Decoupled Complementary Fields on 3D Gaussian Maps for Embodied Navigation and Reasoning

Jiahao Zhao[1] **Yaonan Wang**[1] **Mingtao Feng**[2] **Hui Zhang**[1] **He Xie**[1]
Renjie Ding[1] **Zijie Wu**[1,2*]

[1]National Engineering Research Center of Robot Visual Perception and Control Technology, School of Artificial Intelligence and Robotics, Hunan University
[2]School of Artificial Intelligence, Xidian University

{zhaojiahao, yaonan, mintfeng, zhanghui1983, xiehe}@hnu.edu.cn
{regi, wuzijieeee}@hnu.edu.cn

Abstract

Embodied navigation and reasoning require an agent to actively acquire, organize, and verify task-relevant evidence in previously unseen environments. Recent methods typically rely on semantic maps, scene graphs, vision-language priors, or structured memory, but often underutilize persistent, fine-grained 3D evidence and entangle query-related ambiguity with environment-side reliability. This coupling can make evidence acquisition inefficient and final decisions vulnerable to weak or transient observations. In this work, we propose DCF-Nav, an embodied navigation and reasoning framework built on persistent 3D Gaussian maps. DCF-Nav constructs a query-conditioned object-centric Gaussian navigation state that preserves fine-grained 3D evidence across viewpoints while providing an executable spatial interface for region reasoning and navigation. We further formulate Decoupled Complementary Fields, consisting of a Query-Ambiguity Field that captures where task-relevant evidence remains unresolved and an Environment-Readiness Field that estimates where the environment is sufficiently established for reliable motion and inspection. By coupling these complementary fields, DCF-Nav selects informative frontiers and instantiates executable local goals for hierarchical navigation, while progressively verifying task-relevant candidates using persistent volumetric Gaussian evidence before stopping or answer handoff. Experimental results on three complementary settings demonstrate that DCF-Nav significantly enhances task-relevant evidence acquisition and reliable decision making across embodied reasoning and multimodal lifelong navigation, highlighting the effectiveness of persistent 3D evidence and decoupled complementary fields for embodied agents.

1 Introduction

Embodied navigation and reasoning is a fundamental capability for autonomous agents operating in previously unseen environments, where visual perception must inform action [1, 2] and memory, exploration, and decision making must be coordinated under partial observability [3, 4, 5]. Recent tasks further extend this challenge from locating targets to acquiring and verifying task-relevant evidence across long-horizon interactions, including multimodal navigation [6, 7] and question-conditioned exploration [8, 9]. Such settings require navigation decisions to be grounded in persistent 3D evidence while remaining sensitive to what is still unresolved for the current query. However, existing agents often organize observations through compact semantic maps, symbolic abstractions, or transient vision-language cues [3, 10, 11, 12], which can leave fine-grained volumetric evidence underused

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

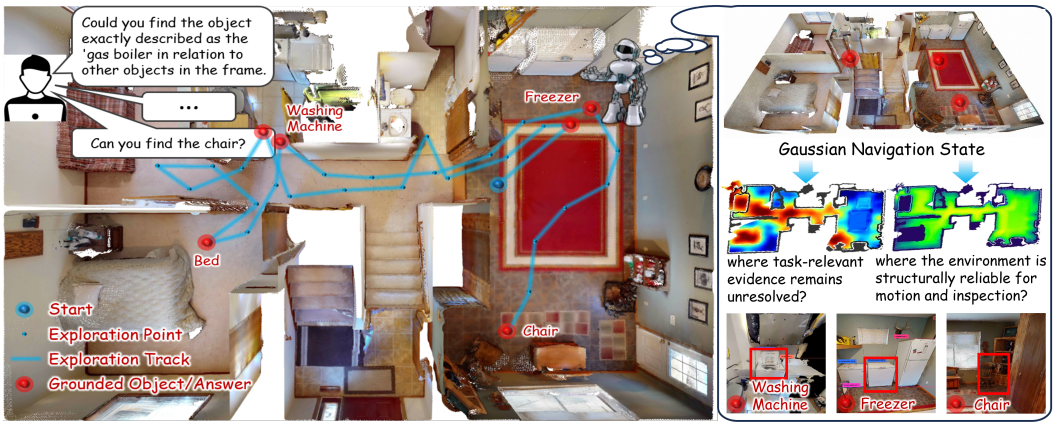


Figure 1: Overview of DCF-Nav for efficient embodied navigation and reasoning. By maintaining a persistent Gaussian Navigation State and decoupling query-related ambiguity from environment readiness, DCF-Nav prioritizes informative and reliable regions, reuses accumulated evidence across sequential queries, and reduces redundant exploration and backtracking before reliable object grounding or answer generation.

and obscure the distinction between query-related ambiguity and environment-side reliability. We propose *Decoupled Complementary Fields on 3D Gaussian Maps (DCF-Nav)*, an embodied navigation and reasoning framework that maintains persistent volumetric evidence, separates query-related ambiguity from environment-side readiness, and grounds exploration, motion, and final decisions through progressive volumetric verification.

To support embodied navigation and reasoning in unseen environments, prior works have first sought to maintain navigation memory beyond the current egocentric observation. Semantic-map-based methods project observations into spatial memories for object search and goal-directed exploration [3, 13], while scene graphs organize objects and their relationships [14, 15, 16, 17], and vision-language maps and grid memories support spatial grounding and navigation reasoning [18, 19]. More recent 3D scene memories preserve multi-view visual context and connect accumulated observations to frontier-based exploration [9]. This development makes 3D Gaussian representations particularly relevant. 3D Gaussian Splatting provides an explicit scene representation [20], Gaussian-based SLAM enables online RGB-D mapping across viewpoints [21, 22], and language- or semantics-enhanced Gaussians support query-relevant grounding and navigation [23, 24, 25]. GSMem further combines persistent Gaussian memory with retrieval-rendering and semantic–geometric exploration [26]. However, richer memory alone does not specify which task evidence remains unresolved, which regions are reliable enough to inspect, or when accumulated evidence is sufficient. Accordingly, another line of work improves query-aware decisions through semantic priors and exploration potentials [27, 13], vision-language-guided navigation [11, 28, 12], and confidence-based stopping or hypothesis verification [29, 30]. Together, these studies advance persistent 3D representation and query-aware decision making, but leave unresolved how accumulated 3D evidence should guide exploration and support candidate verification.

More specifically, existing representations and navigation policies exhibit two related limitations. Semantic maps, scene graphs, and structured memories provide compact abstractions, but they often compress away fine-grained geometry, appearance, and multi-view semantic evidence that are important for verification [18, 14, 9]. Conversely, Gaussian representations preserve rich volumetric evidence, yet existing Gaussian-based methods are often designed for reconstruction, semantic grounding, or memory retrieval rather than for explicitly structuring navigation decisions [20, 24]. Moreover, many navigation policies still couple task-related uncertainty and environment-side reliability within a single exploration or frontier score [31, 32], which can obscure the difference between informative but weakly established regions and structurally reliable but query-irrelevant regions. This coupling can lead to inefficient evidence acquisition, premature commitment to transient cues, and unreliable stopping or answer handoff.

To address these limitations, we introduce DCF-Nav for embodied navigation and reasoning. We first construct a query-conditioned object-centric Gaussian navigation state that integrates persistent

semantic Gaussians, object-centric volumetric units, query-conditioned priors, and query-agnostic BEV support. The state preserves fine-grained evidence across viewpoints while providing a compact object-level reasoning interface and executable spatial support. Built on this state, we formulate Decoupled Complementary Fields to separate two decision signals that are often entangled during embodied exploration: where query-relevant evidence remains unresolved, and where the environment is structurally reliable for motion and inspection. We further couple these fields for hierarchical region-and-frontier navigation, selecting informative frontiers and instantiating executable local goals for closed-loop motion. Progressive volumetric verification evaluates persistent candidates using candidate-specific Gaussian evidence and multi-view consistency before stopping or answer handoff.

Our contributions are threefold:

- We construct a query-conditioned object-centric Gaussian navigation state as the common substrate for embodied navigation and reasoning. Built on online 3D Gaussian Splatting, the state integrates persistent semantic Gaussians, object-centric volumetric units, query-conditioned priors, and query-agnostic BEV support, preserving fine-grained 3D evidence across viewpoints while providing an executable spatial interface for region reasoning and navigation.
- We formulate Decoupled Complementary Fields to separate two decision signals that are often entangled in embodied navigation: query-related ambiguity and environment-side readiness. The Query-Ambiguity Field captures where task-relevant evidence remains unresolved, while the Environment-Readiness Field estimates where the environment is structurally reliable for motion and inspection.
- We couple these complementary fields for hierarchical region-and-frontier navigation, selecting informative frontiers and instantiating executable local goals for closed-loop motion. We further introduce progressive volumetric verification, which evaluates persistent candidates using candidate-specific Gaussian evidence and multi-view consistency before stopping or answer handoff.

2 Related Work

Embodied Navigation and Reasoning. Object-goal navigation has progressed from semantic-map-based exploration [3] and complementary navigation potentials [13] to zero-shot and vision-language-guided reasoning [27, 11, 28, 12]. Recent methods further support instance-specific and heterogeneous goal specifications [33, 34, 10], while GOAT-Bench and OpenEQA extend evaluation to multimodal lifelong navigation and evidence-dependent question answering [6, 8]. Advances in embodied decision making also integrate object grounding with frontier selection [35] and connect semantic planning with geometry-grounded navigation [36]. Beyond navigation planning, reliable decisions require reasoning about uncertain or insufficient evidence. Prior studies address this challenge through target-belief estimation [31], uncertainty-aware Gaussian maps [32], confidence-calibrated answering [29], and hypothesis-driven correction [30]. Recent methods further investigate contextual ambiguity and multi-view verification [37], as well as accumulated evidence for target confirmation [38]. However, identifying unresolved task-relevant evidence does not establish whether regions are reliable to inspect or whether a candidate has sufficient support for a final decision.

Persistent 3D Scene Memory and Gaussian Representations. Persistent spatial memories organize accumulated observations for subsequent grounding and reasoning. Spatial language maps and grid-based memories encode semantic and geometric context [18, 39, 40, 19], while scene graphs organize objects and their spatial relationships across observations [14, 16]. Related research further exploits object-relation knowledge for structured 3D scene generation [41] and uncertainty-aware representations for robust 3D relationship prediction [42, 43]. More recent embodied systems, such as 3D-Mem, preserve visual context through persistent memory snapshots linked to exploration [9]. At the geometric level, hierarchical representations have been investigated for scalable 4D LiDAR generation [44], while 3D Gaussian Splatting provides explicit primitives for scene reconstruction [20, 45]. Gaussian-based SLAM enables online RGB-D mapping across viewpoints [21, 22]. Multimodal 3D modeling has also explored sketch- and text-conditioned geometry and appearance generation [46], while language-aligned Gaussian representations associate observed geometry with semantic features for open-vocabulary understanding [24, 47, 23]. Recent approaches further connect Gaussian representations with open-set semantic grouping [25] and persistent Gaussian memory for retrieval, rendering, and semantic–geometric exploration [26]. However, organizing persistent 3D evidence for both navigation-time reasoning and candidate-level verification remains challenging.

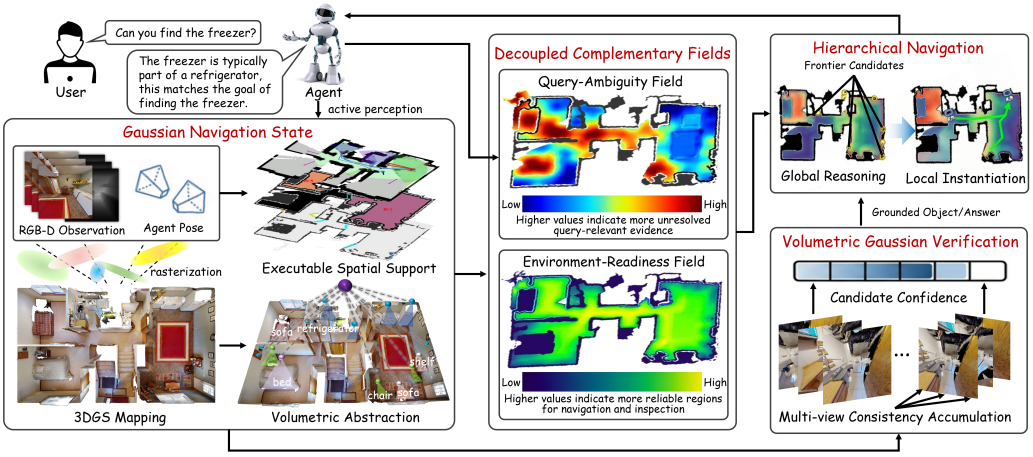


Figure 2: Architecture of DCF-Nav. Given a task query and posed RGB-D observations, DCF-Nav incrementally constructs a query-conditioned object-centric Gaussian navigation state that preserves persistent fine-grained 3D evidence while providing executable spatial support. Decoupled Complementary Fields separate Query-Ambiguity, which captures where task-relevant evidence remains unresolved, from Environment-Readiness, which evaluates whether regions are sufficiently established for reliable navigation and inspection. The coupled fields guide hierarchical region-and-frontier navigation and local goal instantiation, while volumetric verification grounds stopping or answer decisions in persistent multi-view Gaussian evidence.

