SSTG-Nav: Metric-Grounded Spatial-Semantic Topological Graphs for Reusable Object Navigation

Daojie Peng[1], **Bingtao Wang**[2], **Jun Ma**[1*]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Shandong University
dpeng108@connect.hkust-gz.edu.cn, wangbt@mail.sdu.edu.cn, jun.ma@ust.hk

Abstract

Service robots operating for months in the same homes, offices, and facilities should become more reliable with experience instead of searching familiar space from scratch for every request. Yet ObjectNav is predominantly formulated as one-shot exploration, leaving a central deployment challenge unresolved: recognizing an object does not identify a reachable place to stop, and one confident map error can terminate the task. We introduce SSTG-Nav, a reusable metric-semantic memory that turns a one-time survey into actionable object goals, consolidates evidence across viewpoints, and retains spatially distinct recovery standoffs. On 1,000 HM3D-v2 episodes across 36 scenes, our goal-independent topology achieves a 99.4% geometric success ceiling. Holding semantic responses fixed, metric grounding raises SR/SPL from 0.835/0.560 to 0.920/0.603, and source-aware fusion reaches 0.926/0.586. Fusion-aware Top-3 recovery raises Success@1/2/3 to 0.928/0.965/0.975 and reaches 0.601 SPL@3. Model, field-of-view, density, and corruption controls identify where these gains originate, and a ROS 2/Nav2 realization demonstrates the complete reusable query-to-execution pipeline. Together, the results establish pre-exploration as a powerful practical regime for dependable, repeated semantic navigation.

Code — https://github.com/DaojiePENG/sstg-nav-bench
Project — https://daojiepeng.github.io/SSTG-Nav

Introduction

ObjectNav requires an embodied agent to reach an instance of a requested category and declare STOP at a valid location. It is a foundational capability for service robots (Batra et al. 2020). Most benchmarks deliberately begin in an unexplored environment, and semantic exploration (Chaplot et al. 2020), vision-language frontier maps (Yokoyama et al. 2024), commonsense constraints (Zhou et al. 2023), online scene graphs (Yin et al. 2024), and topological memories (Liu et al. 2025) have substantially advanced that setting. Long-lived robots, however, revisit the same homes, offices, hospitals, and facilities across hundreds of requests. Repeating full discovery wastes prior experience and makes task completion less predictable. We study the complementary, practically central question of how a robot can convert one survey into reliable navigation infrastructure for all subsequent requests.

This reusable regime changes the objective from one-episode exploration efficiency to dependable completion over a robot's lifetime. A persistent map can amortize expensive perception, answer future language queries immediately, retain several hypotheses for ambiguous objects, and recover from a rejected candidate without restarting search. The central scientific challenge is making that memory *actionable*: it must encode where the robot can successfully stop, not merely what its cameras once recognized.

This distinction exposes a fundamental observation-to-action gap. A camera may recognize an object through a doorway, across a railing, or from the wrong side of furniture, while the camera pose itself is a poor navigation goal. Moreover, maps sampled from official target viewpoints already contain successful STOP configurations; we call this *target-view coverage* and use it as a controlled ceiling. Our goal-independent setting addresses the harder and more useful problem: discovering navigable space without task goals, then converting visual evidence into object-centric destinations that are reachable, corroborated, and verifiable at arrival.

Figure 1 presents our solution. SSTG-Nav separates a one-time, goal-independent survey from lightweight repeated queries. It converts view-local VLM evidence into reachable object-centric destinations, strengthens them through independent multi-view support, ranks alternatives on a sparse navigation graph, and uses fresh RGB-D evidence to decide whether to STOP or continue. Goal annotations terminate in a separate post-hoc evaluator lane. This design preserves the efficiency of topological memory while supplying the metric grounding and closed-loop recovery needed for dependable execution.

Our contributions are:

- **Scientific formulation:** we isolate reusable ObjectNav as an operating regime and identify the observation-pose/STOP-pose mismatch, together with protocols that separate target-derived coverage from goal-independent mapping.
- **Technical framework:** SSTG-Nav turns view-local VLM detections into reachable object-centric standoffs, consol-

*Corresponding author.

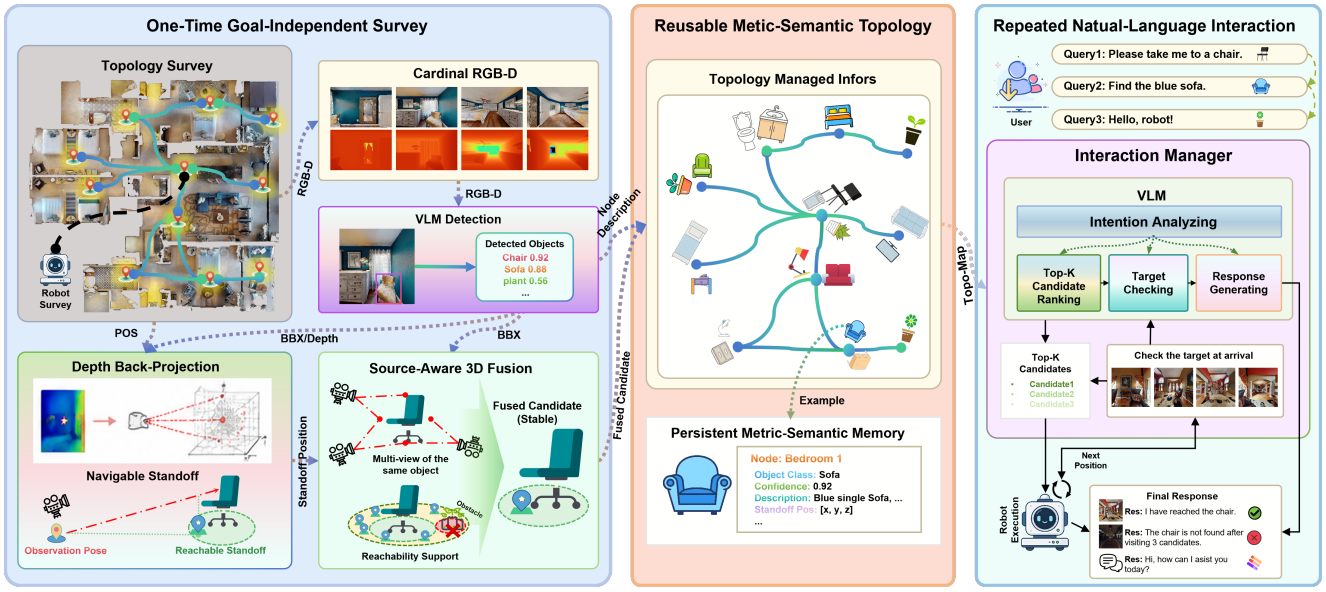

Figure 1: SSTG-Nav converts a one-time, goal-independent survey into a persistent metric-semantic topology that supports repeated natural-language navigation. View-local RGB-D detections are projected into reachable object-centric standoffs and consolidated across viewpoints. For each user request, the interaction manager constructs a structured semantic query, retrieves and ranks multiple candidates, coordinates graph planning and Nav2 execution, and verifies the target from fresh arrival RGB-D. A rejected hypothesis triggers navigation to the next candidate, while clarification, progress, recovery, and verified completion are reported through the bidirectional natural-language interface.

idates them through source-aware 3D fusion, and verifies candidate arrival from fresh RGB-D before STOP.

- **Empirical evidence:** full-validation coverage controls and same-response independent-map ablations distinguish geometry, semantic backend, FoV, metric grounding, fusion, and sequential recovery.
- **System realization:** a modular natural-language-to-map-to-navigation stack instantiates the interfaces in ROS 2, with closed-loop Nav2 target dispatch and physical motion verified on a mobile robot.

Together, the contributions turn pre-exploration from a favorable assumption into a rigorously evaluated navigation paradigm. Tables 3, 4, and 6 show that a task-independent map nearly saturates geometric coverage and that SSTG-Nav converts this potential into 0.926 one-shot SR and 0.975 fusion-aware Top-3 SR. This positions reusable metric-semantic topology as a strong foundation for long-lived robots that must answer many navigation requests in stable environments.

Related Work

Online ObjectNav. Early modular agents learn where to explore through semantic policies or potential maps (Chaplot et al. 2020; Ramakrishnan et al. 2022), while ZSON and CoW broaden recognition toward open-world or language-driven goals (Majumdar et al. 2022; Gadre et al. 2023). Foundation-model systems subsequently connect language and pixel evidence to online exploration through LLM reasoning, pixel goals, value maps, commonsense constraints, or sparse Voronoi structure (Yu, Kasaei, and Cao 2023; Cai et al. 2024; Zhou et al. 2023; Yokoyama et al. 2024; Wu et al. 2024; Peng, Ma, and Ma 2026). Recent methods add generative priors, geometric affordances, scene graphs, Bayesian structure, or predictive world models (Zhang et al. 2024; Yuan et al. 2024; Yin et al. 2024; Zhang et al. 2025b; Yin et al. 2025; Nie et al. 2025), and increasingly organize reasoning, experience retrieval, instruction interpretation, target fusion, frontier planning, topology, learned behavior, and execution consistency (Cao et al. 2025; Wang et al. 2026b; Long et al. 2025; Zhang et al. 2025a; Chabal et al. 2025; Liu et al. 2025; Cai et al. 2026; Wang et al. 2026a). These systems primarily search an initially unknown scene. SSTG-Nav instead assumes a goal-independent prior survey and isolates the metricization, fusion, and recovery errors that remain after exploration has been amortized.

Reusable semantic maps. Reusable abstract models first demonstrated that scene knowledge accumulated online can improve later ObjectNav requests (Campari et al. 2022). VLMaps fuses language features with 3D reconstruction for natural-language map queries (Huang et al. 2023), while ConceptFusion, OpenMask3D, and CLIP-Fields provide complementary open-set 3D or continuous semantic memories (Jatavallabhula et al. 2023; Takmaz et al. 2023; Shafiullah et al. 2023). CARe explicitly studies pre-explored maps and uses confidence and multi-view consistency to revise decisions (Ko et al. 2025). Topological graph memories compress experience into landmarks and connectivity (Kim et al. 2022), and TopoNav further establishes topology as an effective substrate for advanced ObjectNav reasoning (Liu

et al. 2025). Our representation is deliberately lighter than a dense feature field: it stores navigation nodes plus metric object candidates, then evaluates the precise conversion from a recognized observation to a valid ObjectNav STOP pose.

Problem and Protocols

Let a pre-explored scene contain navigable space $\mathcal{X}$ and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. A navigation node $v_i = (\mathbf{p}_i, \mathbf{q}_i, \mathcal{I}_i)$ stores its 3D position, orientation, and RGB-D views. An edge connects locally reachable nodes and is weighted by navmesh geodesic distance. A query supplies category c and start $\mathbf{x}_0$. The system returns a ranked metric candidate $\hat{\mathbf{s}}$ and the shortest feasible path from $\mathbf{x}_0$ to that point.

