SGC-Nav: Scene Graph Conditioned Multi-LLM Consensus for Mapless Zero-Shot Object Navigation

Yikai Xue[1,*], Yachao Wang[1,*], Yuanqi Qiu[1], Yinchuan Wang[1], Yiming Wu[1], Yue Wang[2] Chaoqun Wang[1,†]

***Abstract*— Without a prior metric map, zero-shot open-vocabulary object navigation using only RGB observations remains highly challenging. Existing methods often over-rely on the current observation or encode history as text sequences. This makes topological connectivity, revisit traces, and target-absence cues hard to query and exploit. As a result, robots can suffer from local deadlocks, repeated revisits, and unreliable stopping. To address this, we propose SGC-Nav, a navigation framework that combines structured memory with consensus-based decision-making. SGC-Nav builds scene graphs from RGB observations. It maintains an online scene graph memory that merges revisits, records connectivity, revisit frequency, and absence cues. Building on this structured context, SGC-Nav performs loop-aware multi-LLM consensus with a goal-conditioned stop judge and feasibility correction to stabilize direction selection and termination. On the public navigation benchmark, SGC-Nav achieves higher success rates and greater path efficiency than prior RGB-only, mapless methods. Its effectiveness is further validated through real-robot indoor experiments.**

I. INTRODUCTION

In real indoor environments such as homes and offices, robots often need to navigate and find a target object given a natural-language goal (e.g., "find a chair"). We study zero-shot object navigation under mapless, RGB-only decision conditions that are more amenable to real-world deployment, where an open-vocabulary natural-language goal is given to guide navigation to the target object [1]. Recent advances in large language models (LLMs) and vision-language models (VLMs) have driven reasoning-based navigation [2]. However, when directly using LLMs/VLMs for high-level reasoning and decision-making, policies driven mainly by the current observation often neglect exploration history—such as explored regions and absence cues—which leads to local deadlocks, repeated revisits, and unreliable stopping [3].

To mitigate these issues, a mainstream line of methods formulates "where to explore" as frontier selection on semantic/value maps, leveraging depth and mapping to provide traversability and global planning signals. For example, some methods build occupancy maps and frontiers from depth, construct language-aligned value maps [4]–[7]; others maintain a 3D voxel belief over target existence for global planning [8];Some incorporate graph-structured cues into value maps for spatial-structure support [9], [10]. However, under RGB-only and mapless constraints, these methods often lack key structural signals provided by depth-based mapping, such as explicit frontiers, global reachability and connectivity, and stable traversability constraints, and thus tend to rely more on local semantic cues, leading to repeated revisits, local loops or deadlocks, and unreliable stopping.

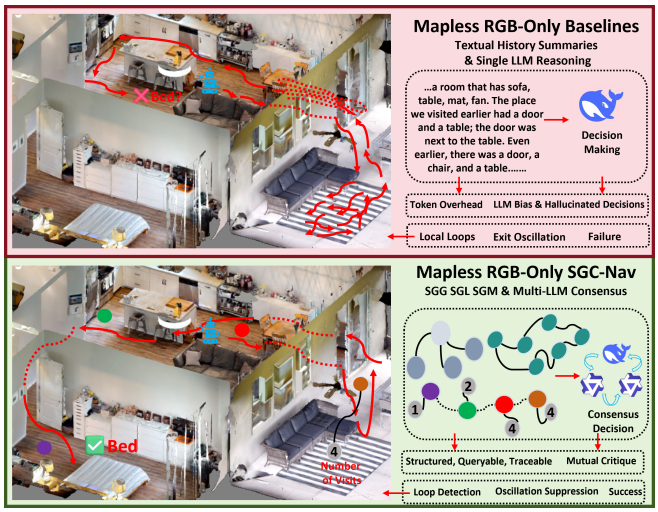

Fig. 1: Qualitative comparison under mapless, RGB-only navigation. Text-based history with single-LLM decision-making (top) can neglect exploration history, leading to local deadlocks, repeated revisits, and unreliable stopping. SGC-Nav (bottom) combines structured scene graph memory with multi-LLM consensus decision-making, reducing loops/oscillation and improving stopping reliability.