3 Method

Overview. We propose *Decoupled Complementary Fields on 3D Gaussian Maps (DCF-Nav)* for embodied navigation and reasoning, as illustrated in Fig. 2. DCF-Nav maintains a persistent object-centric Gaussian navigation state, separates query-related ambiguity from environment readiness for hierarchical navigation, and verifies task-relevant candidates using persistent volumetric evidence. Sec. 3.1 introduces the Gaussian navigation state, Sec. 3.2 formulates the complementary fields, Sec. 3.3 describes hierarchical navigation, and Sec. 3.4 presents volumetric Gaussian verification.

3.1 Query-Conditioned Object-Centric Gaussian Navigation State

Embodied navigation requires evidence that is both persistent for reasoning and executable for motion. Image-space observations provide rich local cues but do not by themselves establish persistent spatial correspondence across viewpoints, while planar maps support planning but compress fine-grained 3D evidence. We therefore maintain a query-conditioned object-centric Gaussian navigation state

$$\mathcal{M}_t^q = (\mathcal{G}_t, \mathcal{U}_t, \Lambda^q, \mathbf{B}_t)\,, \tag{1}$$

where $\mathcal{G}_t$ is a persistent semantic Gaussian map, $\mathcal{U}_t$ is an object-centric volumetric abstraction, Λ^q contains query-conditioned priors, and $\mathbf{B}_t$ is the executable BEV support for region construction and local planning. The Gaussian map, object units, and BEV support share a metric frame. Query conditioning is introduced through Λ^q and the on-demand evaluation of the persistent scene evidence.

Persistent Semantic Gaussian Map. At timestep t, the agent observes $\mathcal{O}_t = (V_t, D_t, \mathbf{T}_t)$, where V_t is the RGB image, D_t is the depth image, and $\mathbf{T}_t \in SE(3)$ is the camera pose. We represent the observed scene as semantic Gaussian primitives

$$\mathcal{G}_t = \{g_i\}_{i=1}^{N_t}\,, \qquad g_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i, \mathbf{p}_i, \mathbf{f}_i, \rho_i)\,. \tag{2}$$

Here $\boldsymbol{\mu}_i \in \mathbb{R}^3$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$ denote the 3D mean and covariance, respectively. The scalar $\alpha_i \in [0,1]$ denotes opacity, and $\mathbf{c}_i$ encodes appearance. The semantic attributes comprise a category distribution $\mathbf{p}_i$, a normalized open-vocabulary feature $\mathbf{f}_i$, and an accumulated reliability score ρ_i.

Given the current RGB-D observation and pose, new Gaussians are initialized in newly observed or insufficiently explained regions, while a local active subset of existing Gaussians is refined through

incremental Gaussian splatting [20, 22]. Let $(\hat{V}_t, \hat{D}_t)$ denote the rendered RGB-D outputs. To maintain photometric and geometric consistency of the persistent Gaussian state, we optimize

$$\mathcal{L}_t^{\text{map}} = (1 - \lambda_{\text{ssim}})\mathcal{L}_t^{\text{rgb}} + \lambda_{\text{ssim}}\mathcal{L}_t^{\text{ssim}} + \lambda_{\text{depth}}\mathcal{L}_t^{\text{depth}}, \tag{3}$$

where $\mathcal{L}_t^{\text{rgb}}$ and $\mathcal{L}_t^{\text{depth}}$ denote RGB and valid-depth L_1 reconstruction losses, respectively, and $\mathcal{L}_t^{\text{ssim}} = 1 - \text{SSIM}(V_t, \hat{V}_t)$. Frame-level semantic predictions and visual-language features are then lifted to visible Gaussians via rendering weights.

Object-Centric Volumetric Abstraction. Reasoning over all primitives is inefficient and sensitive to local noise. We aggregate spatially and semantically compatible Gaussians into persistent object units

$$\mathcal{U}_t = \{u_k\}_{k=1}^{M_t}, \qquad u_k = \left(\mathcal{I}_k,\, \bar{\mathbf{x}}_k,\, \mathbf{b}_k,\, \bar{\mathbf{p}}_k,\, \bar{\mathbf{f}}_k,\, w_k\right), \tag{4}$$

where $\mathcal{I}_k$ is the Gaussian index set, $\bar{\mathbf{x}}_k$ is the 3D center, $\mathbf{b}_k$ is the spatial extent, $\bar{\mathbf{p}}_k$ and $\bar{\mathbf{f}}_k$ are aggregated semantic attributes, and w_k is the support strength. These units serve as stable volumetric anchors for query matching, contextual reasoning, and verification.

Query-Conditioned Evidence Projection. The query branch produces $\Lambda^q = (\mathbf{z}^q, \mathbf{r}^q, \mathcal{C}^q)$, where $\mathbf{z}^q$ is a query embedding, $\mathbf{r}^q$ is a room-level prior, and $\mathcal{C}^q$ encodes contextual expectations. For an object unit u_k, direct query compatibility is $z_t^q(u_k) = \frac{1+\cos(\bar{\mathbf{f}}_k, \mathbf{z}^q)}{2}$. Rather than storing query-conditioned semantics as a dense persistent map, we project evidence to spatial regions on demand. For a candidate region r, its query-conditioned semantic support is

$$S_t(r; q) = \Phi_{\text{sem}}\left(r,\, \mathcal{G}_t,\, \mathcal{U}_t,\, \Lambda^q\right), \tag{5}$$

where Φ_{sem} aggregates Gaussian-level semantics, object-level compatibility, room priors, and contextual cues within r.

Executable Spatial Support. Navigation is executed on a BEV support state

$$\mathbf{B}_t = \left[\mathbf{B}_t^{\text{occ}},\, \mathbf{B}_t^{\text{exp}},\, \mathbf{B}_t^{\text{free}}\right], \tag{6}$$

where the channels encode occupied, explored, and traversable cells. We keep $\mathbf{B}_t$ query-agnostic and use it only for executable spatial support. Semantic and visual evidence remains attached to $\mathcal{G}_t$ and $\mathcal{U}_t$, and query-conditioned region scores are computed on demand through Λ^q rather than stored as a persistent BEV semantic channel.

3.2 Decoupled Complementary Fields

The state $\mathcal{M}_t^q$ provides heterogeneous evidence with different decision roles. Query-conditioned semantic evidence indicates where task-relevant information may exist, while spatial and geometric support indicate whether a region is reliable for motion and inspection. Collapsing these signals into a single score can overcommit to regions that are semantically attractive but structurally unreliable, or ignore regions that are ambiguous but informative. We therefore define Decoupled Complementary Fields over reachable regions.

Let $\mathcal{R}_t = \text{Partition}(\mathbf{B}_t^{\text{free}}, \mathbf{B}_t^{\text{exp}}, \mathbf{x}_t)$ be the set of connected reachable regions from the current ground-plane position $\mathbf{x}_t$. For each $r \in \mathcal{R}_t$, DCF is

$$\text{DCF}_t(r; q) = \left(U_t(r; q),\, A_t(r)\right), \tag{7}$$

where $U_t(r; q)$ is the query-conditioned *Query-Ambiguity Field*, and $A_t(r)$ is the query-agnostic *Environment-Readiness Field*.

Query-Ambiguity Field. The Query-Ambiguity Field measures where evidence relevant to q remains unresolved, not merely where the query score is high. For region r, let $\boldsymbol{\pi}_t(r)$ be a room posterior estimated from Gaussian and object evidence, $\Delta_t(r; q)$ be the disagreement among query-relevant cues, and $h(S_t(r; q))$ be the Bernoulli entropy of the support in Eq. (5). We define

$$U_t(r; q) = \text{Norm}_{\mathcal{R}_t}\left(\bar{H}_t(r) + \Delta_t(r; q) + h(S_t(r; q))\right), \tag{8}$$

where $\bar{H}_t(r)$ is the normalized entropy of $\boldsymbol{\pi}_t(r)$, and $\text{Norm}_{\mathcal{R}_t}$ normalizes scores over $\mathcal{R}_t$. Thus, $U_t(r; q)$ is high when a region's semantic role is uncertain, its evidence sources disagree, or its query support is suggestive but not decisive.

Environment-Readiness Field. The Environment-Readiness Field estimates whether a region is sufficiently established for reliable navigation and inspection, independent of the query. We combine map-level structural support from $\mathbf{B}_t$ with Gaussian-level geometric support from $\mathcal{G}_t$

$$A_t(r) = \mathrm{Norm}_{\mathcal{R}_t}\left(A_t^{\mathrm{map}}(r) + A_t^{\mathrm{gs}}(r)\right). \tag{9}$$

Here $A_t^{\mathrm{map}}(r)$ penalizes frontier exposure, nearby unknown space, and incomplete boundaries, while $A_t^{\mathrm{gs}}(r)$ rewards Gaussian density, coverage, and consistency with traversable free space. This field acts as a structural gate that prevents query-driven exploration from overcommitting to weakly observed or geometrically unstable regions.

3.3 Hierarchical Navigation via Complementary-Field Coupling

DCF is defined over regions, while embodied navigation is executed through short-horizon actions. We therefore use a hierarchical policy: the global layer selects an informative frontier via field coupling, and the local layer instantiates an executable waypoint on $\mathbf{B}_t^{\mathrm{free}}$.

Global Region-and-Frontier Reasoning. For each region r, we define a coupled region value

$$\mathcal{V}_t(r; q) = A_t(r)\frac{S_t(r; q) + U_t(r; q)}{2}. \tag{10}$$

The support term $S_t(r; q)$ favors promising evidence, the ambiguity term $U_t(r; q)$ favors informative disambiguation, and the readiness term $A_t(r)$ gates both by structural reliability. Let $\mathcal{F}_t$ be frontier candidates extracted from the boundary between explored free space and unknown space, and let $r(f)$ be the region associated with frontier f. We select frontiers by

$$f_t^\star = \arg\max_{f\in\mathcal{F}_t}\left[\mathcal{V}_t(r(f); q) + \nu_t(f) - C_t(f)\right], \tag{11}$$

where $\nu_t(f)$ and $C_t(f)$ denote local exploration novelty and geodesic motion cost, respectively.

Local Goal Instantiation for Executable Navigation. The selected frontier is converted into a local goal map $\mathbf{X}_t^{\mathrm{loc}} = \mathrm{GoalMap}(f_t^\star, \mathbf{B}_t^{\mathrm{free}})$, and a Fast Marching planner computes a short-horizon waypoint $\mathbf{x}_t^\star = \mathrm{FMM}(\mathbf{B}_t^{\mathrm{free}}, \mathbf{X}_t^{\mathrm{loc}})$ on traversable free space. After each action, the agent updates $\mathcal{M}_t^q$, recomputes DCF, and replans in closed loop.

3.4 Volumetric Gaussian Verification

Region-level fields guide where to acquire evidence, but final decisions should be grounded in persistent volumetric candidates rather than transient frame-level cues. We therefore verify task-relevant hypotheses using the Gaussian map $\mathcal{G}_t$ and object units $\mathcal{U}_t$.

Volumetric Candidate Grounding. At timestep t, we construct a temporary query-conditioned evidence map

$$\mathbf{E}_t^q = \Psi\left(\mathcal{O}_t,\, \mathbf{B}_t,\, \mathcal{G}_t,\, \mathcal{U}_t,\, \Lambda^q\right), \tag{12}$$

where Ψ aggregates current-view evidence, region-level support, object compatibility, and persistent Gaussian evidence. For each object unit $u \in \mathcal{U}_t$, its overlap with $\mathbf{E}_t^q$ is $O_t(u; q) = \mathrm{Overlap}(\Pi(u), \mathbf{E}_t^q)$, where $\Pi(u)$ is its BEV projection. We form a candidate set $\mathcal{K}_t = \mathrm{Cand}(\mathcal{U}_t, \mathbf{E}_t^q, q)$ from objects with strong evidence overlap or persistent query compatibility.

For each $u \in \mathcal{K}_t$, we aggregate spatial, semantic, query, geometric, and multi-view cues into a candidate-specific evidence score

$$\kappa_t(u; q) = \sum_{j\in\mathcal{J}_{\mathrm{ver}}} \omega_j\, \kappa_t^j(u; q), \qquad \sum_{j\in\mathcal{J}_{\mathrm{ver}}} \omega_j = 1, \tag{13}$$

where $\mathcal{J}_{\mathrm{ver}} = \{O, S, Q, G, M\}$ indexes evidence overlap, persistent semantic support, object query compatibility, Gaussian support mass, and multi-view consistency, respectively. All cues lie in $[0, 1]$, and $\omega_j \geq 0$ for every $j \in \mathcal{J}_{\mathrm{ver}}$. We rank candidates using

$$\widetilde{\kappa}_t(u; q) = \mathrm{Norm}_{\mathcal{K}_t}\big(\kappa_t(u; q)\big). \tag{14}$$

The normalized score is used only for ranking, while $\kappa_t(u; q)$ is retained for temporal verification.

Progressive Verification for Reliable Decisions. When $\mathcal{K}_t \neq \varnothing$, let $u_t^\star = \arg\max_{u \in \mathcal{K}_t} \widetilde{\kappa}_t(u; q)$ be the best current candidate. If $\mathcal{K}_t = \varnothing$, no candidate is selected and the agent continues frontier exploration. We accumulate candidate-specific evidence over time as

$$b_t(u; q) = \lambda_b b_{t-1}(u; q) + (1 - \lambda_b)\, \mathrm{vis}_t(u) \kappa_t(u; q), \tag{15}$$

where $\mathrm{vis}_t(u) \in \{0, 1\}$ indicates whether u is validly observed at timestep t. A candidate is verified when its accumulated confidence and temporal consistency satisfy the verification criteria defined in Appendix A.4. If no candidate is reliable, the agent continues frontier exploration; a plausible but unverified candidate induces an inspection goal, while a verified candidate provides grounded evidence for stopping or answer handoff.