For an official episode with valid goal-viewpoint set $\mathcal{Y}_c$, success follows the ObjectNav task and challenge protocol (Batra et al. 2020; Yadav et al. 2023a):

$$S = \mathbf{1}\left[\min_{\mathbf{y}\in\mathcal{Y}_c} d_{\mathrm{geo}}(\hat{\mathbf{s}}, \mathbf{y}) \le 1\,\mathrm{m}\right]. \tag{1}$$

SPL is $S\ell/\max(\ell, p)$, where ℓ is the official shortest-path distance and p is the planned route length (Anderson et al. 2018). DTG is final geodesic distance to the nearest valid viewpoint.

We use three protocols. *Target-view coverage* selects one high-IoU navigable viewpoint per annotated object instance and measures a coverage-controlled upper bound. *Independent topology oracle* samples the map without goals, then assigns oracle semantic labels only for evaluation. It isolates geometric coverage. *Independent topology with real semantics* keeps goals hidden throughout capture, VLM inference, projection, fusion, and retrieval. This is our main protocol. Reporting the regimes separately makes the benefit of reusable mapping directly interpretable under each information budget.

Table 1: Information boundary of the three evaluation protocols. "Eval." means annotations are exposed only after mapping and candidate selection.

Protocol	Goals in map	Oracle semantics	Real VLM
Target-view coverage	✓	optional	optional
Independent oracle	✗	Eval.	✗
Independent real (ours)	✗	✗	✓

Table 1 makes the decisive distinction explicit: only the coverage control uses official goal viewpoints during map construction. In the independent real protocol, episode starts, categories, object annotations, and valid goal viewpoints remain hidden throughout topology construction, RGB-D capture, VLM inference, projection, fusion, and retrieval; they enter only after candidate selection for success and DTG evaluation. The supplement provides the complete construction audit.

Method

Goal-Independent Spatial Topology

For each scene, we draw a fixed-seed pool of navigable points and apply greedy farthest-point sampling until the empirical pool-cover radius is at most r. Edges connect up to eight nearby nodes when a navmesh path exists and its geodesic length passes a local threshold. This sampling is independent of episode starts, categories, object annotations, and goal viewpoints. The graph may be built once and reused.

At every node, the robot captures four cardinal RGB-D views. We process views separately rather than asking a VLM to localize objects in a stitched panorama, which preserves each camera model and avoids ambiguous panorama coordinates. Each returned detection is $d = (c, b, q, k)$: category, normalized box, confidence, and view index.

From Detection to Navigable Candidate

Let $\tilde{\mathbf{u}}$ be the box center and z the median valid depth in a 7×7 patch. With intrinsics K and world-from-camera transform T_{wk}, the estimated surface point is

$$\hat{\mathbf{o}}_d = T_{wk}\left(zK^{-1}[\tilde{\mathbf{u}}^\top, 1]^\top\right). \tag{2}$$

The observation pose itself can be a poor destination. We therefore move ρ meters from the surface toward the source camera in the ground plane,

$$\mathbf{s}_d^* = \hat{\mathbf{o}}_d + \rho\frac{\Pi(\mathbf{p}_i - \hat{\mathbf{o}}_d)}{\|\Pi(\mathbf{p}_i - \hat{\mathbf{o}}_d)\|_2}, \tag{3}$$

where Π drops the vertical coordinate. We snap $\mathbf{s}_d^*$ to navigable space and discard it if no navmesh path exists from the source node. This operation uses depth only for geometry. Semantic confidence remains the VLM output.

Reachability-Aware Soft Fusion

Detections of category c form a cluster C only if their object estimates are within horizontal radius r_o and vertical tolerance r_h, and their STOP candidates are mutually reachable within geodesic threshold r_g. The last condition prevents Euclidean fusion through walls or between disconnected floor regions. Multiple boxes from one source image must not inflate confidence. We retain the maximum confidence per source topology node and compute a noisy-OR score

$$Q(C) = 1 - \prod_{i\in U(C)}\left(1 - \max_{d\in C_i} q_d\right), \tag{4}$$

where $U(C)$ is the set of independent observing nodes. A cluster is retained with two-node support or a high-confidence singleton; its highest-confidence reachable standoff becomes the representative. We store two linked layers: $\mathcal{H}^F$ contains one Q-scored representative per retained cluster, while $\mathcal{H}^R$ keeps all reachable pre-fusion standoffs with view-local confidence. Fused-only Top-K draws every visit from $\mathcal{H}^F$; our policy draws visit one from $\mathcal{H}^F$ and spatially separated recovery visits from $\mathcal{H}^R$. Thus stable primary ranking does not erase alternative stopping geometry or filtered detections.

Algorithm 1 summarizes the complete offline transformation from calibrated observations to reusable metric candidates; no query or episode field appears in its inputs.

Sequential candidates improve fault tolerance without consulting the evaluator. At candidate k, the robot captures four

Algorithm 1 Goal-independent metric semantic mapping

Require: Graph $\mathcal{G}$, RGB-D views, VLM f, standoff ρ
$\mathcal{H} \leftarrow \emptyset$
for $v_i \in \mathcal{V}$ and view k **do**
  for $d = (c, b, q, k) \in f(I_{ik})$ **do**
    $z \leftarrow \textsc{PatchMedian}(D_{ik}, b)$
    $\hat{\mathbf{o}}_d \leftarrow \textsc{BackProject}(b, z, K, T_{ik})$
    $\mathbf{s}_d^* \leftarrow \textsc{Standoff}(\hat{\mathbf{o}}_d, \mathbf{p}_i, \rho)$
    $\mathbf{s}_d \leftarrow \textsc{NavSnap}(\mathbf{s}_d^*)$
    if $\mathbf{s}_d$ is reachable from v_i **then**
      append $(c, q, \hat{\mathbf{o}}_d, \mathbf{s}_d, i)$ to $\mathcal{H}$
    end if
  end for
end for
for category c **do**
  cluster $\mathcal{H}_c$ by 3D proximity and STOP reachability
  score clusters with independent-source noisy-OR
  retain multi-source clusters or confident singletons
  retain reachable residual standoffs for recovery
end for
return $\mathcal{G}$ with fused representatives and residual standoffs

fresh 120° RGB-D views: view zero faces the stored object estimate and the others rotate by 90°. GPT-5.4 receives these views plus the boxed mapping-time reference and returns target visibility, a stopping-side judgment g_k, an arrival-view box, and confidence q_k. The box indexes aligned depth with central-patch median z_k. We accept STOP exactly when

$$A_k = \mathbf{1}[\text{visible}_k \wedge g_k = \text{valid} \wedge q_k \geq 0.75 \\ \wedge\, 0.25\,\text{m} \leq z_k \leq 2.5\,\text{m}]. \tag{5}$$

On rejection, planning continues from the current pose to the next spatially diverse candidate. Official goal viewpoints enter only after this decision to score success, SPL, and verifier confusion.

Reusable Query and Execution Layer

Let the language normalizer map query q to category $c(q)$, and let

$$\mathcal{C}(q) = \{i : c_i = c(q),\ \mathbf{s}_i \text{ is reachable from } \mathbf{x}_0\} \tag{6}$$

be its fused primary candidates from $\mathcal{H}^F$. Candidate order is the lexicographic permutation

$$\boldsymbol{\pi}_q = \text{LexSort}_{i \in \mathcal{C}(q)} \left(-Q_i,\ d_{\mathcal{G}}(\mathbf{x}_0, \mathbf{s}_i)\right), \tag{7}$$

which makes semantic support primary and path length a deterministic tie-breaker. This choice prevents a nearby unsupported observation from outranking independently corroborated evidence while avoiding a learned query-time policy.

Visit one uses $\pi_q(1)$. Up to two later visits come from the category-matched $\mathcal{R}(q) \subset \mathcal{H}^R$, ordered by confidence and path distance while enforcing $\|\mathbf{s}_i - \mathbf{s}_{\pi_q(j)}\|_2 \geq 2\,\text{m}$ for every prior visit. If $\mathcal{C}(q)$ is empty, the best reachable residual becomes visit one. Hence fused-only tries three cluster representatives, whereas fusion-aware tries one supported representative and two view-specific standoffs.

Execution is a receding candidate sequence rather than independent start-to-goal trials. With $\mathbf{x}_0$ the request start, leg k is

$$P_k = \text{ShortestPath}(\mathcal{G}, \mathbf{x}_{k-1}, \mathbf{s}_{\pi_q(k)}), \qquad \mathbf{x}_k = \mathbf{s}_{\pi_q(k)}, \tag{8}$$

and the executed path is $P(q) = P_1 \oplus \cdots \oplus P_\tau$ for $\tau = \min\{k : A_k = 1\}$. Hence a rejected hypothesis changes the next planning state instead of restarting the episode. For $m_c = |\mathcal{C}(q)|$, ordering costs $O(m_c \log m_c)$ and each graph search costs $O(|\mathcal{E}| + |\mathcal{V}| \log |\mathcal{V}|)$. If one survey serves N requests, its amortized non-motion cost is

$$\bar{C}_N = C_{\text{map}}/N + C_{\text{retrieve}} + C_{\text{plan}}, \tag{9}$$

formalizing the intended advantage: VLM inference and depth grounding are paid once, while repeated requests perform retrieval, graph search, and fresh arrival verification only.

Experiments

Setup

We evaluate the HM3D-Semantics v0.2 ObjectNav episodes from the 2023 Habitat Navigation Challenge (Yadav et al. 2023a,b). The episodes use semantic annotations over HM3D scenes (Ramakrishnan et al. 2021) and are rendered in Habitat-Sim 0.3.3 (Savva et al. 2019). Full validation has 1,000 episodes across 36 scenes and six categories: chair, bed, plant, toilet, TV/monitor, and sofa. Goal-independent farthest-point sampling draws 12,000 navigable proposals per scene and stops at $r = 0.8\,\text{m}$ coverage, producing 6,642 nodes and 21,845 edges. Each node stores four 640×640, 120° RGB-D views at 1.25 m camera height. Controlled representation and model ablations use the official 30-episode, two-scene minival subset on a 339-node graph; these scales are identified explicitly in Tables 3–6.

The primary semantic backend is GPT-5.4 (OpenAI 2026); MiMo-v2.5 (Xiaomi MiMo Team 2026) and local Qwen2.5-VL-3B-Instruct (Bai et al. 2025) are auxiliary model controls. We use $\rho = 0.8\,\text{m}$, $r_o = 1.2\,\text{m}$, $r_h = 1.0\,\text{m}$, $r_g = 3.0\,\text{m}$, and residual separation $\delta = 2.0\,\text{m}$. Full fusion retains two-source clusters or singletons above 0.92; the small-split representation controls retain every valid cluster so category scarcity is not confounded with filtering. Arrival verification uses four fresh 640×640, 120° views, confidence 0.75, and measured depth range $[0.25, 2.5]\,\text{m}$. We report Wilson 95% intervals for SR and 10,000-sample non-parametric bootstrap intervals for SPL. Supplementary Sections 1–2 audit the construction, prompts, and parameters used by all result tables.