Pioneer work has explored injecting history via textual summaries, under the RGB-only mapless setting. For instance, PanoNav [11] reduces deadlocks and improves exploration efficiency through panoramic parsing and a bounded memory queue. However, when history is mainly provided as a natural-language sequence, topological connectivity, revisit cues, and absence cues must be inferred from text, which is more error-prone in confounding scenes and incurs substantial token overhead. Another direction improves the reliability of a single model via multi-round deliberation [12], yet without a queryable and traceable long-term structured memory, single-model policies still struggle to suppress loops and redundant exploration and can be affected by inherent biases that cause occasional misjudgments.

To address the two key gaps, unstable utilization of long-horizon history and bias-prone single-model decisions, we propose SGC-Nav. Fig. 1 provides a qualitative comparison between SGC-Nav and prior mapless RGB-only baselines [11], [13]. Under the strict mapless, RGB-only setting, we

*Equal contribution. †Corresponding author.
[1]School of Control Science and Engineering, Shandong University, China. [2]School of Control Science and Engineering, Zhejiang University, China

This research was in part funded by the Key RD Program of Shandong Province, China, grant number 2024CXGC010210 and 2025CXGC010214.

introduce two key designs: (i) a structured scene graph memory that explicitly encodes spatial connectivity, revisit traces, and absence cues from not observing the target, making history queryable and traceable; and (ii) a multi-LLM consensus decision-making mechanism that treats inter-LLM disagreement as an uncertainty signal. With a limited number of evidence-level critique and aggregation rounds, it yields more robust action selection and stopping decisions. We evaluate SGC-Nav in the Habitat [14] simulator using the HM3D [15] dataset under the ObjectNav benchmark. Under the same conditions, SGC-Nav achieves the best Success Rate (SR) and Success weighted by Path Length (SPL) among RGB-only mapless open-vocabulary baselines. In summary, our main contributions are as follows.

- We design a three-level scene graph representation with a queryable online memory, encoding view-wise traversability, topological connectivity, and absence cues for stable long-horizon reasoning.
- We propose an iterative multi-LLM consensus mechanism where models share voting evidence and perform mutual evaluation, collaboratively resolving disagreements to mitigate single-LLM bias.
- We develop SGC-Nav, an open-vocabulary, mapless, RGB-only navigation framework, achieving state-of-the-art performance on HM3D and demonstrating robust real-robot deployment.

II. RELATED WORK

A. Open-vocabulary Object Navigation

Open-vocabulary object navigation specifies the target category with natural-language and requires cross-category generalization. Accordingly, it often combines VLM-based open-vocabulary recognition with navigation planning to close the loop from goal understanding to exploration and arrival. Mainstream approaches typically rely on depth or other geometric sensing to build metric occupancy maps, semantic maps, or value maps, and perform global planning and exploration prioritization with explicit frontiers and traversability constraints [4]–[8]. To turn the language goal into actionable global search signals, these methods often learn language-aligned value functions or explicitly maintain a belief over target existence (e.g., voxel belief maps), and couple them with reachability and frontier selection to enable more stable region prioritization and path planning at scale [10], [16], [17].

Building on this framework, some works introduce structured representations (e.g., graph-structured context or online scene graph prompting) to strengthen spatial-relation reasoning and long-horizon consistency [9], while others augment frontier-based zero-shot navigation with history-aware VLM prompting, incorporating action/observation history into semantic frontier scoring and decision-making to reduce repeated revisits and improve long-horizon decision stability [18]. However, under RGB-only and mapless conditions, the key structural signals provided by depth-based mapping—such as stable frontiers, global reachability and connectivity, and reliable traversability constraints—are difficult to obtain, weakening the coupling between frontier selection and value/belief signals; as a result, policies often degrade and exhibit unstable exploration, increased repeated revisits, and reduced stopping reliability.