4 Experiments

We evaluate DCF-Nav on three complementary settings: Active Embodied Question Answering (A-EQA) and Episodic-Memory Embodied Question Answering (EM-EQA) from OpenEQA [8], and multimodal lifelong navigation on GOAT-Bench [6]. Together, they evaluate active evidence acquisition, memory-based reasoning, persistent evidence reuse, and grounded final decisions. Detailed protocols and baselines are provided in Appendix B, with repeated-run uncertainty and computational analyses in Appendices D and H.

4.1 Active Embodied Question Answering

Protocol. A-EQA is the active exploration setting of OpenEQA [8], where an agent gathers visual evidence in an unseen environment before answering an open-vocabulary question. Following prior work [9], we evaluate on the commonly used 184-question subset. We report LLM-Match, which evaluates answer correctness against human references using GPT-4, and LLM-Match SPL, which additionally accounts for exploration efficiency. Detailed benchmark and evaluation protocols are provided in Appendix B.1.

Baselines. We compare against language-only, caption-based, direct-VLM, and active exploration methods, including Explore-EQA [29], ConceptGraphs [16], 3D-Mem [9], GSMem [26], and HGR [30]. All VLM-exploration methods use GPT-4o for consistency with prior work.

Table 1: Experiments on A-EQA. "CG" denotes ConceptGraphs. All VLM exploration methods use GPT-4o as the VLM backbone.

Method	LLM-Match ↑	LLM-Match SPL ↑
Blind LLMs		
GPT-4o	35.9	–
LLaMA-3.1-8B-Instruct	32.2	–
LLM with Frame Captions		
GPT-4o w/ LLaVA-1.5-13B	33.1	–
LLaMA-3.1 w/ LLaVA-1.5-13B	30.8	–
VLMs		
GPT-4o	47.6	–
Qwen3-VL-8B-Instruct	44.0	–
VLM Exploration		
Explore-EQA [29]	46.9	23.4
CG [16] w/ Frontier Snapshots	47.2	33.3
3D-Mem [9]	52.6	42.0
GSMem [26]	55.4	43.8
HGR [30]	55.9	45.0
DCF-Nav (Ours)	**56.5**	**45.2**

Results. Table 1 shows that DCF-Nav achieves the strongest overall performance on A-EQA. The gap between passive and exploration-based baselines highlights the importance of active evidence acquisition for open-ended embodied question answering. Compared with snapshot- and graph-based memories, DCF-Nav retains query-relevant observations as persistent volumetric evidence rather than isolated views or discrete object entries. The gains over 3D-Mem and GSMem further suggest that persistent memory alone is insufficient, as accumulated evidence must also be converted into query-conditioned navigation decisions during exploration. Compared with HGR, DCF-Nav grounds candidate evaluation in persistent volumetric evidence while explicitly separating unresolved query evidence from environment-side reliability. These comparisons suggest that the advantage of DCF-Nav lies in making persistent Gaussian evidence actionable through decoupled Query-Ambiguity and Environment-Readiness reasoning, while retaining candidate-level volumetric verification for reliable final decisions.

4.2 Episodic-Memory Embodied Question Answering

Protocol. EM-EQA evaluates question answering from previously collected episodic observations without active exploration. To isolate memory and verification from the navigation policy, we disable active navigation, frontier selection, Query-Ambiguity, and Environment-Readiness, while retaining persistent Gaussian memory and volumetric verification. We report LLM-Match and the average number of visual frames used for answering. Further details are provided in Appendix B.2.

Baselines. We compare against language-only, caption-based, and multi-frame reasoning methods, including 3D-Mem [9].

Results. Table 2 shows that DCF-Nav outperforms 3D-Mem with comparable visual evidence. With active navigation and complementary-field reasoning disabled, the comparison highlights the effectiveness of persistent semantic Gaussians and object-centric volumetric units for organizing previously observed evidence across viewpoints. Volumetric verification further grounds answer generation in this persistent representation rather than isolated episodic frames.

Table 2: Experiments on EM-EQA.

Method	Avg. Frames	LLM-Match
Blind LLM	0.0	35.5
CG Captions	0.0	34.4
SVM Captions	0.0	34.2
Frame Captions	0.0	38.1
Multi-Frame	3.0	48.1
3D-Mem [9]	3.1	57.2
DCF-Nav (Ours)	3.1	**58.5**
Human	Full	86.8

4.3 Multimodal Lifelong Navigation

Protocol. GOAT-Bench [6] evaluates sequential navigation to category-, language-, and image-specified targets while preserving memory across subtasks. We evaluate the full "Val Unseen" split, comprising 36 scenes and 360 episodes with 2,669 navigation subtasks, as well as the 278-subtask comparison subset containing one episode per scene, following prior work [9, 30]. We report Success Rate (SR) and Success weighted by Path Length (SPL), where successful navigation requires stopping within 1 meter of the target and SPL additionally accounts for path efficiency. Detailed evaluation settings and subset composition are provided in Appendix B.3.

Baselines. We compare against the original GOAT-Bench baselines and recent VLM- and memory-based navigation methods, including VLMnav [49], DyNaVLM [50], TANGO [51], MTU3D [35], ConceptGraphs [16], 3D-Mem [9], GSMem [26], and HGR [30].

Table 3: Experiments on the GOAT-Bench.

	Full Validation Set		Subset	
Method	SR ↑	SPL ↑	SR ↑	SPL ↑
Modular GOAT [4]	24.9	17.2	-	-
Modular CLIP on Wheels [48]	16.1	10.4	-	-
SenseAct-NN Monolithic [6]	12.3	6.8	-	-
SenseAct-NN Skill Chain [6]	29.5	11.3	-	-
VLMnav [49]	20.1	9.6	-	-
DyNaVLM [50]	25.5	10.2	-	-
TANGO [51]	32.1	16.5	-	-
MTU3D [35]	47.2	27.7	-	-
ConceptGraphs [16]	61.2	44.3	67.8	48.1
3D-Mem w/o memory [9]	58.6	38.5	66.2	44.1
3D-Mem [9]	62.9	44.7	69.1	48.9
GSMem [26]	67.2	46.9	-	-
HGR [30]	64.1	50.1	72.4	56.2
DCF-Nav (Ours)	**68.2**	**50.7**	**75.9**	**56.4**

Results. Table 3 shows that DCF-Nav achieves the strongest performance on both the full "Val Unseen" split and the comparison subset. The improvements in both success and path efficiency indicate that the gain is associated with more effective target-directed navigation rather than simply longer exploration. Compared with ConceptGraphs and 3D-Mem, DCF-Nav preserves reusable multi-view geometric, semantic, and appearance evidence in an object-centric Gaussian state instead of relying on discrete object nodes or sparse visual snapshots. Compared with GSMem, the results suggest that persistent Gaussian memory is further strengthened when its accumulated evidence is explicitly converted into navigation-time region decisions. Compared with HGR, DCF-Nav performs this reasoning through decoupled Query-Ambiguity and Environment-Readiness fields, allowing each new goal to reinterpret the persistent scene state without rebuilding the underlying spatial memory. Together, these comparisons highlight the benefit of combining persistent volumetric evidence with query-conditioned, decoupled navigation reasoning for long-horizon embodied navigation.

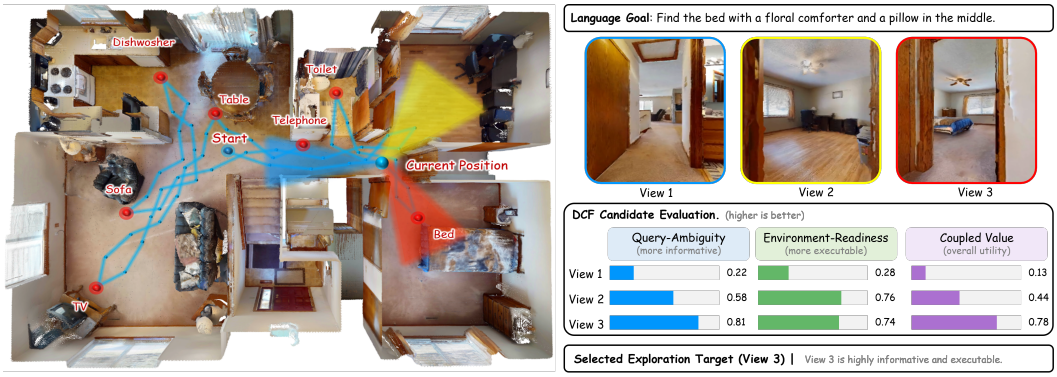


Figure 3: Qualitative example of DCF-guided region selection. Given the persistent Gaussian navigation state and the current query, DCF-Nav evaluates candidate views using Query-Ambiguity and Environment-Readiness. Their coupled values favor regions that are both informative for resolving the query and reliable for navigation and inspection, leading the agent to select View 3 for subsequent exploration.

4.4 Ablation and Mechanism Analysis

Component Ablations. Table 4 evaluates the main components of DCF-Nav on A-EQA and the GOAT-Bench comparison subset. Removing the object-centric units consistently degrades performance, supporting their role as a compact interface between fine-grained Gaussian primitives and region-level reasoning. The two field ablations reveal distinct decision roles. Without Query-Ambiguity, the policy loses an explicit signal for regions where task-relevant evidence remains unresolved, whereas removing Environment-Readiness more strongly affects navigation efficiency by weakening the distinction between semantically promising and structurally reliable regions. More importantly, the Single Coupled Score performs worse than either single-field ablation, indicating that the benefit of DCF arises from preserving query-related ambiguity and environment-side readiness as separate decision variables rather than aggregating all cues into a single exploration score. Removing volumetric verification further degrades both tasks, showing that candidate selection benefits from persistent evidence-based verification before stopping or answer handoff. Five-run uncertainty analysis in Appendix D further confirms the stability of these component-level trends.

Table 4: Ablation study on A-EQA and the GOAT-Bench subset, evaluating object-centric abstraction, complementary-field decoupling, and volumetric verification.

	A-EQA		GOAT-Bench	
Variant	LLM-Match ↑	LLM-Match SPL ↑	SR ↑	SPL ↑
Full DCF-Nav	**56.5**	**45.2**	**75.9**	**56.4**
w/o Object-Centric Units	53.6	41.9	70.5	50.8
w/o Query-Ambiguity Field	52.9	41.5	69.8	50.6
w/o Environment-Readiness Field	54.1	40.3	70.9	48.7
Single Coupled Score	51.8	39.4	68.3	47.9
w/o Volumetric Verification	49.7	38.6	68.7	49.6

Validity of the Complementary Fields. The component ablations establish the utility of the two fields, but do not directly verify whether they exhibit their intended decision roles. We therefore evaluate Query-Ambiguity on 914 pre-discovery candidates using subsequent task-relevant evidence as an offline diagnostic. Query-Ambiguity achieves an AUC of 0.621 for subsequent target visibility, while target visibility increases from 11.65% in the lowest ambiguity quartile to 18.80% in the highest. Environment-Readiness exhibits a complementary behavioral effect, with the proposed field coupling reducing dead-end and backtracking rates to 10.32% and 21.00%, respectively. These results provide predictive and behavioral evidence that Query-Ambiguity identifies regions where additional task-relevant evidence is useful, whereas Environment-Readiness suppresses structurally unproductive exploration. Detailed constituent-field diagnostics and alternative coupling formulations are provided in Appendix E.

Volumetric Verification. We further examine whether volumetric verification actively revises candidate decisions rather than merely accumulating confidence. Across the diagnostic episodes, the verifier revises 37 of 205 initial VLM-selected candidates, showing that persistent semantic, geometric, and multi-view evidence can revise preliminary candidate selections before stopping or answer handoff. Representative sequences of candidate refinement and recovery after rejection

are presented in Appendix C.4. A controlled comparison with sampled render-to-VLM evidence is provided in Appendix F, and robustness to pose and depth perturbations is reported in Appendix G.

Qualitative Analysis. Figure 3 illustrates how DCF-Nav translates persistent Gaussian evidence into navigation-time region decisions. Query-Ambiguity highlights regions where task-relevant evidence remains unresolved, while Environment-Readiness evaluates whether the corresponding spatial evidence is sufficiently established for reliable motion and inspection. Their coupling discourages selection based solely on semantic relevance and instead favors regions with both task-relevant evidence and structural support. This example demonstrates how the complementary fields connect persistent volumetric evidence with executable exploration decisions. Additional full-episode trajectories are provided in Appendix C.2, with analyses of challenging and successful cases in Appendices C.3 and C.4.

5 Conclusion

We presented *Decoupled Complementary Fields on 3D Gaussian Maps* for embodied navigation and reasoning in previously unseen environments. DCF-Nav maintains persistent fine-grained 3D evidence in a query-conditioned object-centric Gaussian state, separates query-related ambiguity from environment readiness for hierarchical navigation, and grounds final decisions through volumetric verification. Experiments across three complementary benchmarks demonstrate its effectiveness for active embodied question answering, episodic-memory reasoning, and multimodal lifelong navigation, while ablation and mechanism analyses validate the distinct roles of the proposed components.