Comparison Across Operating Regimes

The central result is that a reusable survey moves ObjectNav to a substantially stronger operating point. Among the unknown-scene HM3D-v2 systems listed in Table 2, the best SR and SPL are 0.842 and 0.479. A single fused SSTG-Nav destination already reaches 0.926/0.586, and fusion-aware recovery reaches 0.975/0.601, margins of 0.133 SR and 0.122 SPL over those strongest listed values. More importantly, this reliability does not require a task-specific navigation policy: one goal-independent survey builds metric-semantic infrastructure that can serve many later requests.

Table 2: ObjectNav context on HM3D-v2. Unknown-scene methods explore per episode, whereas reusable-map methods amortize a survey across repeated queries. SSTG-Nav achieves the highest SR and SPL among the listed results. TF denotes no task-specific policy training; CARe reports a custom MP3D success metric (†).

Method	Scene at start	Evaluation	Policy	SR	SPL
Unknown-scene HM3D-v2					
SG-Nav (Yin et al. 2024; Wang et al. 2026b)	unknown	HM3D-v2	TF	0.496	0.255
InstructNav (Long et al. 2025)	unknown	HM3D-v2	TF	0.580	0.209
VLFM (Yokoyama et al. 2024; Wang et al. 2026b)	unknown	HM3D-v2	TF	0.636	0.325
ApexNav (Zhang et al. 2025a)	unknown	HM3D-v2	TF	0.762	0.380
FOM-Nav (Chabal et al. 2025)	unknown	HM3D-v2	learned	0.758	0.479
TrajRAG (Wang et al. 2026b)	unknown	HM3D-v2	TF	0.781	0.402
IntentNav (Cai et al. 2026)	unknown	HM3D-v2	learned	0.822	0.385
ConsistNav (Wang et al. 2026a)	unknown	HM3D-v2	TF	0.842	0.412
Pre-explored reusable maps					
CARe+VLMaps (Ko et al. 2025)	pre-explored	custom MP3D	TF	0.827†	–
SSTG-Nav fused, single	pre-explored	HM3D-v2	TF	0.926	0.586
SSTG-Nav fusion-aware Top-3	pre-explored	HM3D-v2	TF	**0.975**	**0.601**

Table 3: Full-validation coverage–semantics decomposition over 1,000 episodes. Target-view maps use goal-derived capture poses; independent maps do not.

Map protocol	Semantics	Nodes	SR	SPL	DTG
Target-view	Oracle	1,168	0.990	0.802	**0.073**
Target-view	Qwen	1,168	0.644	0.438	2.016
Independent	Oracle	6,642	**0.994**	**0.992**	0.622
Independent	GPT-5.4 fused	6,642	0.926	0.586	0.747

Table 4: Cumulative full-validation component ablation on the same 1,000 episodes, 6,642-node topology, and 6,642 GPT-5.4 responses. SR/SPL are measured after at most K destinations; DTG is shown for one-shot variants.

Variant	Depth	Fusion	Recovery	K	SR	SPL	DTG
Camera-node baseline	✗	✗	✗	1	0.835	0.560	1.039
+ metric grounding	✓	✗	✗	1	0.920	**0.603**	0.879
+ source-aware fusion	✓	✓	✗	1	0.926	0.586	**0.747**
+ residual recovery	✓	✓	✓	3	**0.975**	0.601	–

In recurring environments such as homes, offices, and care facilities, perception and mapping are paid once; subsequent natural-language queries reduce to retrieval, graph planning, and targeted recovery. The analyses below isolate how coverage, grounding, fusion, and residual candidates produce this advantage. Complete HM3D-v1/v2 context appears in the supplement.

Coverage Is Necessary but Not Sufficient

This operating point first requires a survey that covers useful stopping regions without knowing future goals. The full-validation decomposition in Table 3 shows 0.990 SR for target-view oracle coverage and 0.994 SR/0.992 SPL for the goal-independent 0.8 m topology with evaluator-assigned semantics. Independent sampling therefore provides nearly complete success-region coverage without privileged target views. The remaining gap to real GPT-5.4 semantics localizes the main challenge to semantic grounding and STOP placement rather than map density; confidence intervals and failure counts appear in the supplement.

Coverage alone, however, does not solve retrieval. Even on privileged target-view poses, real Qwen semantics reach only 0.644/0.438 in Table 3. Supplementary Table 2 shows the progression behind that result: nearest-node retrieval gives 0.426/0.324, primary-label filtering gives 0.574/0.415, and confidence ranking reduces wrong selections from 549 to 331. Confidence clearly extracts value from fixed observations, but the remaining oracle gap motivates the proposed multi-view support and metric reachability.

Metric Grounding and Fusion

Having established sufficient coverage, we next ask which components convert observations into successful destinations. The cumulative ablation in Table 4 assigns a distinct role to each stage. Metric grounding supplies the largest one-shot gain under fixed perception, changing 93 failures to successes and eight in reverse and raising SR/SPL by 0.085/0.042 (exact McNemar $p < 10^{-18}$). Fusion compresses the primary index from 20,107 candidates to 1,329 supported hypotheses, lowers DTG from 0.879 to 0.747 m, and reaches 0.926 one-shot SR. Residual recovery then raises completion to 0.975 SR/0.601 SPL. The gains therefore come from actionable geometry, source support, and candidate diversity rather than a change in VLM responses.

To test whether fusion is merely a GPT-specific prompting effect, we repeat the same-response comparison across backend and FoV. In all six triplets in Table 5, fusion raises raw SR: 0.920 to 0.926 at full scale, 0.867 to 0.967 and 0.933 to 1.000 for GPT controls, 0.333 to 0.500 for Qwen, and 0.833 to 0.900 and 0.567 to 0.633 for MiMo. This consistency, together with the 20,107-to-1,329 primary-index reduction, supports fusion as a representation component rather than a model-specific artifact; uncertainty and paired discordance appear in supplementary Tables 5 and 7.

These controls also reveal when a wider camera is useful. The 120° view exposes a low toilet missed by the standard

Table 5: Same-response RGB-D and fusion ablations across scale, FoV, and backend. Each entry is SR/SPL; bold marks the best SR within each triplet. The full row uses 1,000 episodes/36 scenes; the remaining rows use 30 episodes/two scenes.

Backend	FoV	Camera node	Raw RGB-D	Soft fusion
GPT-5.4 (full)	120°	0.835/0.560	0.920/0.603	**0.926**/0.586
GPT-5.4 (mini)	90°	0.733/0.568	0.867/0.681	**0.967**/0.665
GPT-5.4 (mini)	120°	0.933/0.704	0.933/0.713	**1.000**/0.691
Qwen2.5-VL-3B (mini)	90°	0.400/0.233	0.333/0.188	**0.500**/0.218
MiMo-v2.5 (mini)	90°	0.200/0.129	0.833/0.614	**0.900**/0.664
MiMo-v2.5 (mini)	120°	0.567/0.429	0.567/0.398	**0.633**/0.367

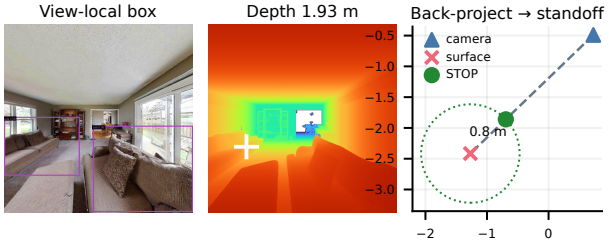

Figure 2: An audited sofa observation. A view-local VLM box indexes aligned depth; its median surface point is back-projected and shifted to a reachable 0.8 m standoff. ObjectNav goals are not used in this construction.

camera, while fusion replaces an unsupported sofa singleton with a 22-source cluster, as illustrated in Figure 4(a,b). Accordingly, the GPT-5.4 wide row in Table 5 removes both raw failures and reaches 1.000 SR, trading only 0.022 SPL for perfect completion. MiMo instead falls from 0.900 to 0.633 fused SR when widened. Wider FoV is therefore not a free gain; visibility, projection geometry, and VLM localization must be calibrated together.

The largest full-scale gain has a direct geometric explanation. In Figure 2, the VLM box selects a metric surface and depth back-projection moves the goal from the observation pose to a reachable 0.8 m standoff. With identical detections, this operation raises SR from 0.835 to 0.920 in Table 4. The improvement therefore comes from closing the observation-to-STOP gap, not from changing semantic predictions.

Density and Robustness Analysis

Once detections are grounded into reachable standoffs, adding topology nodes quickly stops helping. The left panel of Figure 3 shows that 0.25/0.50/0.75/1.00 topology fractions use about 85/170/255/339 nodes yet reach oracle SR 0.900/1.000/1.000/1.000. Half the topology already covers every minival success region, making semantic localization the higher-value target beyond this density.

Semantic reliability behaves differently. In the right panel of Figure 3, 25% and 50% dropout reduce mean SR from 1.000 to 0.907 and 0.763, while 15% false-positive probability reduces it to 0.840. The full 48-condition grid (28,800 evaluations) falls to 0.602 SR when 50% node retention, 25%

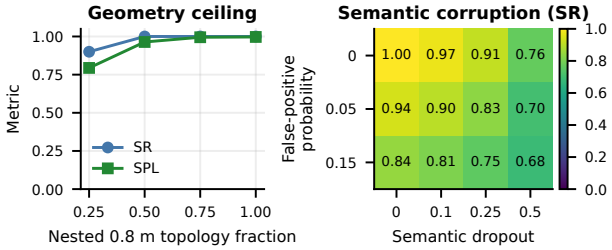

Figure 3: Geometric density saturates before semantic reliability on minival. Left: oracle semantics on nested goal-independent topologies. Right: mean SR over 20 controlled oracle-label corruptions with all nodes retained.

dropout, and 15% false positives are combined. The contrast is the useful observation: geometry saturates early, whereas modest semantic corruption remains damaging. These controlled perturbations do not model a particular VLM.

The mechanism-level audit closes the loop between these controls and execution. Fresh RGB-D rejects an incorrect rank-1 plant pose and accepts a spatially distinct rank-2 candidate in Figure 4(c), the behavior quantified by Table 6. The physical sequence in Figure 4(d) traces the implemented robot pipeline from metric RGB-D observations and a surveyed semantic topology to language-conditioned retrieval and Nav2 target arrival.

Sequential Fusion-Aware Recovery

This failure mode motivates sequential fusion-aware recovery. We evaluate the first one, two, and three destinations in Table 6; Success@k asks whether the first k visits contain a valid stopping pose, and SPL@k accumulates executed legs through the first success. The protocol directly tests recovery from an incorrect primary hypothesis, while the deployed loop supplies the fresh RGB-D decision in Eq. 5; verifier calibration remains in the supplement.

Table 6: Sequential recovery on full validation. Fused-only draws all visits from $\mathcal{H}^F$; ours draws one primary from $\mathcal{H}^F$ and 2 m-separated backups from $\mathcal{H}^R$. ΔSR uses the strict single fused target.