B. Mapless Object Navigation

Mapless object navigation does not explicitly maintain a metric map but makes decisions directly from observations and history. Early representative work such as ZSON [19] enables open-world object-goal navigation via multimodal goal embeddings. Under strict RGB-only, mapless constraints, PixNav [13] drives short-horizon execution with pixel-level goal cues and combines it with LLM-based planning for long-horizon search. PanoNav [11] leverages panoramic parsing and a bounded memory queue to inject history as textual summaries, mitigating local deadlocks. More recently, VLM-centric mapless visual planners further develop this direction: ImagineNav [20] turns long-horizon navigation into iterative "future-view imagination and best-view selection," reducing reliance on explicit mapping pipelines, and ImagineNav++ [21] builds on it with a more compact memory mechanism to maintain long-horizon spatial consistency.

However, when history is encoded only implicitly, topological connectivity, revisit cues, and absence cues still have to be inferred indirectly from the representation, which is more error-prone in confounding scenes and can incur additional token overhead. Even with multi-round deliberation or stronger prompting, single-model decision-making can still neglect exploration history due to the lack of a queryable and traceable long-term structured memory, leading to bias-driven occasional misjudgments [12]. In contrast, we focus on zero-shot navigation under strict mapless, RGB-only constraints by explicitly encoding connectivity and revisit traces with a structured scene graph memory, and combining it with multi-LLM consensus decision-making that leverages inter-model disagreement as an uncertainty signal to improve long-horizon exploration stability and stopping reliability.

III. METHODOLOGY

In this paper, we study zero-shot object navigation under mapless and RGB-only constraints. The navigation framework is shown in Fig. 2. At each step t, the robot receives a six-view panoramic RGB images $I_t = \{I_t^{(v)}\}_{v=0}^{5}$ together with a natural-language goal g, and selects an action from the discrete action set $A = \{\texttt{Stop}, \texttt{MoveAhead}, \texttt{TurnLeft}, \texttt{TurnRight}, \texttt{LookUp}, \texttt{LookDown}\}$. To enable stable long-horizon navigation and reliable stopping without metric maps or localization priors, SGC-Nav decomposes decision-making into three components. (a) Panoramic Structured Parsing: A VLM parses the current panorama augmented with a coordinate scaffold conditioned on g, producing the scene graph of global information SGG_t and the scene graph of local information SGL_t. (b) Scene Graph Memory Maintenance: We update an online structured scene graph memory SGM_t that merges revisited places and records connectivity, revisit traces, and absence cues over time.

