TV/console region, so the Inspector returns `same` and the Router answers `#4`. The final answer is correct.

The table preserves the recorded decision sequence and its decision-relevant retrieval scores, geometric LLRs, semantic verdict margins, and budgets. Candidate `#5` is omitted because it is never selected or inspected. The evidence descriptions are edited only to name features visible in the displayed images. In particular, the episode demonstrates a justified semantic override: inspection corrects an erroneous geometric leader rather than merely confirming the candidate already preferred by geometry.