Candidate policy	K	SR	ΔSR	SPL	Successes
Single fused target (no recovery)	1	0.926	–	0.586	926
Fusion-aware primary only	1	0.928	+0.002	0.588	928
Fused-only Top-3 control	3	0.964	+0.038	0.596	964
Primary + one diverse residual	2	0.965	+0.039	0.598	965
Primary + two diverse residuals	3	**0.975**	**+0.049**	**0.601**	**975**

The recovery ablation confirms that success comes from preserving useful alternatives, not simply revisiting more fused clusters. As Table 6 shows, a single target succeeds in 926 episodes; residual fallback adds two, and the first and second diverse residuals recover 37 and ten more. Final SR reaches 0.975 (+0.049) with 0.601 SPL, whereas fused-only Top-3 saturates at 0.964. The paired comparison yields eleven gains and no losses (nine sofa, two bed; exact McNe-

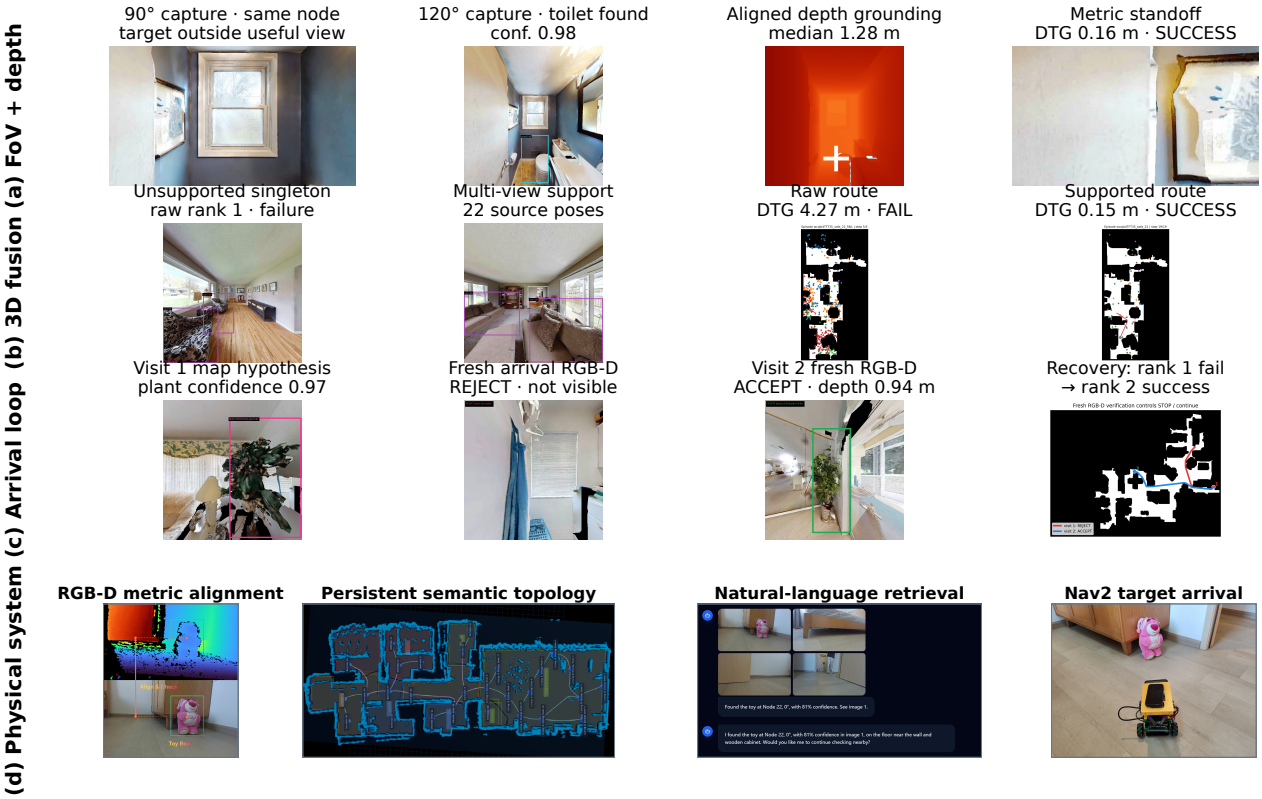

Figure 4: Mechanism and deployment audit. (a) Wide capture and depth ground a successful toilet standoff. (b) Source-aware fusion replaces an unsupported sofa singleton with a 22-view hypothesis. (c) Fresh RGB-D rejects a failed plant candidate and accepts rank 2. (d) The physical-system sequence shows onboard RGB-D alignment, the persistent semantic topology produced by the survey, natural-language retrieval of a toy at Node 22, and the robot's Nav2 arrival at the selected target.

mar $p < 0.001$); matched raw and separation controls remain in the supplement.

Physical-System Realization

We deploy SSTG-Nav on a Yahboom X3 mobile robot with ROS 2 Humble (Macenski et al. 2022) and Nav2 (Macenski et al. 2020), as shown in Figure 4(d). A Gemini 336L supplies aligned RGB-D observations, an RPLIDAR S2 supplies laser scans, and `slam_toolbox` (Macenski and Jambrecic 2021) provides occupancy mapping and localization. The persistent manager retains $\mathcal{G}$ and interaction state across legs, linking natural-language category parsing, Eq. 7 planning, Nav2 execution, arrival verification, and candidate recovery. The illustrated query retrieves the toy at Node 22 before dispatching the selected target to Nav2; the corresponding real-robot run is provided separately as multimedia material.

Discussion and Limitations

Across protocols, goal-independent geometry reaches 0.994 SR, while metric grounding, fusion, and recovery convert that potential into 0.926 one-shot and 0.975 Top-3 SR (Tables 3, 4, and 6). These gains establish topology coverage, object-centric metricization, source-aware evidence, and sequential recovery as a coherent route to reliable repeated ObjectNav.

Limitations. The full study covers one HM3D-v2 split, one hosted primary VLM, six categories, and static scenes. Continuous navmesh paths abstract control noise and map change, and relational dialogue lies beyond the category-query benchmark. Future work will extend the persistent-map and arrival interfaces to dynamic environments and broader real-camera evaluation.

Conclusion

SSTG-Nav turns reusable RGB-D observations into reachable, fused metric-semantic goals. Grounding raises SR from 0.835 to 0.920, fusion reaches 0.926 one-shot SR, and residual recovery reaches 0.975 Top-3 SR (Tables 4 and 6). Together with the FoV/model controls and ROS 2/Nav2 realization, the result provides a reproducible foundation for high-reliability navigation over persistent semantic maps.

Supplementary Material

This appendix provides self-contained protocol definitions, complete quantitative results, qualitative analyses, and physical-system details that complement the main paper.

1 Protocol Definitions

1.1 Task and Evaluator

The evaluation uses the HM3D-Semantics v0.2 ObjectNav episodes from the 2023 Habitat Navigation Challenge (Yadav et al. 2023a,b). The episodes use HM3D scenes (Ramakrishnan et al. 2021) and are rendered in Habitat-Sim 0.3.3 (Savva et al. 2019). For episode e, let c_e be the requested category, $\mathbf{x}_e$ the official start, ℓ_e the official shortest path, and $\mathcal{Y}_{c_e}$ all official navigable goal viewpoints for the category in that scene. A selected candidate $\hat{\mathbf{s}}_e$ succeeds when its navmesh geodesic distance to $\mathcal{Y}_{c_e}$ is at most 1 m. The evaluator records

$$S_e = \mathbf{1}\left[\min_{\mathbf{y}\in\mathcal{Y}_{c_e}} d_{\mathrm{geo}}(\hat{\mathbf{s}}_e, \mathbf{y}) \le 1\right], \tag{1}$$

$$\mathrm{SPL}_e = S_e \frac{\ell_e}{\max(\ell_e, p_e)}, \tag{2}$$

$$\mathrm{DTG}_e = \min_{\mathbf{y}\in\mathcal{Y}_{c_e}} d_{\mathrm{geo}}(\hat{\mathbf{s}}_e, \mathbf{y}). \tag{3}$$

If no semantic candidate exists, SPL is zero and DTG is the official start-to-goal distance. Episode keys concatenate scene, category, and original episode ID because HM3D reuses numeric IDs across categories. Path and goal-distance caches include candidate coordinates rather than map-local node IDs, and predicted categories never shortcut distance computation; only explicitly evaluator-created oracle nodes can do so. These implementation safeguards prevent semantic predictions from being treated as ground truth. SPL follows the standard embodied-navigation definition (Anderson et al. 2018); the success interpretation follows ObjectNav recommendations (Batra et al. 2020).

Main-paper Table 1 separates map supervision from post-hoc scoring. The independent real protocol is the primary result: episode starts, categories, object annotations, and valid goal viewpoints remain unavailable to topology construction, RGB-D capture, VLM inference, projection, fusion, and retrieval. The tables below expand that information boundary through construction counts, model-response completion statistics, and episode-level results.

Table 1 preserves the broad comparison while allowing the main paper to foreground the most relevant HM3D-v2 block. Across 16 published unknown-scene rows spanning both HM3D versions, the reusable-map operating point remains visibly distinct; the controlled internal tables below then isolate where SSTG-Nav obtains its gains.

1.2 Protocol A: Target-View Coverage

For every annotated object instance with available viewpoints, the mapper chooses one navigable high-IoU official viewpoint and captures four RGB views at yaw offsets $0°$, $-45°$, $45°$, and $180°$. Across the 36 validation scenes, this creates 1,168 nodes and 6,446 edges. Because official goal viewpoints determine capture locations, this protocol measures semantic retrieval and planning under controlled coverage. It is neither goal-independent nor deployable.

Table 2 shows that target-view coverage is not exactly one. Ten episodes have no retained coverage node: four TV/monitor episodes in scene `BAbdmeyTvMZ`, five plant episodes in `q5QZSEeHe5g`, and one chair episode in the same scene. These cases are retained. Qwen inference completes on 1,167 of 1,168 node images; the unsuccessful response is also retained as a mapping failure. The 0.346 SR gap between oracle and Qwen under the same privileged poses isolates semantic retrieval error from coverage.

The per-category target-view audit is incorporated into Table 6 alongside the full goal-independent results. Under confidence-ranked Qwen semantics, TV/monitor is lowest at 0.430 SR whereas bed reaches 0.818. These values describe semantic retrieval under controlled coverage and are not used as class weights or tuning signals.

1.3 Protocol B: Goal-Independent Topology

For each scene, Habitat's pathfinder draws 12,000 navigable points using seed 20260719 plus the deterministic scene index. Greedy farthest-point selection starts at the pool point nearest its centroid and repeatedly adds the point farthest from the selected set until every pool point is within 0.8 m. Up to eight Euclidean-nearest nodes are considered for edges. An edge is kept only when the Euclidean distance is at most 2.4 m, a navmesh route exists, and its geodesic length is at most 3.84 m. The mapper uses the split only to enumerate scene assets; no episode start, category, object annotation, or goal viewpoint is consulted by sampling, capture, inference, or fusion.

Table 3 verifies that the independent construction covers both compact and large scenes with one radius rule rather than a fixed node count. Oracle semantics are assigned only after mapping by measuring each node against the official category goal sets. This evaluator-side labeling produces a geometric coverage ceiling of 0.994 SR/0.992 SPL on all 1,000 validation episodes (994 successes; SR 95% CI [0.987,0.997]). Real-semantic variants never receive those labels.