Limitations. The current framework primarily targets static or quasi-static indoor environments and depends on sufficiently accurate RGB-D observations and camera poses for persistent Gaussian fusion. Severe geometric corruption or persistent scene changes may introduce stale or inconsistent evidence, while end-to-end latency remains dominated by VLM-mediated reasoning. Future work will explore dynamic scene modeling, more efficient reasoning, and deployment under imperfect real-world sensing and actuation.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grants 62293510, 62293515, 62573184, 62373293, 62463020, 62472336, 62302369, and U2570205, in part by the Science and Technology Project of Jiangxi Province under Grants 20261BCE310009, 20261BCE310010, and 20252BAC200185, in part by the Young Science and Technology Star Program of Shaanxi Province under Grant 2025ZC-KJXX-25, in part by the Ganpo Talent Support Plan under Grant gpjc20260121, in part by the Key Program of the National Natural Science Foundation of China under Grant 62433010, and in part by the Major Program of the National Natural Science Foundation of China under Grant 62595801.

The authors declare no competing interests.

References

[1] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018.

[2] Jiahao Zhao, Yaonan Wang, Zijie Wu, Mingtao Feng, Renjie Ding, and Hui Zhang. Staged modulation diffusion policy with complementary visual fusion for robotic workpiece assembly. *IEEE Transactions on Industrial Informatics*, 22(8):7506–7517, 2026.

[3] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.

[4] Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*, 2023.

[5] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634, 2024.

[6] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024.

[7] Bo Miao, Weijia Liu, Jun Luo, Lachlan Shinnick, Jian Liu, Thomas Hamilton-Smith, Yuhe Yang, Zijie Wu, Vanja Videnovic, Feras Dayoub, and Anton van den Hengel. LangMap: A human-verified benchmark for hierarchical open-vocabulary goal navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

[8] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024.

[9] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17294–17303, 2025.

[10] Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19057–19066, 2025.

[11] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR, 2023.

[12] Yuxuan Kuang, Hai Lin, and Meng Jiang. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 338–351, 2024.

[13] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022.

[14] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021.

[15] Mingtao Feng, Haoran Hou, Liang Zhang, Zijie Wu, Yulan Guo, and Ajmal Mian. 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9182–9191, 2023.

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

[17] Bo Miao, Mingtao Feng, Zijie Wu, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Referring human pose and mask estimation in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[18] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[19] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 15625–15636, 2023.

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[21] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21357–21366, 2024.

[22] Pengchong Hu and Zhizhong Han. Vtgaussian-slam: Rgbd slam for large scale scenes with splatting view-tied 3d gaussians. In *International Conference on Machine Learning*, pages 24347–24370. PMLR, 2025.

[23] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024.

[24] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

[25] Jianzhe Gao, Rui Liu, and Wenguan Wang. 3d gaussian map with open-set semantic grouping for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9252–9262, 2025.

[26] Yiren Lu, Yi Du, Disheng Liu, Yunlai Zhou, Chen Wang, and Yu Yin. Gsmem: 3d gaussian splatting as persistent spatial memory for zero-shot embodied exploration and reasoning. *arXiv preprint arXiv:2603.19137*, 2026.

[27] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352, 2022.

[28] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48. IEEE, 2024.

[29] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024.

[30] Peixin Chen, Guoxi Zhang, Jianwei Ma, and Qing Li. Hypothesis graph refinement: Hypothesis-driven exploration with cascade error correction for embodied navigation. *arXiv preprint arXiv:2604.04108*, 2026.

[31] Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[32] Jianzhe Gao, Rui Liu, Yuxuan Xu, Tongtong Cao, Yingxue Zhang, Zhanguang Zhang, Sida Peng, Yi Yang, and Wenguan Wang. Uncertainty-aware gaussian map for vision-language navigation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[33] Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*, 2022.

[34] Xinyu Sun, Lizhao Liu, Hongyan Zhi, Ronghe Qiu, and Junwei Liang. Prioritized semantic learning for zero-shot instance navigation. In *European Conference on Computer Vision*, pages 161–178. Springer, 2024.

[35] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8120–8132, 2025.

[36] Wenyuan Xie, Shaokai Wu, Yijin Zhou, Yanbiao Ji, Guodong Zhang, Bayram Bayramli, Qiuchang Li, Xunchu Zhou, Yue Ding, and Hongtao Lu. MVP-Nav: Multi-layer Value Map Planner Navigator. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026.

[37] Taeyun Kim, Alvin Jinsung Choi, Dasol Hong, and Hyun Myung. Clue: Adaptively prioritized contextual cues by leveraging a unified semantic map for effective zero-shot object-goal navigation. *arXiv preprint arXiv:2605.19206*, 2026.

[38] Guanlin Liu, Shaobin Ling, Renyuan Liu, Zeying Gong, and Junjie Hu. Aecnav: Active evidence consolidation for efficient zero-shot open-vocabulary object navigation. *arXiv preprint arXiv:2608.10817*, 2026.

[39] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Multimodal spatial language maps for robot navigation and manipulation. *The International Journal of Robotics Research*, page 02783649251351658, 2025.

[40] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10968–10980, 2023.

[41] Zijie Wu, Mingtao Feng, Yaonan Wang, He Xie, Weisheng Dong, Bo Miao, and Ajmal Mian. External knowledge enhanced 3d scene generation from sketch. In *European Conference on Computer Vision*, pages 286–304. Springer, 2024.

[42] Mingtao Feng, Chenbo Yan, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Hyperrectangle embedding for debiased 3d scene graph prediction from rgb sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6410–6426, 2025.

[43] Qitong Yang, Mingtao Feng, Zijie Wu, Shijie Sun, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Learning coherent matrixized representation in latent space for volumetric 4d generation. *Pattern Recognition*, page 113531, 2026.

[44] Zijie Wu and Na Zhao. Hierascaffold: Learning compact hierarchical representations for scalable 4d liDAR generation. In *Forty-third International Conference on Machine Learning*, 2026.

[45] Renjie Ding, Yaonan Wang, Min Liu, Jialin Zhu, Jiazheng Wang, Jiahao Zhao, Wenting Shen, Feixiang He, and Xiang Chen. A step to decouple optimization in 3dgs. In *The Fourteenth International Conference on Learning Representations*, 2026.

[46] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8895–8905. IEEE, 2023.

[47] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024.

[48] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023.

[49] Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *arXiv preprint arXiv:2411.05755*, 2024.

[50] Zihe Ji, Huangxuan Lin, and Yue Gao. Dynavlm: Zero-shot vision-language navigation system with dynamic viewpoints and self-refining graph memory. *arXiv preprint arXiv:2506.15096*, 2025.

[51] Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. Tango: training-free embodied ai agents for open-world tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24603–24613, 2025.

[52] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

A Implementation Details of Method

A.1 Query-Conditioned Gaussian State

This section provides implementation details for the query-conditioned object-centric Gaussian navigation state in Sec. 3.1. We describe how frame-level semantics are lifted to persistent Gaussian primitives, how object units are aggregated and associated over time, how region-level query support is computed, and how executable BEV support is maintained.

Semantic lifting. For each visible Gaussian g_i, rasterization provides the contributing pixel set $\mathcal{P}_t(i)$ and rendering weights $\omega_{i,u,t}$. Given pixel-wise category posterior $\mathbf{s}_t(u)$, open-vocabulary feature $\mathbf{e}_t(u)$, and confidence $c_t(u)$, we compute the visibility mass $m_{i,t} = \sum_{u\in\mathcal{P}_t(i)} \omega_{i,u,t} c_t(u)$ and frame-level semantic evidence

$$\widetilde{\mathbf{p}}_{i,t} = \frac{\sum_{u\in\mathcal{P}_t(i)} \omega_{i,u,t} c_t(u)\mathbf{s}_t(u)}{m_{i,t}+\epsilon}, \qquad \widetilde{\mathbf{f}}_{i,t} = \mathrm{Normalize}_2\left(\frac{\sum_{u\in\mathcal{P}_t(i)} \omega_{i,u,t} c_t(u)\mathbf{e}_t(u)}{m_{i,t}+\epsilon}\right). \tag{16}$$

The persistent semantic state is updated by reliability-weighted fusion:

$$\mathbf{p}_{i,t} = \mathrm{Normalize}_1\left(\lambda_\rho \rho_{i,t-1}\mathbf{p}_{i,t-1} + m_{i,t}\widetilde{\mathbf{p}}_{i,t}\right), \mathbf{f}_{i,t} = \mathrm{Normalize}_2\left(\lambda_\rho \rho_{i,t-1}\mathbf{f}_{i,t-1} + m_{i,t}\widetilde{\mathbf{f}}_{i,t}\right), \tag{17}$$

with $\rho_{i,t} = \lambda_\rho \rho_{i,t-1} + m_{i,t}$. The frame-level evidence in Eq. (16) is normalized by the visibility mass. The fusion in Eq. (17) retains visibility-mass weighting and gives greater influence to repeated, reliable observations.

Object-unit aggregation and association. Reliable Gaussians are grouped by spatial proximity and semantic compatibility. For an object unit u_k with index set $\mathcal{I}_k$, let $W_k = \sum_{i\in\mathcal{I}_k} \rho_i\alpha_i$. The aggregated center and semantics are

$$\bar{\mathbf{x}}_k = \frac{1}{W_k+\epsilon}\sum_{i\in\mathcal{I}_k} \rho_i\alpha_i\boldsymbol{\mu}_i, \bar{\mathbf{p}}_k = \mathrm{Normalize}_1\left(\sum_{i\in\mathcal{I}_k} \rho_i\alpha_i\mathbf{p}_i\right), \bar{\mathbf{f}}_k = \mathrm{Normalize}_2\left(\sum_{i\in\mathcal{I}_k} \rho_i\alpha_i\mathbf{f}_i\right). \tag{18}$$

The support strength is $w_k = W_k$, and the extent $\mathbf{b}_k$ is computed from the spatial support of member Gaussians. To maintain temporal persistence, a newly induced local unit u' is associated with an existing unit u by combining 3D overlap and semantic consistency:

$$\mathcal{A}(u', u) = \lambda_{\mathrm{geo}}\,\mathrm{IoU}_{3D}(\mathbf{b}', \mathbf{b}) + \lambda_{\mathrm{feat}}\cos(\bar{\mathbf{f}}_{u'}, \bar{\mathbf{f}}_u) + \lambda_{\mathrm{sem}}\left(1 - D_{\mathrm{JS}}(\bar{\mathbf{p}}_{u'}, \bar{\mathbf{p}}_u)\right). \tag{19}$$

The local unit is merged into the best-matching persistent unit when the association is reliable; otherwise, it initializes a new object unit.

Region-level semantic support. For a region r, define associated Gaussian and object sets $\mathcal{G}_t(r) = \{g_i \in \mathcal{G}_t \mid \Pi(\boldsymbol{\mu}_i) \in r\}$ and $\mathcal{U}_t(r) = \{u_k \in \mathcal{U}_t \mid \Pi(u_k) \cap r \neq \varnothing\}$, where $\Pi(\cdot)$ projects 3D support to the BEV interface. We compute Gaussian-level query support as

$$S_G(r; q) = \frac{\sum_{g_i\in\mathcal{G}_t(r)} \rho_i\alpha_i s_i^q}{\sum_{g_i\in\mathcal{G}_t(r)} \rho_i\alpha_i + \epsilon}, \tag{20}$$

where $s_i^q = \frac{1+\cos(\mathbf{f}_i, \mathbf{z}^q)}{2}$ for open-vocabulary queries and $s_i^q = \mathbf{p}_i(c(q))$ when a category label $c(q)$ is specified. Object-level support and contextual support are

$$S_U(r; q) = \frac{\sum_{u_k\in\mathcal{U}_t(r)} w_k z_t^q(u_k)}{\sum_{u_k\in\mathcal{U}_t(r)} w_k + \epsilon}, \qquad S_C(r; q) = \frac{\sum_{u_k\in\mathcal{U}_t(r)} w_k c_t^{\mathrm{ctx}}(u_k; q)}{\sum_{u_k\in\mathcal{U}_t(r)} w_k + \epsilon}. \tag{21}$$

Here $z_t^q(u_k) = \frac{1+\cos(\bar{\mathbf{f}}_k, \mathbf{z}^q)}{2}$, and $c_t^{\mathrm{ctx}}(u_k; q)$ is obtained from query-conditioned contextual expectations over nearby object units. With room compatibility $S_R(r; q) = \langle \boldsymbol{\pi}_t(r), \mathbf{r}^q \rangle$, we instantiate Eq. (5) as

$$S_t(r; q) = \mathrm{Norm}_{\mathcal{R}_t}\left(S_G(r; q) + S_U(r; q) + S_C(r; q) + S_R(r; q)\right), \tag{22}$$

where unavailable cues are omitted.

Spatial support fusion. For each BEV channel $\zeta \in \{\mathrm{occ}, \mathrm{exp}, \mathrm{free}\}$, the local estimate $\widetilde{\mathbf{B}}_t^\zeta$ is pose-aligned and fused by

$$\mathbf{B}_t^\zeta = \mathrm{Fuse}^{\mathrm{map}}\left(\mathbf{B}_{t-1}^\zeta, \mathcal{W}_{\mathbf{T}_t}\left(\widetilde{\mathbf{B}}_t^\zeta\right)\right). \tag{23}$$

The BEV support remains query-agnostic; query-conditioned semantics are projected from $\mathcal{G}_t$ and $\mathcal{U}_t$ only during region evaluation.