2 Metric Semantic Mapping Details

2.1 RGB-D Capture and Projection

Every independent node stores four cardinal views at yaw offsets $0°$, $90°$, $180°$, and $270°$. The standard camera is 640×360 with $90°$ horizontal FoV; the wide camera is 640×640 with $120°$ horizontal FoV. RGB and depth share a pose 1.25 m above the agent base. Habitat depth is camera-forward Z depth. For image width W, height H, and horizontal FoV ϕ_h, vertical FoV is

$$\phi_v = 2\tan^{-1}\left(\frac{H}{W}\tan\frac{\phi_h}{2}\right). \tag{4}$$

The standard camera consequently has an approximately $59°$ vertical FoV, whereas the square wide camera has $120°$ in both directions.

The VLM returns category, view index, normalized box, and confidence. We use the box center and the median of valid depth values in a 7×7 patch, accepting 0.2–6.0 m. Let (u, v) be normalized center coordinates and z depth. Camera-frame coordinates are

$$x_c = 2(u - 0.5)z\tan(\phi_h/2), \tag{5}$$

$$y_c = -2(v - 0.5)z\tan(\phi_v/2), \qquad z_c = -z. \tag{6}$$

Table 1: Complete published ObjectNav context retained outside the compact main table. HM3D-v1 and HM3D-v2 are shown separately; reusable-map methods amortize a survey over repeated queries. TF denotes no task-specific policy training, and CARe reports a custom MP3D success metric (†).

Method	Scene at start	Evaluation	Policy	SR	SPL
Unknown-scene HM3D-v1					
VoroNav (Wu et al. 2024)	unknown	HM3D-v1	TF	0.420	0.260
VLFM (Yokoyama et al. 2024)	unknown	HM3D-v1	TF	0.525	0.304
GAMap (Yuan et al. 2024)	unknown	HM3D-v1	TF	0.531	0.260
SG-Nav (Yin et al. 2024)	unknown	HM3D-v1	TF	0.540	0.249
FBN (Zhang et al. 2025b)	unknown	HM3D-v1	TF	0.588	0.312
ApexNav (Zhang et al. 2025a)	unknown	HM3D-v1	TF	0.596	0.330
TrajRAG (Wang et al. 2026b)	unknown	HM3D-v1	TF	0.625	0.339
CogNav (Cao et al. 2025)	unknown	HM3D-v1	TF	0.725	0.262
Unknown-scene HM3D-v2					
SG-Nav (Yin et al. 2024; Wang et al. 2026b)	unknown	HM3D-v2	TF	0.496	0.255
InstructNav (Long et al. 2025)	unknown	HM3D-v2	TF	0.580	0.209
VLFM (Yokoyama et al. 2024; Wang et al. 2026b)	unknown	HM3D-v2	TF	0.636	0.325
ApexNav (Zhang et al. 2025a)	unknown	HM3D-v2	TF	0.762	0.380
FOM-Nav (Chabal et al. 2025)	unknown	HM3D-v2	learned	0.758	0.479
TrajRAG (Wang et al. 2026b)	unknown	HM3D-v2	TF	0.781	0.402
IntentNav (Cai et al. 2026)	unknown	HM3D-v2	learned	0.822	0.385
ConsistNav (Wang et al. 2026a)	unknown	HM3D-v2	TF	0.842	0.412
Pre-explored reusable maps					
CARe+VLMaps (Ko et al. 2025)	pre-explored	custom MP3D	TF	0.827†	–
SSTG-Nav fused, single	pre-explored	HM3D-v2	TF	0.926	0.586
SSTG-Nav fusion-aware Top-3	pre-explored	HM3D-v2	TF	**0.975**	**0.601**

Table 2: Full HM3D-v2 validation under target-view coverage (1,000 episodes, 36 scenes).

Variant	SR	SR 95% CI	SPL	SPL 95% CI	Failures
Oracle labels	0.990	[0.982,0.995]	0.802	[0.791,0.812]	10 missing
Qwen all, nearest	0.426	[0.396,0.457]	0.324	[0.299,0.349]	549 wrong, 25 missing
Qwen primary	0.574	[0.543,0.604]	0.415	[0.389,0.439]	265 wrong, 161 missing
Qwen all, confidence	0.644	[0.614,0.673]	0.438	[0.415,0.461]	331 wrong, 25 missing

Table 3: Independent topology construction. The full split has 6,642 nodes over 36 scenes; per-scene counts range from 74 to 399.

Split / scene	Nodes	Edges	Empirical cover radius
Full validation	6,642	21,845	≤ 0.800 m
Minival: `TEEsavR23oF`	132	415	≤ 0.8 m
Minival: `wcojb4TFT35`	207	691	≤ 0.8 m
Minival total	339	1,106	–

The stored node rotation and view yaw transform this point to world coordinates. A desired 0.8 m standoff is placed toward the source camera in the horizontal plane, assigned the source base height, snapped to the navmesh, and rejected when unreachable from the observing node.

2.2 Fusion Parameters and Decision Rules

Candidates are first partitioned by category. Two candidates are connected when their object estimates are within 1.2 m horizontally and 1.0 m vertically, and their snapped STOP poses are mutually reachable within 3.0 m geodesic distance. Connected components define clusters. Within cluster C, duplicate detections from one source node contribute only their maximum confidence q_i, producing

$$Q(C) = 1 - \prod_{i \in U(C)} (1 - \min(0.99, q_i)). \tag{7}$$

For the full-validation map, clusters with at least two unique source nodes are retained and a singleton is retained when its confidence is at least 0.92. The minival representation/model controls retain every valid cluster (minimum support one), because applying the full-map filter to only two scenes would confound representation with category deletion. In both cases the highest-confidence member supplies the representative STOP pose; clustering geometry and noisy-OR scoring are unchanged. The main paper therefore compares representations within, not across, each scale/backend triplet.

The reachability condition is necessary. Euclidean proximity alone can merge observations across walls or between disconnected navmesh islands. Likewise, repeated boxes from one frame are not independent confirmation and must not increase support.

2.3 VLM Prompt and Parser

Both hosted backends receive four separate RGB images in one node-level request. The effective instruction is:

> Inspect the four indexed views from one robot pose. Detect every visible instance of chair, bed, plant, toilet, TV/monitor, and sofa. For each detection return JSON with `view_index`, canonical `category`, normalized `bbox_norm=[x1,y1,x2,y2]`, and `confidence`. Boxes must be local to the named view; do not use panorama coordinates or infer an object that is not visibly grounded.

The parser accepts coordinates in $[0, 1]$ or $[0, 1000]$, converts the latter, clamps bounds, rejects invalid boxes and categories, and records the raw response unchanged. GPT-5.4 (OpenAI 2026) and MiMo-v2.5 (Xiaomi MiMo Team 2026) use the same parser version, `bbox-norm-or-1000-v2`, so model comparisons do not confound response scaling.

Local Qwen2.5-VL-3B (Bai et al. 2025) uses the same four 90° RGB-D views in a 2×2 grounding panorama and returns an absolute box confined to one tile. Valid empty lists are retained. A subset of panoramas produces degenerate non-JSON text under deterministic decoding; those failed nodes are repaired by applying a simpler, identical category-and-box prompt independently to the four original views. If a view remains token-degenerate, the same RGB is re-encoded at a nearby patch grid for one final retry. The cache records `repair_mode=four_view_local_grounding`. Camera-node, raw, and fused Qwen variants all consume the resulting immutable detection cache, so the repair changes inference packaging but does not confound the representation comparison.

Table 4 establishes completion and representation scale: full fusion compresses 20,107 valid depth candidates to 1,329, while every node-level request completes. The complete independent run uses model string `gpt-5.4` through a hosted chat-compatible endpoint.

3 Complete Results

3.1 Independent RGB-D Variants

Table 5 gives the complete uncertainty record behind the compact main-paper table. The full rows use the same 6,642-node goal-independent topology; minival rows use its separately sampled 339-node control graph. Within each localized backend/FoV group, camera-node, raw, and fused variants reuse identical cached detections. The category-only Qwen panorama row is retained as a separate legacy coverage result and is not part of the localized triplet. Qwen camera-to-raw has 0/2 gains/losses, raw-to-fused has 8/3, and camera-to-fused has 6/3; inaccurate single boxes and useful repeated evidence therefore coexist, and fusion is not monotonic. Full GPT fusion changes 28 raw failures to successes and 22 successes to failures, so its net SR gain is modest even though it reduces candidate count by 93.4%. GPT-5.4 wide raw fails two sofa episodes and fusion succeeds on all episodes; MiMo failures show that wide input and depth do not compensate for a weaker localization response distribution.

Table 6 retains the complete full-scale category workload without mixing it with two-scene counts. Metric grounding improves most categories; fusion raises plant SR from 0.868 to 0.941 and chair SR to 0.995, while bed and sofa expose the remaining ranking-efficiency tradeoff.

3.2 Paired Analyses

We use exact two-sided McNemar tests for paired success outcomes and a 10,000-resample paired episode bootstrap for SPL differences. These analyses were added after the primary summaries and do not select a model or hyperparameter.

Table 7 shows that the same-response camera-node contrasts are not merely a GPT-versus-Qwen model effect. On full validation, depth grounding supplies the dominant gain; fusion compresses the map and slightly raises SR, but reduces SPL. GPT-5.4's mini 90° camera-to-fusion success change is detectable, Qwen's fused SR advantage comes from 8/3 raw-to-fused discordances, and MiMo exposes both a strong standard-FoV metricization gain and significant wide-FoV degradation. No multiple-testing correction is applied, so these values are targeted diagnostics rather than a confirmatory family of tests.

3.3 Topology Density

The topology is nested: the first quarter, half, and three quarters of the farthest-point order are strict subsets of the full map. Official goals are loaded after mapping to label candidates for this geometric ceiling only.

Table 8 shows that oracle SR saturates at half density, whereas SPL continues to improve as candidates approach shorter routes. DTG is not monotonic after SR saturates because any point within the 1 m goal set counts as successful, and added candidates can be closer to the episode start without being the closest point to the target. SPL is computed with the official $\max(\ell, p)$ denominator and remains bounded by one.

3.4 Sequential Candidates

Table 9 deliberately uses official goal distance after every visit. It measures candidate-list potential and remains separate from the autonomous verifier evaluated next.

The auxiliary Qwen sweep covers five separation values; its endpoints and best S@3 are shown here. On mini, GPT-5.4's two raw errors are recovered by the second candidate. At full scale, raw spatial diversity reaches the same 0.975 three-visit ceiling but starts at 0.920, whereas fusion-aware recovery reaches 0.928/0.965 after one/two visits and preserves 0.975 at visit three. The fused-only row draws all three destinations from the 1,329 cluster representatives. The fusion-aware row draws its first destination from that representative index, then selects 2 m-separated backups from the 20,107 pre-fusion metric standoffs. This linked residual bank retains alternative stopping geometry suppressed by one-representative-per-cluster fusion and also provides a fallback when no cluster is retained. It yields eleven paired gains and no losses over fused-only Top-3 (nine sofa and two bed episodes). The official evaluator supplies the metric stop signal in every Top-K row.

3.5 Fresh-Arrival RGB-D/VLM Verification

The deployable loop never reads the goal set. For each unique ranked candidate it captures four fresh square 120° RGB-D views after arrival. View zero faces the stored 3D object estimate; the remaining views rotate by 90°. GPT-5.4 receives these four images and a boxed mapping-time reference, then returns target visibility, stopping-side validity, confidence, an arrival-view index, and a normalized box. The central half of that box indexes aligned depth. Strict acceptance requires visible target, valid stopping side, confidence at least 0.75, and median depth in $[0.25, 2.5]$ m. On rejection, the next geodesic starts at the current candidate.

Table 4: Completed API and candidate-generation records for the independent topology. Each request contains four RGB views. Token totals are provider-reported and are not directly comparable across providers.

Backend	Sensor	Requests	Completed	Raw detections	Valid depth candidates	Fused candidates	Tokens
GPT-5.4 full	120°, 640 × 640	6,642	6,642	21,579	20,107	1,329	44,499,374
GPT-5.4	90°, 640 × 360	339	339	746	674	129	2,031,743
GPT-5.4	120°, 640 × 640	339	339	971	922	130	2,305,616
MiMo-v2.5	90°, 640 × 360	339	339	794	735	125	461,903
MiMo-v2.5	120°, 640 × 640	339	339	1,022	969	108	715,700
Qwen2.5-VL-3B	90°, 640 × 360	339	339	1,064	959	315	local

Table 5: Independent-topology results. "Full" rows use 1,000 episodes/36 scenes; remaining rows use 30 episodes/two scenes. SR intervals are Wilson; SPL intervals use 10,000 episode-level bootstrap resamples.

Backend	Representation	Sensor	SR	SR 95% CI	SPL	SPL 95% CI	DTG
Oracle	evaluator geometry	full, 0.8 m	0.994	[0.987,0.997]	0.992	[0.987,0.996]	0.622
GPT-5.4 full	camera node	120°	0.835	[0.811,0.857]	0.560	[0.539,0.582]	1.039
GPT-5.4 full	raw RGB-D	120°	0.920	[0.902,0.935]	0.603	[0.584,0.621]	0.879
GPT-5.4 full	soft fusion	120°	0.926	[0.908,0.941]	0.586	[0.568,0.604]	0.747
Qwen2.5-VL-3B	category panorama	90°	0.400	[0.246,0.577]	0.344	[0.195,0.501]	2.812
Qwen2.5-VL-3B	localized camera	90°	0.400	[0.246,0.577]	0.233	[0.125,0.347]	5.412
Qwen2.5-VL-3B	raw RGB-D	90°	0.333	[0.192,0.512]	0.188	[0.091,0.294]	5.970
Qwen2.5-VL-3B	soft fusion	90°	0.500	[0.332,0.668]	0.218	[0.128,0.319]	5.791
GPT-5.4	camera node	90°	0.733	[0.556,0.858]	0.568	[0.422,0.706]	1.436
GPT-5.4	raw RGB-D	90°	0.867	[0.703,0.947]	0.681	[0.569,0.779]	0.560
GPT-5.4	soft fusion	90°	0.967	[0.833,0.994]	0.665	[0.580,0.743]	0.375
GPT-5.4	camera node	120°	0.933	[0.787,0.982]	0.704	[0.606,0.796]	0.424
GPT-5.4	raw RGB-D	120°	0.933	[0.787,0.982]	0.713	[0.616,0.801]	0.440
GPT-5.4	soft fusion	120°	1.000	[0.886,1.000]	0.691	[0.609,0.771]	0.136
MiMo-v2.5	camera node	90°	0.200	[0.095,0.373]	0.129	[0.039,0.232]	2.418
MiMo-v2.5	raw RGB-D	90°	0.833	[0.664,0.927]	0.614	[0.482,0.736]	1.414
MiMo-v2.5	soft fusion	90°	0.900	[0.744,0.965]	0.664	[0.550,0.765]	1.011
MiMo-v2.5	camera node	120°	0.567	[0.392,0.726]	0.429	[0.275,0.581]	5.979
MiMo-v2.5	raw RGB-D	120°	0.567	[0.392,0.726]	0.398	[0.257,0.539]	5.981
MiMo-v2.5	soft fusion	120°	0.633	[0.455,0.781]	0.367	[0.245,0.488]	3.840

The 1,000 raw-list episodes reference 656 unique candidates, producing 2,624 fresh RGB-D views and 656/656 completed verifier requests. They consume 4,731,659 tokens with mean/median/95th-percentile latency 20.52/19.45/32.31 seconds per unique candidate. Goal annotations are absent from capture, prompt construction, inference, and STOP/continue; they are read only afterward to form the confusion matrix and ObjectNav metrics.

Table 10 calibrates the arrival classifier separately from candidate quality. Strict raw verification ends 912 episodes successfully, with candidate ranks 1/2/3 accounting for 888/55/3 accepts. The collapsed-representative fused ablation uses 572 unique candidates and 2,288 arrival views; exact observations reuse 314 raw-list responses and 258 new requests. This ablation exposed the implementation issue corrected by the final hierarchy: fusion now controls the primary hypothesis while residual metric standoffs remain available for Top-3 recovery.

3.6 Coverage-Controlled Corruption

The target-view oracle minival topology is perturbed across retained-node fractions $\{1, 0.75, 0.5, 0.25\}$, semantic dropout probabilities $\{0, 0.1, 0.25, 0.5\}$, and false-positive probabilities $\{0, 0.05, 0.15\}$. Every condition uses 20 seeds, giving 960 map perturbations and 28,800 episode evaluations. A false positive adds a randomly selected incorrect benchmark category to a retained node; dropout independently removes a correct label.

Table 11 shows that dropout and false positives compound rather than substitute for one another: at 50% retained nodes, 0.25 dropout and 0.15 false positives reduce mean SR to 0.602. This experiment is a sensitivity analysis on a privileged target-view map. It does not estimate the error distribution of Qwen, GPT-5.4, or MiMo.

4 Qualitative and Visual Audit

Figure 1 shows the compression effect spatially: fusion removes repeated projected points while retaining category coverage across floors. Figure 2 then ties three ag-

Table 6: Consolidated full-validation per-category audit. Entries are SR/SPL over the same 1,000 episodes. Target-view Qwen isolates semantic retrieval under controlled coverage; the three goal-independent GPT-5.4 columns isolate spatial representation.

	Target-view semantic isolation	Goal-independent metric-semantic mapping		
Category (n)	Qwen confidence	Camera node	Raw RGB-D	Soft fusion
Chair (195)	0.708/0.454	0.841/0.415	0.974/0.488	0.995/0.498
Bed (165)	0.818/0.550	0.891/0.673	0.988/0.716	0.976/0.631
Plant (152)	0.724/0.462	0.862/0.473	0.868/0.481	0.941/0.532
Toilet (166)	0.633/0.440	0.934/0.771	0.940/0.746	0.940/0.673
TV/monitor (135)	0.430/0.356	0.644/0.455	0.800/0.578	0.778/0.571
Sofa (187)	0.524/0.361	0.807/0.572	0.914/0.611	0.893/0.613

Table 7: Paired changes. "Gain/loss" counts success discordances from the first variant to the second.

Comparison	Gain/loss	p	ΔSPL	95% CI
Full camera → raw	93/8	1.74×10^{-19}	0.042	[0.028,0.057]
Full raw → fused	28/22	0.480	−0.017	[−0.032,−0.002]
Full camera → fused	113/22	6.09×10^{-16}	0.026	[0.005,0.047]
GPT 90° raw → fused	3/0	0.250	-0.016	[-0.074,0.047]
GPT 90° camera → raw	7/3	0.344	0.113	[-0.014,0.243]
GPT 90° camera → fused	7/0	0.0156	0.097	[-0.014,0.211]
GPT fused 90° → 120°	1/0	1.000	0.026	[-0.038,0.111]
Qwen raw → fused	8/3	0.227	0.030	[-0.103,0.167]
MiMo fused 90° → 120°	0/8	0.0078	-0.297	[-0.439,-0.157]
MiMo 90° camera → fused	21/0	< 0.0001	0.535	[0.402,0.662]

Table 8: Independent topology density ceiling on minival.

Node fraction	Approx. nodes	SR	SPL	DTG
0.25	85	0.900	0.794	0.784
0.50	170	1.000	0.964	0.376
0.75	255	1.000	0.995	0.508
1.00	339	1.000	0.997	0.555

gregate findings to raw evidence—wider visibility, source-supported fusion, and verifier-controlled recovery—and separately records the physical mapping-query-execution sequence. Figure 3 makes the same-response comparison visual across model and FoV controls, whereas Figure 4 restores full-validation scale and separates raw single/Top-3, fused single, and fusion-aware recovery. Together, the figures move from map structure, through mechanism audits, to aggregate outcome rather than serving as interchangeable qualitative examples.

Representative stills in Figures 1–4 make the qualitative mechanisms directly inspectable in this document. Additional demonstrations are available from the project page linked after the abstract.

5 Physical-System Implementation

The physical platform is a Yahboom X3 mobile robot running ROS 2 Humble (Macenski et al. 2022). A Gemini 336L RGB-D camera supplies visual observations, an RPLIDAR S2 supplies planar laser scans, `slam_toolbox` (Macenski and Jambrecic 2021) maintains the occupancy map and localization estimate, and Nav2 (Macenski et al. 2020) executes metric goals. Figure 5 expands the physical sequence shown in the main paper; additional demonstrations are available on the project page.

Perception supplies aligned RGB-D observations, the persistent manager owns the reusable graph shown in Figure 5(b), the language interface normalizes the request and exposes its supporting views, and semantic/topological planning dispatches the selected standoff through Nav2. Interaction state retains the remaining candidate list across navigation legs. Quantitative SR/SPL results in the paper come from the simulator benchmark; the physical sequence documents the implemented robot interface.

6 Additional Limitations and Intended Scope

The main paper evaluates category-level ObjectNav. The natural-language interface is part of the reusable robot system, but relational queries such as "the vase next to the sofa" are not represented in HM3D ObjectNav-v2 and are not quantitatively evaluated here. The six-class prompts likewise do not establish open-vocabulary performance beyond the benchmark categories.

The independent map is static. Object motion, furniture rearrangement, appearance change, and map aging are not evaluated. The implemented verifier uses fresh simulator RGB-D and benchmark-category prompts; physical-world calibration, occlusion change, and sensor noise may alter its precision/recall. The oracle-feedback Top-K table remains only a candidate-list ceiling and is never used by the autonomous loop. The simulator uses continuous navmesh shortest paths, so collision recovery, actuator noise, localization drift, camera motion blur, and discrete action limits are absent. GPT-5.4 and MiMo-v2.5 are accessed through hosted endpoints whose weights and future serving behavior are not controlled by the authors. The full independent RGB-D result covers one 1,000-episode HM3D-v2 validation split and one primary semantic backend. Its confidence intervals quantify episode sampling uncertainty, not cross-dataset, model-serving, or real-world variation.

Table 9: Sequential candidate and separation sensitivity. A positive separation greedily removes candidates near earlier hypotheses. The full-scale policy block exposes both the matched 2 m comparison and nearby sensitivity controls.