A.2 Decoupled Complementary Fields

For a scalar score s on a nonempty finite index set $\mathcal{D}$, we define

$$\mathrm{Norm}_{\mathcal{D}}(s(d)) = \frac{s(d) - \min_{d' \in \mathcal{D}} s(d')}{\max_{d' \in \mathcal{D}} s(d') - \min_{d' \in \mathcal{D}} s(d') + \epsilon}, \qquad \epsilon > 0. \tag{24}$$

The same definition is used for $\mathcal{D} = \mathcal{R}_t$, $\mathcal{F}_t$, and $\mathcal{K}_t$. When all scores are equal, the normalized scores are zero.

Reachable region construction. Let Ω_t^{free} and Ω_t^{exp} be the traversable and explored cells from $\mathbf{B}_t$. We first restrict planning to the reachable component from the current ground-plane position $\mathbf{x}_t$:

$$\Omega_t^{\mathrm{reach}} = \mathrm{Reach}\left(\Omega_t^{\mathrm{free}} \cap \Omega_t^{\mathrm{exp}}, \mathbf{x}_t\right). \tag{25}$$

The reachable component $\Omega_t^{\mathrm{reach}}$ is further partitioned into connected, nonempty subregions, $\mathcal{R}_t = \mathrm{Partition}(\Omega_t^{\mathrm{reach}})$. This operation subdivides the reachable component rather than merely extracting its connected components.

Room posterior and cue disagreement. The room posterior $\boldsymbol{\pi}_t(r)$ is estimated from category evidence in $\mathcal{G}_t(r)$ and $\mathcal{U}_t(r)$ using a category-to-room compatibility table, followed by softmax normalization. Let $\mathcal{L}$ be the set of room categories, with $|\mathcal{L}| \geq 2$, and let $\pi_t(r, \ell)$ denote the ℓ-th component of $\boldsymbol{\pi}_t(r)$. Its normalized entropy is

$$\bar{H}_t(r) = -\frac{1}{\log |\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \pi_t(r, \ell) \log \pi_t(r, \ell). \tag{26}$$

Let $\mathcal{J}_t(r; q)$ index the available normalized cues among Gaussian support, object support, contextual support, and room compatibility, and let $s_j(r; q)$ denote cue j. Their disagreement is the cue variance

$$\Delta_t(r; q) = \frac{1}{|\mathcal{J}_t(r; q)|} \sum_{j \in \mathcal{J}_t(r;q)} \left(s_j(r; q) - \bar{s}(r; q)\right)^2, \tag{27}$$

where, for $|\mathcal{J}_t(r; q)| \geq 1$, $\bar{s}(r; q) = |\mathcal{J}_t(r; q)|^{-1} \sum_{j \in \mathcal{J}_t(r;q)} s_j(r; q)$. This term is high when, for example, Gaussian features suggest query relevance but object-level or room-level evidence does not.

Readiness terms. Map-level readiness is computed from frontier exposure $F_t(r)$, nearby unknown-space ratio $\Xi_t(r)$, and boundary incompleteness $\Gamma_t(r)$:

$$A_t^{\mathrm{map}}(r) = 1 - \mathrm{Norm}_{\mathcal{R}_t}\left(F_t(r) + \Xi_t(r) + \Gamma_t(r)\right). \tag{28}$$

Here $F_t(r)$ measures the fraction of the region boundary adjacent to frontier cells, $\Xi_t(r)$ measures the amount of unknown space near r, and $\Gamma_t(r)$ measures unresolved boundary support in the BEV map.

Gaussian-level readiness summarizes volumetric stability:

$$A_t^{\mathrm{gs}}(r) = \mathrm{Norm}_{\mathcal{R}_t}\left(D_t^{\mathrm{gs}}(r) + C_t^{\mathrm{gs}}(r) + Q_t^{\mathrm{gs}}(r) - R_t^{\mathrm{gs}}(r)\right), \tag{29}$$

where D_t^{gs}, C_t^{gs}, Q_t^{gs}, and R_t^{gs} denote Gaussian density, Gaussian coverage, free-space consistency, and geometric risk, respectively. In practice, density and coverage reward stable volumetric support, free-space consistency checks whether projected Gaussians agree with $\mathbf{B}_t^{\mathrm{free}}$, and geometric risk penalizes obstacle-dominated or repeatedly visited regions.

A.3 Hierarchical Navigation

Frontier extraction and scoring. Frontier cells are reachable free cells adjacent to unknown space, and frontier candidates $\mathcal{F}_t$ are their connected components. Each frontier f is associated with the region $r(f)$ that shares the largest boundary contact with it. The novelty and cost terms in Eq. (11) are

$$\nu_t(f) = \mathrm{Norm}_{\mathcal{F}_t}\left(\frac{|\mathcal{N}(f) \cap \mathcal{X}_t^{\mathrm{unk}}|}{|\mathcal{N}(f)| + \epsilon}\right), \qquad C_t(f) = \mathrm{Norm}_{\mathcal{F}_t}\left(\mathrm{GeoDist}\left(\mathbf{B}_t^{\mathrm{free}}, \mathbf{x}_t, \mathbf{x}(f)\right)\right), \tag{30}$$

where $\mathcal{N}(f)$ is a local neighborhood around f, $\mathcal{X}_t^{\mathrm{unk}}$ is the unknown-cell set, and $\mathbf{x}(f)$ is the frontier centroid. The novelty term favors frontiers that expose new space, while the cost term discourages unnecessarily long detours.

Local goal and closed-loop execution. The selected frontier $f_t^\star$ induces a local goal set around $\mathbf{x}(f_t^\star)$ within reachable free space. FMM computes the geodesic field from this local goal set over $\mathbf{B}_t^{\mathrm{free}}$, and the short-horizon waypoint is obtained by descending the field from $\mathbf{x}_t$. The low-level action is selected according to the relative bearing from the current agent pose to the waypoint. After execution, the agent receives a new observation, updates $\mathcal{M}_t^q$, recomputes DCF, and replans. This closed-loop procedure allows the frontier choice to change as new Gaussian evidence and object units are incorporated.

A.4 Volumetric Gaussian Verification

Evidence map construction. The temporary evidence map in Eq. (12) combines current-view query evidence, region-level support, Gaussian support, and object compatibility. A compact instantiation is

$$\mathbf{E}_t^q(x) = \mathrm{Norm}\left(\mathbf{E}_t^{\mathrm{img}}(x;q) + S_t(r(x);q) + \mathbf{E}_t^G(x;q) + \mathbf{E}_t^U(x;q)\right), \tag{31}$$

where $r(x)$ is the region containing BEV cell x, $\mathbf{E}_t^{\mathrm{img}}$ is current-view query evidence projected to the BEV plane, $\mathbf{E}_t^G$ is Gaussian-level query evidence, and $\mathbf{E}_t^U$ is object-level query evidence.

Candidate grounding and scoring. For candidate u, we instantiate overlap as

$$O_t(u;q) = \frac{\sum_{x\in\Pi(u)} \mathbf{E}_t^q(x)}{|\Pi(u)| + \epsilon}. \tag{32}$$

The candidate set $\mathcal{K}_t$ contains object units with strong evidence overlap or strong persistent query compatibility. The verification cues in Eq. (13) are: evidence overlap $\kappa_t^O = O_t$, persistent semantic support κ_t^S from member Gaussians, object query compatibility $\kappa_t^Q = z_t^q(u)$, Gaussian support mass κ_t^G, and multi-view consistency κ_t^M over recent visible observations. For u with Gaussian index set $\mathcal{I}(u)$, the semantic cue can be written as

$$\kappa_t^S(u;q) = \frac{\sum_{i\in\mathcal{I}(u)} \rho_i \alpha_i s_i^q}{\sum_{i\in\mathcal{I}(u)} \rho_i \alpha_i + \epsilon}, \tag{33}$$

where s_i^q follows Eq. (20). The multi-view cue aggregates candidate evidence over recent valid observations, reducing the influence of single-view false positives.

Temporal consistency and decision interface. Let $\mathcal{W}_t$ denote a recent observation window and $\mathrm{vis}_\tau(u) \in \{0,1\}$ indicate whether candidate u is validly observed at timestep τ. Temporal consistency is defined as

$$\chi_t(u;q) = \mathbf{1}\left[\frac{\sum_{\tau\in\mathcal{W}_t} \mathrm{vis}_\tau(u)\mathbf{1}\left[\kappa_\tau(u;q) \geq \tau_{\mathrm{obs}}\right]}{\sum_{\tau\in\mathcal{W}_t} \mathrm{vis}_\tau(u) + \epsilon} \geq \tau_\chi\right], \tag{34}$$

where τ_{obs} is the per-observation evidence threshold and τ_χ is a threshold on the regularized support ratio above. The ratio is zero when there are no valid observations; for $\tau_\chi > 0$, such a window cannot pass verification. The temporal-consistency test uses the candidate-wise evidence score $\kappa_t(u;q)$ rather than the ranking score $\widetilde{\kappa}_t(u;q)$. The final verification condition is

$$\mathrm{Verified}_t(u;q) = \mathbf{1}\left[b_t(u;q) \geq \tau_b\right] \chi_t(u;q), \tag{35}$$

where τ_b is the accumulated-confidence threshold. For navigation tasks, termination additionally requires the benchmark-specific spatial stopping condition; for embodied question answering, verification triggers grounded evidence handoff for answer generation.

B Detailed Evaluation Protocols and Baselines

This section provides additional benchmark statistics, evaluation protocols, metric definitions, and baseline details omitted from the main paper for brevity.

B.1 Active Embodied Question Answering

Benchmark Protocol. A-EQA is the active exploration setting of OpenEQA [8]. The full benchmark contains 557 questions drawn from 63 HM3D scenes [52], covering diverse reasoning skills including object recognition, object localization, spatial understanding, functional reasoning, and attribute recognition. For each episode, the agent is initialized at a specified location in a previously unexplored environment and must actively gather sufficient visual evidence before producing a natural-language answer. Following prior embodied exploration work [9, 26], we evaluate on the commonly used 184-question A-EQA subset.

Metrics. We report LLM-Match and LLM-Match SPL following OpenEQA [8] and prior work [9, 26]. LLM-Match evaluates answer correctness by using GPT-4 to compare the predicted answer with the human-annotated reference answer on a 1–5 scale, which is subsequently rescaled to 0–100. LLM-Match SPL additionally incorporates exploration efficiency by weighting answer correctness according to path efficiency. For episodes in which an exploration method fails to produce a grounded answer, the path-efficiency term is set to zero following the adopted evaluation protocol.

Baseline Details. The A-EQA comparison contains four groups of methods. *Blind LLM* baselines answer questions without visual observations and therefore quantify the contribution of language priors alone. *LLM with Frame Captions* methods convert observed images into textual captions before answer generation. *VLM* baselines directly process visual observations without explicit active exploration. Finally, *VLM Exploration* methods actively move through the environment before answering and include Explore-EQA [29], ConceptGraphs [16] with Frontier Snapshots, 3D-Mem [9], GSMem [26], and HGR [30]. For consistency with prior work, all VLM-exploration methods in this comparison use GPT-4o as the VLM backbone.

B.2 Episodic-Memory Embodied Question Answering

Benchmark Protocol. EM-EQA evaluates embodied question answering from previously collected episodic observations without allowing the agent to actively explore the environment for the current question [8]. It therefore provides a complementary setting to A-EQA for evaluating how effectively persistent spatial memory organizes, retrieves, and verifies accumulated evidence. For this evaluation, DCF-Nav operates in a memory-only configuration: active navigation, frontier selection, Query-Ambiguity, and Environment-Readiness are disabled, while the Gaussian memory, object-centric volumetric representation, query-conditioned evidence retrieval, and volumetric verification components are retained. This configuration evaluates persistent Gaussian memory and volumetric verification without DCF-guided active exploration.

Metrics and Baselines. We report LLM-Match together with the average number of visual frames used by each method. The comparison includes language-only and caption-based baselines, Multi-Frame reasoning, 3D-Mem [9], and the proposed memory-only DCF-Nav configuration. DCF-Nav uses 3.1 visual frames on average, comparable to the average frame usage reported for 3D-Mem.

B.3 GOAT-Bench

Benchmark Protocol. GOAT-Bench [6] evaluates sequential navigation to heterogeneous goals specified by category names, language descriptions, or target images. Unlike episodic single-goal navigation, the environment is not reset between subtasks, requiring previously collected observations to remain available for subsequent goals. The official "Val Unseen" split contains 36 scenes with 10 lifelong-navigation episodes per scene, resulting in 360 episodes in total. Each episode contains between 5 and 10 sequential navigation subtasks. We evaluate all 360 episodes without filtering any scene, episode, goal modality, or failed subtask, using the official episode budget and stopping criterion.

The ten episode-indexed groups contain 278, 255, 277, 272, 257, 251, 276, 279, 265, and 259 subtasks, respectively, yielding 2,669 navigation subtasks in total. The comparison subset used in prior work contains 278 subtasks corresponding to one episode per scene. Because the number of subtasks varies across episodes, the full-set count cannot be obtained by simply multiplying the subset size by ten.