Map	Separation	Success@1	Success@2	Success@3	SPL@1	SPL@2	SPL@3
Qwen panorama	none	0.400	0.500	0.500	0.344	0.396	0.396
Qwen panorama	5.0 m	0.400	0.567	0.700	0.344	0.392	0.425
GPT-5.4 raw 120°	none	0.933	1.000	1.000	0.713	0.745	0.745
GPT-5.4 full raw	none	0.920	0.955	0.957	0.603	0.610	0.611
GPT-5.4 full raw	2 m	0.920	0.963	**0.975**	0.603	0.612	**0.616**
GPT-5.4 full raw	3 m	0.920	0.965	0.975	0.603	0.614	0.616
GPT-5.4 full fused	none	0.926	0.952	0.964	0.586	0.594	0.596
GPT-5.4 full fused	2 m	0.926	0.952	0.964	0.586	0.594	0.596
GPT-5.4 full fused	3 m	0.926	0.950	0.962	0.586	0.593	0.596
GPT-5.4 fusion-aware residual	2 m	**0.928**	**0.965**	**0.975**	0.588	0.598	0.601

Table 10: Autonomous full-validation arrival verification. S@k and SPL@k execute at most k candidates. P/R compares verifier decisions with the post-hoc official success test over attempted visits. The RGB-D-only rule reparses the same cached responses but ignores the VLM stopping-side flag.

Candidates	Decision rule	S@1	S@2	S@3	SPL@1	SPL@2	SPL@3	P	R	Attempts
Raw, 3 m	strict dual geometry	0.854	0.909	0.912	0.560	0.573	0.573	0.964	0.915	1.110
Raw, 3 m	category + RGB-D range	0.865	0.903	0.905	0.569	0.578	0.578	0.956	0.924	1.091
Fused	strict dual geometry	0.821	0.901	0.901	0.526	0.546	0.546	0.947	0.891	1.087

Table 11: Complete corruption grid: mean SR over 20 seeds. Each three-column group is false-positive probability 0/0.05/0.15.

Nodes	Dropout 0			Dropout 0.10			Dropout 0.25			Dropout 0.50		
	FP 0	FP 0.05	FP 0.15	FP 0	FP 0.05	FP 0.15	FP 0	FP 0.05	FP 0.15	FP 0	FP 0.05	FP 0.15
1.00	1.000	0.937	0.840	0.970	0.898	0.805	0.907	0.828	0.752	0.763	0.700	0.680
0.75	0.907	0.893	0.768	0.885	0.853	0.730	0.833	0.807	0.713	0.690	0.653	0.642
0.50	0.768	0.715	0.682	0.738	0.695	0.653	0.678	0.647	0.602	0.570	0.545	0.520
0.25	0.530	0.497	0.497	0.488	0.455	0.458	0.430	0.407	0.432	0.347	0.333	0.363

Scene 1: raw projected candidates

Scene 1: 3D-fused candidates

Scene 2: raw projected candidates

Scene 2: 3D-fused candidates

Figure 1: Actual semantic maps for both minival scenes. Each row uses a common Habitat top-down frame. Raw panels show every valid depth-projected candidate; fused panels show the candidates remaining after 3D and reachability clustering. Official goal regions are absent.

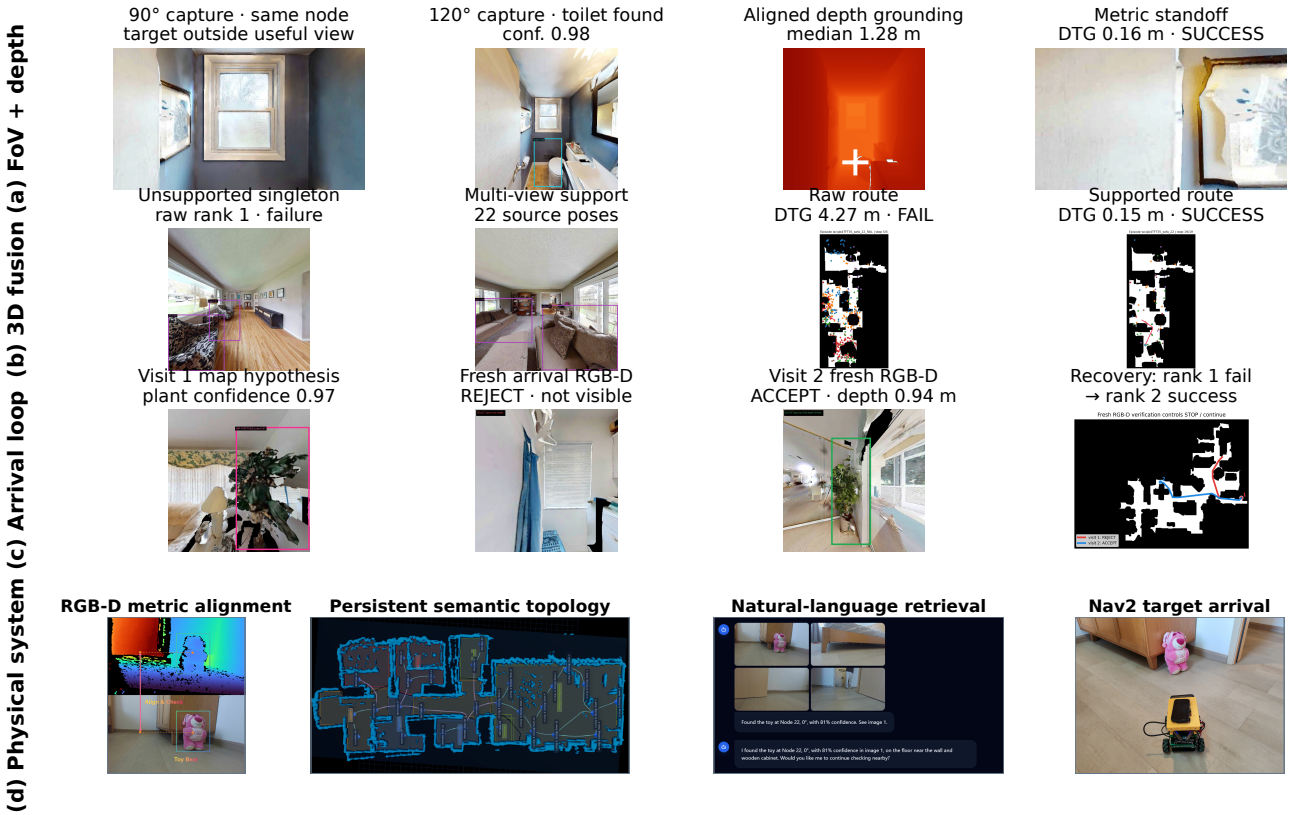

Figure 2: Source-traceable FoV, fusion, closed-loop arrival, and physical-system audit. Panels (a)–(c) link to their original RGB/depth file, pose, VLM box, candidate, and evaluated episode. The third row is a full-validation plant case: rank 1 is truly outside the goal set and rejected from fresh RGB-D; rank 2 is accepted at 0.94 m measured depth and succeeds. Panel (d) records the physical pipeline: RGB-D alignment, the surveyed semantic topology, language-conditioned toy retrieval, and Nav2 target arrival.

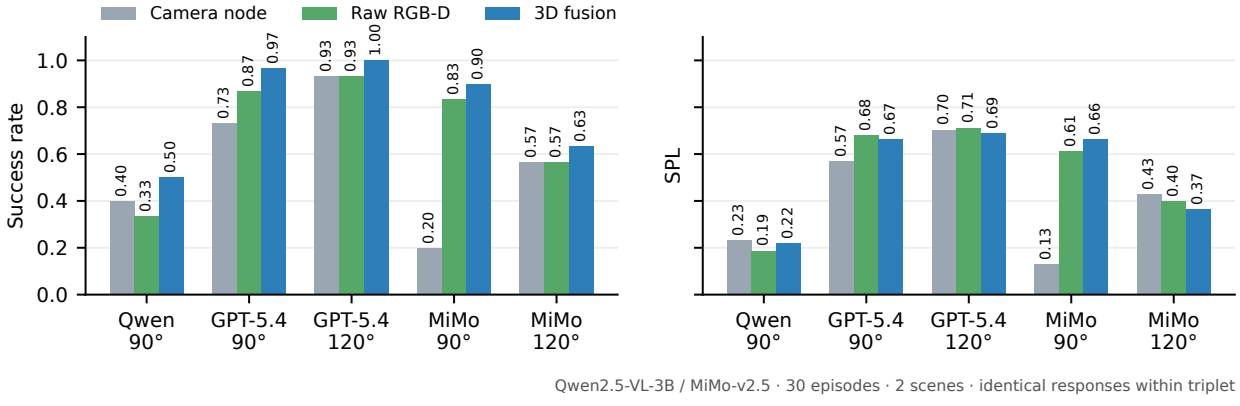

Figure 3: Same-response representation and model ablations on independent-topology minival. Each backend/FoV triplet reuses exact cached VLM responses and changes only whether category evidence is attached to the camera, back-projected with depth, or fused in 3D.

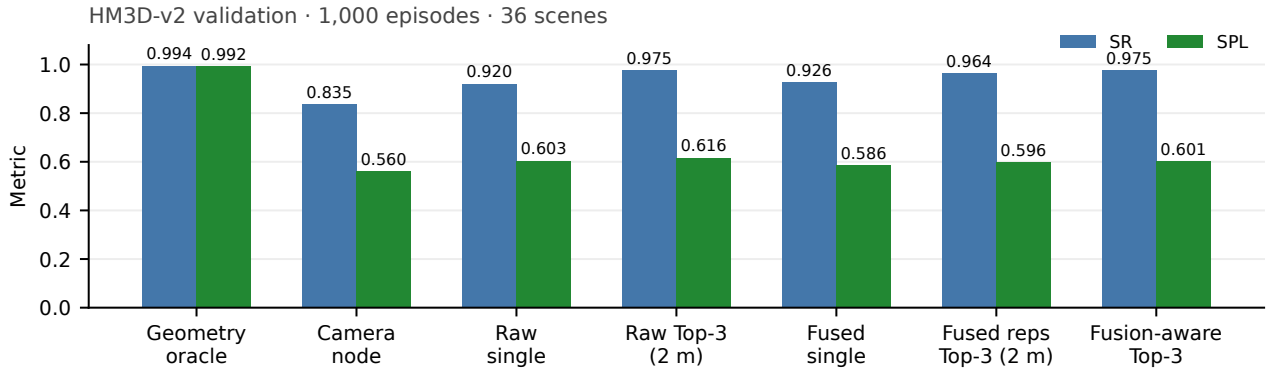

Figure 4: Full-validation independent-topology results. Geometry is the evaluator-labeled ceiling; camera, raw, and fused variants reuse identical GPT-5.4 mapping responses. Under matched 2 m diversity, raw and fusion-aware Top-3 both reach 0.975 S@3; Table 9 resolves the stronger early recovery of the fusion-aware policy.