Metrics. We report Success Rate (SR) and Success weighted by Path Length (SPL) following GOAT-Bench [6]. A subtask is considered successful when the agent stops within 1 meter of the target object. SR measures the fraction of successfully completed subtasks, whereas SPL additionally weights successful navigation according to the ratio between the shortest-path distance and the executed path length.

Baseline Details. The comparison includes the original GOAT-Bench baselines: Modular GOAT [4], Modular CLIP on Wheels [48], SenseAct-NN Monolithic, and SenseAct-NN Skill Chain [6]. We additionally compare against recent VLM- and memory-based navigation methods, including VLMnav [49], DyNaVLM [50], TANGO [51], MTU3D [35], ConceptGraphs [16], 3D-Mem [9], GSMem [26], and HGR [30]. DCF-Nav uses GPT-4o as the VLM backbone in this evaluation.

C Additional GOAT-Bench Results

C.1 Results by Goal Modality and Evaluation Subset

Tables 5–10 provide a finer-grained breakdown of GOAT-Bench performance across category-, language-, and image-specified goals and across the ten episode-indexed subsets. "Number" denotes the number of navigation subtasks in each group. DCF-Nav maintains strong performance across heterogeneous goal modalities, while language-specified goals remain relatively more challenging due to their greater reliance on contextual grounding, spatial descriptions, and instance-level disambiguation.

Table 5: Detailed results on the GOAT-Bench "Val Unseen" full validation set, broken down by goal modality. "Number" denotes the number of navigation subtasks.

	Full Validation Set		
	SR ↑	SPL ↑	Number
Total	68.2	50.7	2669
Category	73.0	50.3	991
Language	61.8	47.0	856
Image	69.1	55.0	822

Table 6: Detailed results on GOAT-Bench subsets 1 and 2, broken down by goal modality.

	Subset 1			Subset 2		
	SR ↑	SPL ↑	Number	SR ↑	SPL ↑	Number
Total	**75.9**	**56.4**	278	64.3	48.7	255
Category	80.8	55.9	99	68.9	48.5	90
Language	69.2	52.0	91	59.7	45.1	77
Image	77.3	61.4	88	63.6	52.1	88

Table 7: Detailed results on GOAT-Bench subsets 3 and 4, broken down by goal modality.

	Subset 3			Subset 4		
	SR ↑	SPL ↑	Number	SR ↑	SPL ↑	Number
Total	74.0	54.6	277	65.1	46.8	272
Category	76.2	52.4	105	73.0	47.7	100
Language	69.7	52.9	89	54.3	41.2	92
Image	75.9	59.3	83	67.5	52.1	80

Table 8: Detailed results on GOAT-Bench subsets 5 and 6, broken down by goal modality.

	Subset 5			Subset 6		
	SR ↑	SPL ↑	Number	SR ↑	SPL ↑	Number
Total	69.3	51.5	257	68.1	49.7	251
Category	73.5	50.6	98	70.7	49.0	99
Language	64.4	48.7	73	62.8	46.5	78
Image	68.6	54.8	86	70.3	54.0	74

Table 9: Detailed results on GOAT-Bench subsets 7 and 8, broken down by goal modality.

	Subset 7			Subset 8		
	SR ↑	SPL ↑	Number	SR ↑	SPL ↑	Number
Total	67.8	51.7	276	65.6	48.9	279
Category	73.2	52.4	112	72.2	47.9	108
Language	60.4	47.9	91	57.5	46.7	87
Image	68.5	55.5	73	65.5	52.4	84

Table 10: Detailed results on GOAT-Bench subsets 9 and 10, broken down by goal modality.

	Subset 9			Subset 10		
	SR ↑	SPL ↑	Number	SR ↑	SPL ↑	Number
Total	67.9	50.2	265	63.3	47.8	259
Category	71.1	50.0	90	68.9	47.9	90
Language	63.2	46.8	95	56.6	41.7	83
Image	70.0	54.6	80	64.0	53.5	86

C.2 Qualitative Lifelong Navigation Results

Figures 4–8 present representative qualitative results on GOAT-Bench, covering both fully completed and partially completed lifelong navigation episodes. For the single-scene examples in Figs. 4–7, panels (a)–(d) show the representative target views from selected subtasks, while panel (e) summarizes the final navigation result, including the current task query, RGB-D observation, online 3DGS reconstruction, top-down navigation map, and accumulated global Gaussian map. The top-down map further visualizes the executed trajectory together with the initial position and the locations at which individual subtasks are completed.

These examples illustrate how DCF-Nav incrementally maintains and reuses persistent object-centric spatial information across sequentially changing goals. In Scene 00800 (Fig. 4), the white pillow located during an early subtask remains encoded in the accumulated navigation state, allowing the agent to directly return to the previously observed target when it reappears in the final subtask. Scene 00877 (Fig. 7) further demonstrates sustained navigation over a long sequence of ten subtasks, all of which are successfully completed. Figures 5, 6, and 8 provide additional representative cases with partial task completion across different scenes, illustrating the behavior of DCF-Nav under more challenging long-horizon navigation sequences. The complete navigation trajectories and execution processes corresponding to these examples are provided in the supplementary video.

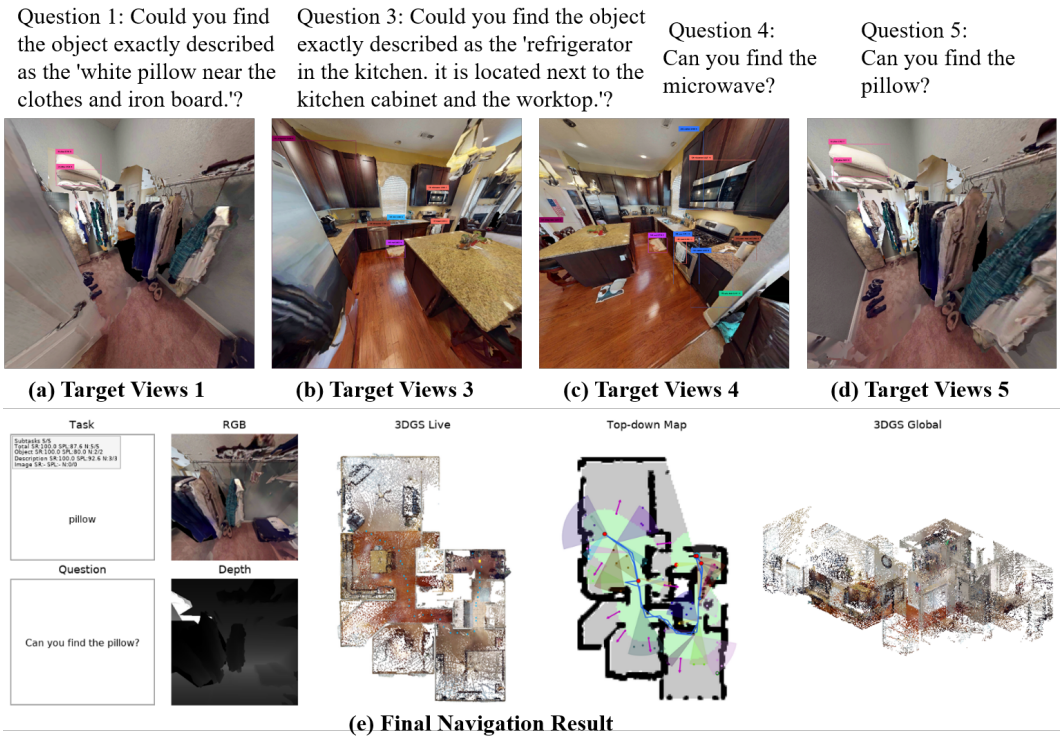


Figure 4: Qualitative result on GOAT-Bench Scene 00800. (a)–(d) Representative target views from selected subtasks. (e) Final navigation result, including the task query, RGB-D observation, online 3DGS reconstruction, top-down navigation map, and accumulated global Gaussian map. After locating the white pillow in an early subtask, DCF-Nav leverages the previously established persistent navigation state to directly return to it when the same target reappears in the final subtask. DCF-Nav successfully completes all five sequential subtasks, achieving 100.0% SR and 87.6 SPL. The complete navigation trajectory and execution process are provided in the supplementary video.

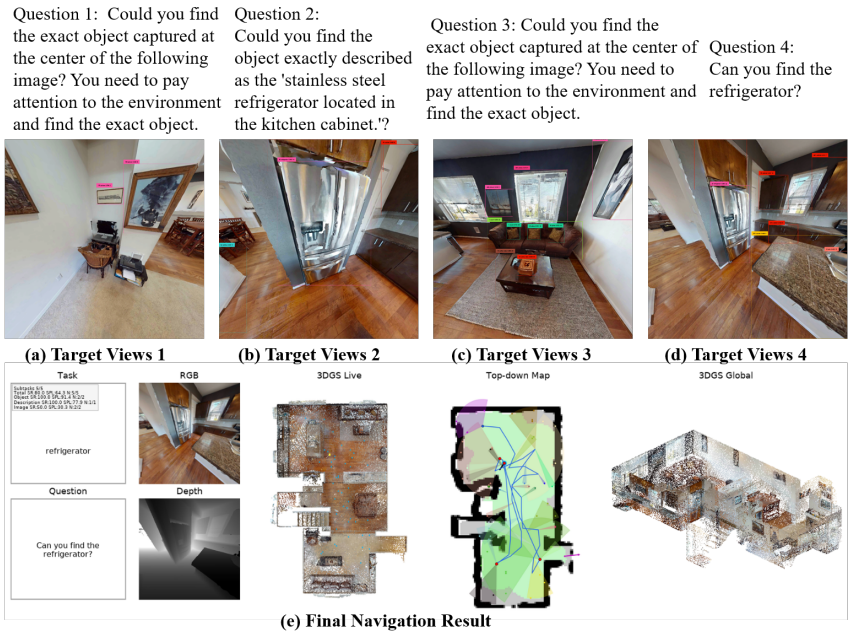


Figure 5: Qualitative result on GOAT-Bench Scene 00832. (a)–(d) Representative target views from selected subtasks. (e) Final navigation result. DCF-Nav completes four of five subtasks, achieving 80.0% SR and 64.3 SPL. The complete navigation trajectory and execution process are provided in the supplementary video.

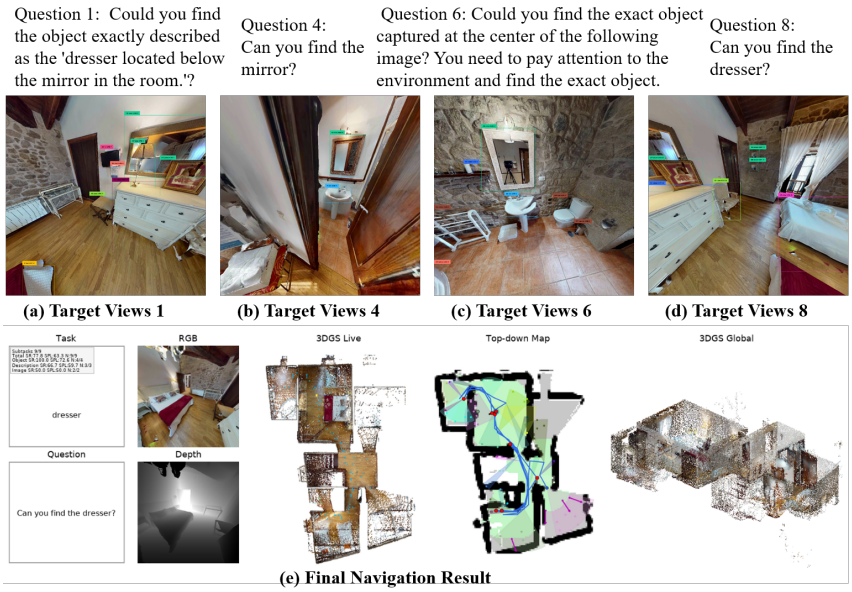


Figure 6: Qualitative result on GOAT-Bench Scene 00849. (a)–(d) Representative target views from selected subtasks. (e) Final navigation result. DCF-Nav completes seven of nine subtasks, achieving 77.8% SR and 63.3 SPL. The complete navigation trajectory and execution process are provided in the supplementary video.

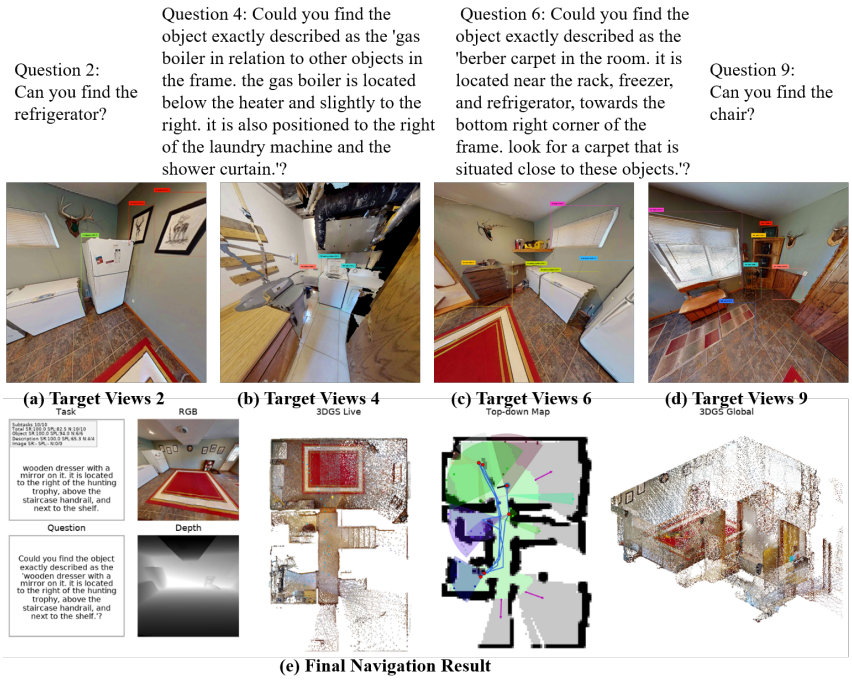


Figure 7: Qualitative result on GOAT-Bench Scene 00877. (a)–(d) Representative target views from selected subtasks. (e) Final navigation result. DCF-Nav successfully completes all ten sequential subtasks, achieving 100.0% SR and 82.5 SPL. The complete navigation trajectory and execution process are provided in the supplementary video.

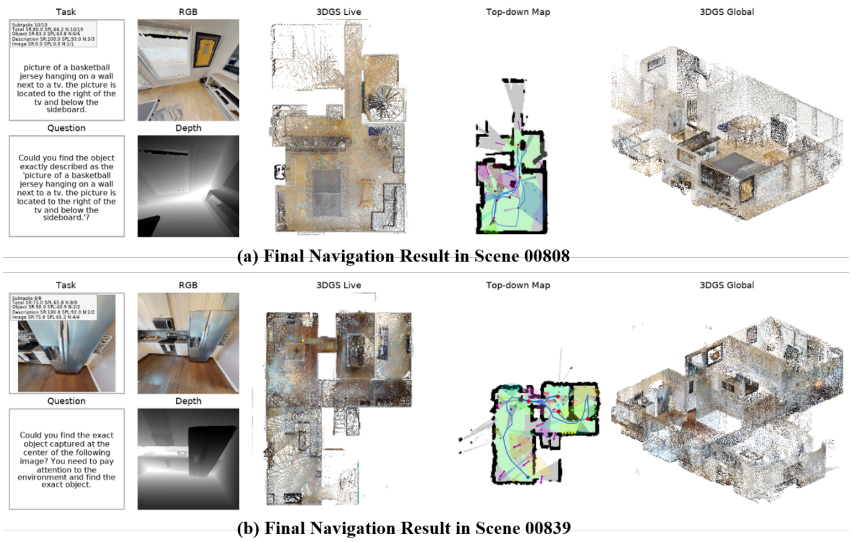


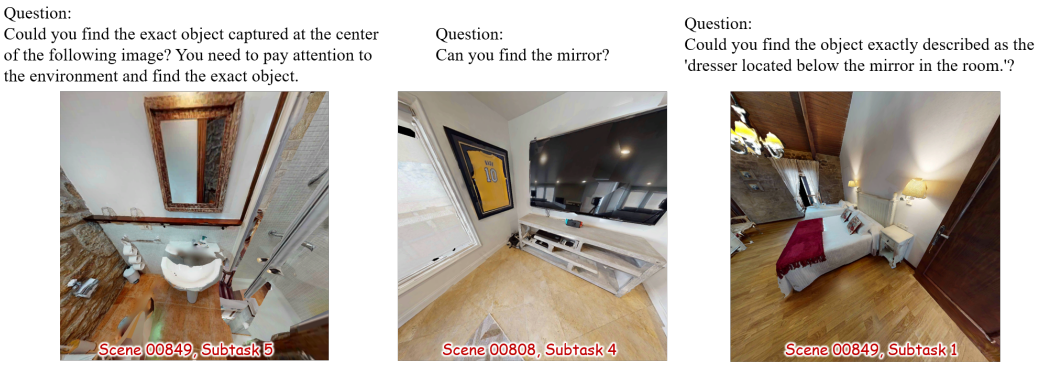
Figure 8: Qualitative results on two GOAT-Bench scenes. (a) Scene 808, where DCF-Nav achieves 80.0% SR and 66.2 SPL. (b) Scene 839, where DCF-Nav achieves 75.0% SR and 65.8 SPL. The complete navigation trajectories and execution processes are provided in the supplementary video.

C.3 Analysis of Challenging Cases

Figure 9 presents three representative unsuccessful cases on GOAT-Bench, covering image-, object-, and description-specified goals. These examples primarily involve ambiguous visual references, appearance-driven semantic confusion, and fine-grained object-category or relational ambiguity.

In Fig. 9(a), an image-goal query for a mirror in Scene 00849 is repeatedly associated with a nearby sink appearing in the same local context. The reference therefore provides a visually relevant scene cue, but the salient contextual object dominates the target interpretation. In Fig. 9(b), a reflective region in Scene 00808 is interpreted as a mirror under an object-category query. Reflection provides a plausible local cue, but it is insufficient to establish the semantic identity of the queried object. Figure 9(c) shows a description-goal query for a dresser located below a mirror in Scene 00849, where a nearby nightstand in the same room is selected instead. This example reflects the difficulty of distinguishing semantically similar furniture under fine-grained relational descriptions.

Together, these cases indicate that ambiguous target references and imperfect upstream semantic evidence can propagate to downstream target selection, particularly when visually similar or spatially adjacent objects coexist in the same local scene.



(a) Image Goal: Mirror **(b) Object Goal: Mirror** **(c) Description Goal: Dresser**

Figure 9: Representative challenging cases on GOAT-Bench across three goal modalities. (a) Image-goal case in Scene 00849 (Subset 1, Subtask 5), where a mirror reference is repeatedly associated with a nearby sink in the same local context. (b) Object-goal case in Scene 00808 (Subset 7, Subtask 4), where a reflective region provides a misleading semantic cue for the queried mirror category. (c) Description-goal case in Scene 00849 (Subset 1, Subtask 1), where a nearby nightstand is selected for the target dresser specified by its relation to a mirror. The highlighted annotations are added only for visualization and are not provided as input to the navigation system.

C.4 Analysis of Representative Successful Cases

Figure 10 highlights three representative successful behaviors of DCF-Nav on GOAT-Bench, including persistent evidence reuse across sequential goals, target-oriented candidate refinement, and recovery after rejecting an inconsistent candidate.

In Fig. 10(a), the same microwave in Scene 00891 is successfully retrieved across multiple sequential subtasks with different query forms. Earlier stored observations are reused for subsequent description-specified goals, while a later image-specified goal is associated with the same persistent object. All three subtasks satisfy the navigation success criterion defined in Appendix B.3. This sequence demonstrates that the accumulated navigation state preserves task-relevant object evidence across both time and goal modality, avoiding repeated rediscovery of previously observed targets.

Figure 10(b) shows a description-specified picture query in Scene 00808. An initial candidate corresponds to a cabinet in the relevant local context, whereas subsequent candidate refinement identifies the requested picture. The agent ultimately reaches the correct target with a final distance of 0.241 m. This example demonstrates how object-level candidate reasoning can refine a broadly relevant scene interpretation into a target-consistent navigation decision.

In Fig. 10(c), an image-specified refrigerator query in Scene 00839 initially produces a chair candidate. The inconsistent candidate is rejected, after which continued exploration identifies the refrigerator at the next decision step and successfully completes the task at a final distance of 0.199 m. The complete sequence shows that



Subtask 3, Description, Question:
Could you find the object exactly described as the 'microwave on the kitchen shelf, which is located to the left of the kettle and below the coffee machine. the microwave is made of stainless steel.'?

Subtask 5, Description, Question:
Could you find the object exactly described as the 'microwave on the kitchen shelf, which is located to the left of the kettle and below the coffee machine. the microwave is made of stainless steel.'?

Subtask 6, Image, Question:
Could you find the exact object captured at the center of the following image? You need to pay attention to the environment and find the exact object.

(a) Cross-Task and Cross-Modal Reuse



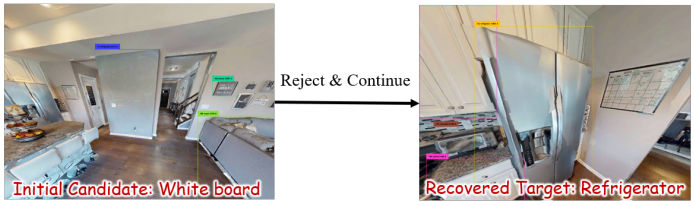
Subtask 3, Description, Question:
Could you find the object exactly described as the 'picture in the kitchen with a fireplace and a painting on the wall. the picture is located on the kitchen counter, to the right of the fireplace and above the ornament.'?



(b) Target-Oriented Candidate Refinement



Subtask 1, Description, Question:
Could you find the exact object captured at the center of the following image? You need to pay attention to the environment and find the exact object.



(c) Rejection and Closed-Loop Recovery

Figure 10: Representative successful cases on GOAT-Bench illustrating complementary behaviors of DCF-Nav. (a) Cross-task and cross-modal evidence reuse in Scene 00891 (Subset 1, Subtasks 3, 5, and 6), where the same microwave remains accessible across sequential description- and image-specified goals through the accumulated navigation state. (b) Target-oriented candidate refinement in Scene 00808 (Subset 7, Subtask 3), where an initial cabinet candidate is refined to the requested picture, leading to successful navigation. (c) Rejection and closed-loop recovery in Scene 00839 (Subset 1, Subtask 1), where an inconsistent chair candidate for an image-specified refrigerator goal is rejected, continued exploration produces the correct refrigerator candidate, and the task is subsequently completed.

candidate rejection is integrated with continued navigation rather than acting as an isolated filtering operation, enabling recovery from an incorrect intermediate selection.

Together, these examples demonstrate complementary behaviors of the evaluated DCF-Nav pipeline: persistent object-centric evidence remains reusable across long-horizon subtasks, candidate-level reasoning improves target specificity, and closed-loop exploration allows the agent to recover from intermediate semantic errors.

D Repeated-Run Statistical Reliability

The main benchmark and ablation tables retain the point estimates used for direct comparison with prior work. To additionally quantify run-to-run variability, we independently evaluate Full DCF-Nav and each ablation

variant over five runs under otherwise identical settings on A-EQA and the fixed GOAT-Bench subset. Table 11 reports the mean and sample standard deviation across these runs. These repeated evaluations are reported separately from the point estimates in the main paper and are intended to assess the stability of the observed component-level trends.

Table 11: Repeated-run uncertainty analysis over five independent runs. Values are mean ± sample standard deviation.

	A-EQA		GOAT-Bench (Subset)	
Variant	LLM-Match ↑	LLM-Match SPL ↑	SR ↑	SPL ↑
Full DCF-Nav	**56.25 ± 0.52**	**44.51 ± 1.14**	**74.52 ± 1.53**	**56.17 ± 0.83**
w/o Object-Centric Units	53.12 ± 0.70	41.60 ± 0.49	69.91 ± 1.23	50.57 ± 0.93
w/o Query-Ambiguity Field	52.63 ± 0.59	41.31 ± 0.46	69.21 ± 1.76	49.50 ± 2.65
w/o Environment-Readiness Field	54.23 ± 0.64	40.49 ± 0.56	70.32 ± 1.27	48.86 ± 0.54
Single Coupled Score	51.31 ± 0.77	39.26 ± 0.59	67.35 ± 1.24	47.67 ± 0.62
w/o Volumetric Verification	49.64 ± 0.80	39.14 ± 0.74	68.60 ± 1.42	49.40 ± 0.61

Full DCF-Nav achieves the highest mean performance across all four metrics, while every ablation remains below the complete method under repeated evaluation. The repeated-run results therefore preserve the component-level trends observed from the point estimates in the main paper rather than depending on a single favorable evaluation. The reported sample standard deviations further quantify the variability associated with repeated runs.

E Additional Analysis of Decoupled Complementary Fields

E.1 Predictive Validity of Query-Ambiguity

The component ablation in the main paper establishes the utility of Query-Ambiguity, but does not by itself determine whether the field identifies regions that are likely to provide additional task-relevant evidence. We therefore evaluate Query-Ambiguity on 914 pre-discovery candidates using offline oracle-shadow target-evidence gain and subsequent target visibility. These diagnostic signals are used only for analysis and are unavailable to the navigation policy.

Table 12 reports the association between each ambiguity signal and subsequent evidence acquisition. The complete Query-Ambiguity Field achieves an evidence-gain correlation of $\rho_{\mathrm{EG}} = 0.203$ and an AUC of 0.621 for subsequent target visibility. Target visibility increases from 11.65% in the lowest within-decision ambiguity quartile to 18.80% in the highest, corresponding to a 7.15-point gap. The complete field also outperforms its individual constituent terms. After retaining only the first observation of each unique query–frontier pair, the association remains positive, indicating that the trend is not solely caused by repeatedly observing the same candidates.

Table 12: Predictive validity of Query-Ambiguity. "Top" and "Bottom" denote target visibility in the highest and lowest within-decision ambiguity quartiles, respectively.