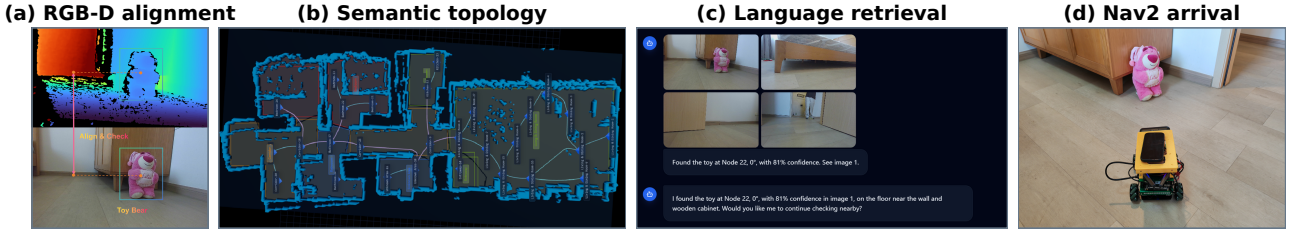

Figure 5: Physical-system realization on the Yahboom X3. (a) Aligned Gemini 336L depth and RGB localize the toy observation. (b) The one-time survey produces a persistent spatial-semantic topology over the mapped environment. (c) The language interface retrieves the toy at Node 22 with 0.81 confidence and presents the supporting views. (d) Nav2 executes the selected target and brings the robot to the toy. These panels document qualitative system integration; SR/SPL results are measured in simulation.

References

Anderson, P.; Chang, A.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; and Zamir, A. R. 2018. On Evaluation of Embodied Navigation Agents. *arXiv preprint arXiv:1807.06757*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Batra, D.; Gokaslan, A.; Kembhavi, A.; Maksymets, O.; Mottaghi, R.; Savva, M.; Toshev, A.; and Wijmans, E. 2020. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv preprint arXiv:2006.13171*.

Cai, W.; Huang, S.; Cheng, G.; Long, Y.; Gao, P.; Sun, C.; and Dong, H. 2024. Bridging Zero-Shot Object Navigation and Foundation Models through Pixel-Guided Navigation Skill. In *ICRA*, 5228–5234.

Cai, Y.; Li, Z.; Wang, M.; Bao, M.; Zhu, H.; Bai, R.; Zhao, D.; Li, Z.; Wang, W.; Yau, W.-Y.; Zhang, J.; and Lv, C. 2026. IntentNav: Learning Spatial-Visual Object Navigation from Human Demonstrations. *arXiv preprint arXiv:2606.08029*.

Campari, T.; Lamanna, L.; Traverso, P.; Serafini, L.; and Ballan, L. 2022. Online Learning of Reusable Abstract Models for Object Goal Navigation. In *CVPR*, 14870–14879.

Cao, Y.; Zhang, J.; Yu, Z.; Liu, S.; Qin, Z.; Zou, Q.; Du, B.; and Xu, K. 2025. CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs. In *ICCV*, 9550–9560.

Chabal, T.; Chen, S.; Ponce, J.; and Schmid, C. 2025. FOM-Nav: Frontier-Object Maps for Object Goal Navigation. *arXiv preprint arXiv:2512.01009*.

Chaplot, D. S.; Gandhi, D.; Gupta, A.; and Salakhutdinov, R. 2020. Object Goal Navigation Using Goal-Oriented Semantic Exploration. In *NeurIPS*, volume 33, 4247–4258.

Gadre, S. Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; and Song, S. 2023. CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. In *CVPR*, 23171–23181.

Huang, C.; Mees, O.; Zeng, A.; and Burgard, W. 2023. Visual Language Maps for Robot Navigation. In *ICRA*, 10608–10615.

Jatavallabhula, K. M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; Tewari, A.; Tenenbaum, J. B.; de Melo, C. M.; Krishna, M.; Paull, L.; Shkurti, F.; and Torralba, A. 2023. ConceptFusion: Open-Set Multimodal 3D Mapping. In *Robotics: Science and Systems*.

Kim, N.; Kwon, O.; Yoo, H.; Choi, Y.; Park, J.; and Oh, S. 2022. Topological Semantic Graph Memory for Image-Goal Navigation. *arXiv preprint arXiv:2209.08274*.

Ko, P.-C.; Su, H.-T.; Chen, C.-Y.; Yeh, J.-F.; Sun, M.; and Hsu, W. H. 2025. Context-Aware Replanning with Pre-Explored Semantic Map for Object Navigation. In *CoRL*, volume 270, 4253–4267.

Liu, P.; Zhang, Q.; Peng, D.; Zhang, L.; Qin, Y.; Zhou, H.; Ma, J.; Xu, R.; and Ji, Y. 2025. TopoNav: Topological Graphs as a Key Enabler for Advanced Object Navigation. *arXiv preprint arXiv:2509.01364*.

Long, Y.; Cai, W.; Wang, H.; Zhan, G.; and Dong, H. 2025. InstructNav: Zero-Shot System for Generic Instruction Navigation in Unexplored Environment. In *CoRL*, volume 270, 2049–2060.

Macenski, S.; Foote, T.; Gerkey, B.; Lalancette, C.; and Woodall, W. 2022. Robot Operating System 2: Design, Architecture, and Uses in the Wild. *Science Robotics*, 7(66): eabm6074.

Macenski, S.; and Jambrecic, I. 2021. SLAM Toolbox: SLAM for the Dynamic World. *Journal of Open Source Software*, 6(61): 2783.

Macenski, S.; Martín, F.; White, R.; and Ginés Clavero, J. 2020. The Marathon 2: A Navigation System. In *IROS*, 2718–2725.

Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; and Batra, D. 2022. ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings. In *NeurIPS*, volume 35, 32340–32352.

Nie, D.; Guo, X.; Duan, Y.; Zhang, R.; and Chen, L. 2025. WMNav: Integrating Vision-Language Models into World Models for Object Goal Navigation. *arXiv preprint arXiv:2503.02247*.

OpenAI. 2026. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/.

Peng, D.; Ma, F.; and Ma, J. 2026. Structured observation language for efficient and generalizable vision-language navigation. *arXiv preprint arXiv:2603.27577*.

Ramakrishnan, S. K.; Chaplot, D. S.; Al-Halah, Z.; Malik, J.; and Grauman, K. 2022. PONI: Potential Functions for ObjectGoal Navigation with Interaction-Free Learning. In *CVPR*, 18890–18900.

Ramakrishnan, S. K.; Gokaslan, A.; Wijmans, E.; Maksymets, O.; Clegg, A.; Turner, J.; Undersander, E.; Galuba, W.; Westbury, A.; Chang, A. X.; Savva, M.; Zhao, Y.; and Batra, D. 2021. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-Scale 3D Environments for Embodied AI. In *NeurIPS Datasets and Benchmarks Track*, volume 1.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; Parikh, D.; and Batra, D. 2019. Habitat: A Platform for Embodied AI Research. In *ICCV*, 9339–9347.

Shafiullah, N. M. M.; Paxton, C.; Pinto, L.; Chintala, S.; and Szlam, A. 2023. CLIP-Fields: Weakly Supervised Semantic Fields for Robotic Memory. In *Robotics: Science and Systems*.

Takmaz, A.; Fedele, E.; Sumner, R. W.; Pollefeys, M.; Tombari, F.; and Engelmann, F. 2023. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, volume 36, 68367–68390.

Wang, H.; Li, Z.; Zhang, Y.; He, Z.; Jiang, L.; Li, K.; Zhao, Y.; Fan, L.; Hou, W.; Liang, T.; Wen, Y.; and Gu, D. 2026a. ConsistNav: Closing the Action Consistency Gap in Zero-Shot Object Navigation with Semantic Executive Control. *arXiv preprint arXiv:2605.09869*.

Wang, Y.; Zhang, S.; Zhang, K.; Song, X.; Du, S.; and Jiang, S. 2026b. TrajRAG: Retrieving Geometric-Semantic Experience for Zero-Shot Object Navigation. In *CVPR*, 15166–15176.

Wu, P.; Mu, Y.; Wu, B.; Hou, Y.; Ma, J.; Zhang, S.; and Liu, C. 2024. VoroNav: Voronoi-Based Zero-Shot Object Navigation with Large Language Model. In *ICML*, volume 235, 53757–53775.

Xiaomi MiMo Team. 2026. MiMo-V2.5. https://huggingface.co/collections/XiaomiMiMo/mimo-v25.

Yadav, K.; Krantz, J.; Ramrakhya, R.; Ramakrishnan, S. K.; Yang, J.; Wang, A.; Turner, J.; Gokaslan, A.; Berges, V.-P.; Mottaghi, R.; Maksymets, O.; Chang, A. X.; Savva, M.; Clegg, A.; Chaplot, D. S.; and Batra, D. 2023a. Habitat Challenge 2023. https://aihabitat.org/challenge/2023/.

Yadav, K.; Ramrakhya, R.; Ramakrishnan, S. K.; Gervet, T.; Turner, J.; Gokaslan, A.; Maestre, N.; Chang, A. X.; Batra, D.; Savva, M.; Clegg, A. W.; and Chaplot, D. S. 2023b. Habitat-Matterport 3D Semantics Dataset. In *CVPR*, 4927–4936.

Yin, H.; Xu, X.; Wu, Z.; Zhou, J.; and Lu, J. 2024. SG-Nav: Online 3D Scene Graph Prompting for LLM-Based Zero-Shot Object Navigation. In *NeurIPS*, volume 37, 10012–10036.

Yin, H.; Xu, X.; Zhao, L.; Wang, Z.; Zhou, J.; and Lu, J. 2025. UniGoal: Towards Universal Zero-Shot Goal-Oriented Navigation. In *CVPR*, 19057–19066.

Yokoyama, N.; Ha, S.; Batra, D.; Wang, J.; and Bucher, B. 2024. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *ICRA*, 42–48.

Yu, B.; Kasaei, H.; and Cao, M. 2023. L3MVN: Leveraging Large Language Models for Visual Target Navigation. In *IROS*, 3554–3560.

Yuan, S.; Huang, H.; Hao, Y.; Wen, C.; Tzes, A.; and Fang, Y. 2024. GAMap: Zero-Shot Object Goal Navigation with Multi-Scale Geometric-Affordance Guidance. In *NeurIPS*, volume 37, 39386–39408.

Zhang, M.; Du, Y.; Wu, C.; Zhou, J.; Qi, Z.; Ma, J.; and Zhou, B. 2025a. ApexNav: An Adaptive Exploration Strategy for Zero-Shot Object Navigation with Target-Centric Semantic Fusion. *IEEE Robotics and Automation Letters*, 10(11): 11530–11537.

Zhang, S.; Yu, X.; Song, X.; Wang, X.; and Jiang, S. 2024. Imagine Before Go: Self-Supervised Generative Map for Object Goal Navigation. In *CVPR*, 16414–16425.

Zhang, S.; Yu, X.; Song, X.; Wang, Y.; and Jiang, S. 2025b. Function-centric Bayesian Network for Zero-Shot Object Goal Navigation. In *ICCV*, 19535–19545.

Zhou, K.; Zheng, K.; Pryor, C.; Shen, Y.; Jin, H.; Getoor, L.; and Wang, X. E. 2023. ESC: Exploration with Soft Commonsense Constraints for Zero-Shot Object Navigation. In *ICML*, volume 202, 42829–42842.