Signal / Subset	N_{cand}	ρ_{EG}	AUC	Top (%)	Bottom (%)	Δ (pp)
Full ambiguity U_t	914	0.203	0.621	18.80	11.65	+7.15
Room entropy $\bar{H}_t$	914	0.018	0.521	14.66	13.53	+1.13
Cue disagreement Δ_t	914	0.094	0.562	16.17	12.78	+3.39
Query-support entropy $h(S_t)$	914	0.156	0.594	17.67	11.65	+6.02
Full U_t, unique query–frontier pairs	173	0.142	0.587	16.62	10.04	+6.58

E.2 Environment-Readiness and Field Coupling

Environment-Readiness is a query-agnostic soft reliability field rather than a binary feasibility mask. Because it is recomputed after every action, regional readiness evolves as frontier exposure, unknown-space extent, Gaussian density and coverage, free-space consistency, and geometric support change with newly acquired observations. The proposed multiplicative coupling weights query-dependent semantic support and ambiguity by Environment-Readiness, while the independent novelty term preserves the ability to explore previously unobserved regions.

We compare the proposed multiplicative coupling with additive coupling $[S_t + U_t + A_t]/3$, a soft gate $[\eta_{\text{gate}} + (1 - \eta_{\text{gate}})A_t][S_t + U_t]/2$ with $\eta_{\text{gate}} = 0.2$, and a no-readiness variant $[S_t + U_t]/2$. Table 13 shows that multiplicative coupling achieves the strongest SR and SPL in this controlled diagnostic evaluation.

Table 13: Sensitivity to alternative DCF coupling formulations in the controlled diagnostic evaluation.

Coupling	SR ↑	ΔSR	SPL ↑	ΔSPL
Multiplicative	**75.90**	–	**56.40**	–
Additive	66.15	-9.75	50.32	-6.08
Soft gate, $\eta_{\text{gate}} = 0.2$	68.04	-7.86	54.83	-1.57
No readiness	64.49	-11.41	50.78	-5.62

To examine whether Environment-Readiness suppresses necessary exploration, we additionally analyze candidates in the lowest within-decision readiness quartile. Only one of 84 target-relevant candidates falls into this lowest-readiness group, and this candidate remains selectable by the complete policy. This result is consistent with the novelty term preserving informative exploration even when local readiness is low.

Table 14 further reports executed-action outcomes. Multiplicative coupling achieves the lowest dead-end and backtracking rates, at 10.32% and 21.00%, respectively, while producing the largest explored-map gain and mean ground-truth progress. These results are consistent with Environment-Readiness reducing structurally unproductive branches while Query-Ambiguity continues to resolve competition among task-relevant candidates.

Table 14: Executed-action outcomes under alternative readiness couplings.

Coupling	Dead-end ↓	Backtracking ↓	Mean Explored Gain ↑	Mean GT Progress ↑
Multiplicative	**29/281 = 10.32%**	**46/219 = 21.00%**	**102.26 cells**	**0.112 m**
Additive	67/369 = 18.16%	114/305 = 37.38%	76.41 cells	0.055 m
Soft gate, $\eta_{\text{gate}} = 0.2$	73/362 = 20.17%	100/297 = 33.67%	78.67 cells	0.023 m
No readiness	70/398 = 17.59%	119/331 = 35.95%	80.62 cells	0.037 m

F Additional Analysis of Volumetric Verification

The main ablation shows that removing volumetric verification degrades both A-EQA and GOAT-Bench performance. We further examine whether the verifier actively revises candidate decisions and how it differs from synthesizing additional views for VLM reasoning.

GSMem-style retrieval and rendering improve the visual evidence presented to the VLM by synthesizing informative viewpoints, whereas DCF-Nav uses persistent volumetric evidence to evaluate whether a candidate is sufficiently supported by semantic, geometric, multi-view, and temporal evidence. We therefore conduct a controlled comparison between volumetric verification and sampled render-to-VLM evidence on the GOAT-Bench diagnostic subset.

As shown in Table 15, volumetric verification alone achieves the strongest SR and SPL in this controlled comparison without additional rendering latency. More importantly, the verifier changes 37 of 205 initial VLM-selected candidates, corresponding to 18.0% of the evaluated decisions. This shows that verification revises candidate selections rather than only accumulating confidence in an unchanged VLM prediction.

Table 15: Controlled comparison of volumetric verification and sampled render-to-VLM evidence. "Revisions" denotes initial VLM-selected candidates changed by the verifier.

Variant	SR ↑	SPL ↑	Added Render Latency ↓	Revisions
Volumetric verification only	**75.9**	**56.4**	**0.00 s**	37/205 (18.0%)
Render-to-VLM only	67.1	50.4	1.51 s	–
Rendering + verification	69.3	51.3	1.68 s	26/205 (12.7%)

These results indicate that synthesized views and volumetric verification serve different roles. Rendering modifies the visual evidence supplied to the VLM, whereas volumetric verification evaluates whether a candidate is consistently supported by the accumulated 3D evidence before final commitment.

G Robustness to Geometric Perturbations

DCF-Nav relies on persistent fusion of RGB-D observations and camera poses, making robustness to geometric estimation errors important for the reliability of the Gaussian navigation state. We therefore conduct a controlled perturbation study on a fixed GOAT-Bench diagnostic subset by injecting pose and depth noise at the Gaussian-mapping interface.

Pose perturbations are sampled in the local camera frame as

$$\delta\mathbf{t} \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \mathbf{I}_3), \qquad \delta\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \sigma_r^2 \mathbf{I}_3), \tag{36}$$

where $\delta\boldsymbol{\omega}$ is mapped to a valid rotation through the $\mathrm{SO}(3)$ exponential map. Depth noise is modeled using

$$\sigma_z(z) = \sqrt{\sigma_{\mathrm{add}}^2 + (\sigma_{\mathrm{rel}} z)^2}. \tag{37}$$

To isolate the effect of corrupted geometry on persistent Gaussian fusion and verification, simulator state, navigation execution, TSDF integration, and metric computation retain GT geometry.

Table 16: Robustness to pose and depth perturbations on the fixed GOAT-Bench diagnostic subset. Δ is measured relative to the clean setting.

Variant	Noise	SR ↑	ΔSR	SPL ↑	ΔSPL
DCF-Nav (Clean)	None	**75.90**	–	**56.40**	–
Pose-only (Low)	1.5 cm, 0.75°	67.92	-7.98	51.74	-4.66
Pose-only (Medium)	3.0 cm, 1.5°	64.15	-11.75	50.31	-6.09
Pose-only (High)	6.0 cm, 3.0°	62.26	-13.64	49.21	-7.19
Depth-only (Low)	0.5 cm + 1.5%	69.83	-6.07	52.99	-3.41
Depth-only (Medium)	1.0 cm + 3.0%	71.70	-4.20	52.17	-4.23
Depth-only (High)	2.0 cm + 6.0%	65.71	-10.19	50.18	-6.22
Pose + Depth (Low)	1.5 cm, 0.75°, 0.5 cm + 1.5%	64.27	-11.63	50.25	-6.15
Pose + Depth (Medium)	3.0 cm, 1.5°, 1.0 cm + 3.0%	65.14	-10.76	49.00	-7.4
Pose + Depth (High)	6.0 cm, 3.0°, 2.0 cm + 6.0%	60.24	-15.66	46.07	-10.33

Table 16 shows an overall degradation as geometric corruption increases. Pose perturbations generally cause larger degradation than depth perturbations at comparable levels, consistent with cross-view misalignment directly affecting persistent Gaussian fusion. Under the medium combined perturbation, DCF-Nav retains approximately 89% of its clean SR and SPL, whereas severe combined pose and depth corruption produces a substantially larger decline. These results indicate that the persistent Gaussian evidence and verification pipeline tolerate moderate geometric noise but remain sensitive to severe cross-view misalignment.

H Runtime, Memory, and Scalability Analysis

We analyze the computational characteristics of DCF-Nav at three levels: module-level runtime, persistent-map memory growth, and online-update scalability. The results show that Gaussian-side mapping and DCF computation support online incremental operation, whereas the current end-to-end control loop is dominated by high-level VLM-mediated reasoning rather than by persistent Gaussian-map maintenance.

H.1 Module-Level Runtime

Table 17 reports the median and P90 latency of the major components. The complete frame-level perception and mapping pipeline requires 330.87 ms at the median. Within this pipeline, online 3DGS updating takes 44.09 ms per accepted frame and remains below 73.40 ms at P90. Conditional Gaussian optimization takes 18.70 ms when triggered and is included in the reported 3DGS-update latency. Frame-level CLIP encoding and detection-to-3D lifting require 9.69 ms and 17.45 ms, respectively.

At the high-level reasoning stage, persistent Gaussian semantic lifting requires 866.17 ms per control step, whereas frontier geometry computation and DCF query ranking require only 3.71 ms and 1.75 ms, respectively. The complete high-level control step requires 5.73 s at the median, primarily due to prompt preprocessing and the remote GPT-4o request, which require 833.12 ms and 3.34 s, respectively. Therefore, the current implementation is not end-to-end real-time, but the primary bottleneck lies in VLM-mediated reasoning rather than Gaussian-map updating or complementary-field computation.

Table 17: Module-level runtime profiling of DCF-Nav. Conditional Gaussian optimization is triggered only when required and is included in the reported online 3DGS-update latency.

Module	Median	P90
Frame-level perception and mapping		
Map-pipeline component sum	330.87 ms	479.86 ms
YOLO detection	12.82 ms	14.25 ms
SAM segmentation	156.65 ms	174.73 ms
Frame-level CLIP encoding	9.69 ms	13.16 ms
Detection-to-3D lifting	17.45 ms	33.11 ms
Object association and fusion	16.07 ms	29.92 ms
TSDF integration	12.65 ms	13.97 ms
Online 3DGS update	**44.09 ms**	73.40 ms
Conditional Gaussian optimization	18.70 ms	24.35 ms
High-level semantic reasoning and control		
Persistent Gaussian semantic lifting	866.17 ms	957.62 ms
Frontier geometry computation	3.71 ms	5.54 ms
DCF query ranking	**1.75 ms**	2.09 ms
Prompt preprocessing	833.12 ms	963.48 ms
GPT-4o API wait	3.34 s	5.12 s
Control-step wall-clock latency	**5.73 s**	**7.93 s**

Table 18: Persistent-map and GPU-memory growth over a long-horizon GOAT-Bench episode.

Accepted Frame	Total Gaussians	Active Gaussians	Gaussian Tensors	Torch Allocated	Torch Reserved	CUDA Used
0	0.102M	0.102M	6.28 MiB	2,733.12 MiB	3,022 MiB	5,242.88 MiB
90	7.330M	0.127M	447.41 MiB	3,252.35 MiB	3,576 MiB	5,842.69 MiB
180	14.122M	0.116M	861.98 MiB	3,680.38 MiB	4,050 MiB	6,347.56 MiB
270	21.062M	0.114M	1,285.52 MiB	4,101.96 MiB	4,470 MiB	6,696.81 MiB
360	28.175M	0.131M	1,719.72 MiB	4,593.53 MiB	5,026 MiB	7,208.13 MiB

H.2 Persistent-Map Memory Growth

Because DCF-Nav maintains a persistent Gaussian representation over long-horizon exploration, we additionally measure GPU-memory growth as the accumulated scene representation expands. Table 18 reports a representative long-horizon GOAT-Bench episode. As the map grows from 0.102M to 28.175M Gaussians, the storage occupied by Gaussian tensors increases from 6.28 MiB to 1,719.72 MiB, corresponding to approximately 61 MiB per million Gaussians. Torch-allocated memory increases from 2.73 GiB to 4.59 GiB, while total device-level CUDA usage increases from 5.24 GiB to 7.21 GiB.

Importantly, the number of Gaussians involved in each local update remains much smaller than the accumulated map. Across the same trace, the active set remains approximately 0.11M–0.13M Gaussians even when the full persistent map grows beyond 28M primitives. This separation between persistent storage and the local active working set allows the system to retain long-horizon evidence without optimizing the complete map at every frame.

H.3 Online-Update Scalability

We next examine whether Gaussian-map update latency increases as the persistent map becomes larger. Table 19 groups accepted frames according to the size of the accumulated map. Although the global map expands from below 1M to more than 28M Gaussians, the median active working set remains within 0.163M–0.186M Gaussians. Correspondingly, the median online-update latency remains within 41.27–49.15 ms, while the P90 latency stays within 71.99–75.80 ms.

These results indicate that persistent storage grows approximately with the explored environment, whereas per-frame map-update cost is primarily determined by the bounded local active set rather than the full accumulated map. The persistent Gaussian representation therefore incurs increasing storage cost over long-horizon exploration but does not produce a proportional increase in online update latency. Together with the 1.75-ms

DCF ranking cost, this analysis shows that the principal computational limitation of the current implementation is high-level VLM reasoning rather than Gaussian-state maintenance or complementary-field evaluation.

Table 19: Online 3DGS-update latency as the accumulated persistent map expands.

Accumulated Map Size	Median Active Gaussians	Update Median	Update P90
0.90M–7.33M	0.177M	44.84 ms	71.99 ms
7.45M–14.12M	0.163M	43.38 ms	75.80 ms
14.21M–21.06M	0.186M	41.27 ms	72.83 ms
21.12M–28.18M	0.177M	49.15 ms	73.75 ms

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are carefully aligned with the paper's actual contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the proposed method at the end of the main text.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include any theoretical assumptions or proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the necessary details in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release all necessary code upon acceptance of the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the necessary details in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the usual format used in previous related works to report and compare the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the relevant information in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of the assets used in this paper are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use GPT-4o as the VLM backbone for query-conditioned reasoning, exploration decisions, and answer handoff in A-EQA, as described in the method and experimental setup. Any LLM use for writing or editing is not part of the scientific contribution.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.