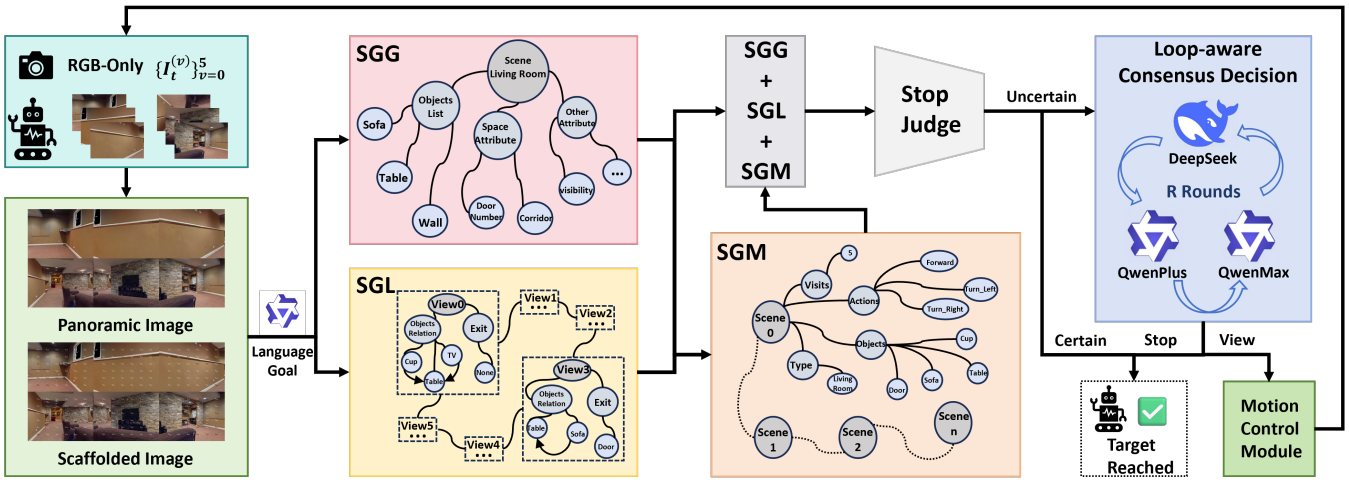

Fig. 2: System overview of SGC-Nav. A VLM parses a six-view panoramic observation augmented with a coordinate scaffold together with the natural-language goal g to produce SGG_t and SGL_t. An online structured scene graph memory SGM_t is then updated to merge revisited places and to record connectivity, revisit traces, and absence cues. The decision module first applies a goal-conditioned Stop Judge. If stopping is not triggered, it uses loop cues and performs loop-aware multi-LLM consensus decision-making to select the next viewing direction, which is executed by a low-level motion controller.

(c) Decision Module: Given (SGG_t, SGL_t, SGM_t), the policy first applies a goal-conditioned Stop Judge. If stopping is not triggered, it derives loop cues from SGM_t and performs loop-aware multi-LLM consensus decision-making to select the next viewing direction, which is executed by a low-level motion controller.

A. Panoramic Structured Parsing

At each decision step t, the robot receives a six-view panoramic RGB images $I_t^{(v)}$ and a natural-language goal g. To provide a reusable spatial reference under mapless, RGB-only constraints, we follow the panoramic input design of [11] and augment each view with a discretized coordinate scaffold, yielding scaffolded views $\tilde{I}_t^{(v)}$. We then concatenate the original six views and the scaffolded six views into two composite images with a fixed layout, denoted as C_t and $\tilde{C}_t$, respectively. During concatenation, we preserve the view identifier and relative orientation for each sub-view to help the model align view indices with headings and semantic cues. Given $(C_t, \tilde{C}_t, g)$, the vision-language model performs structured parsing and outputs two queryable scene graph representations:

$$(SGG_t, SGL_t) = \text{VLM}(C_t, \tilde{C}_t, g). \tag{1}$$

A snippet of the VLM prompting template and the required schema for producing SGG_t and SGL_t is shown in Fig. 3. Here, the term SG denotes a schema-constrained structured scene graph record designed for explicit querying and online updates, rather than a strict graph-theoretic formalization.

The panoramic summary $SGG_t \triangleq \{r_t, \mathcal{L}_t, a_t, v_t\}$, provides a global semantic anchor and coarse structural cues, including an estimated room type r_t, a panoramic object list $\mathcal{L}_t$, coarse structural attributes a_t and an information sufficiency score v_t.

The view-wise action-centric $SGL_t = \{SGL_t^{(v)}\}_{v=0}^{5}$ provides per-view information for action selection. $SGL_t^{(v)} \triangleq \{\mathcal{E}_t^{(v)}, \mathcal{O}_t^{(v)}, \mathcal{R}_t^{(v)}, \hat{r}_t^{(v)}\}$ includes an exit set $\mathcal{E}_t^{(v)}$ that describes candidate traversable exits in view v with confidence, an object set $\mathcal{O}_t^{(v)}$, relation triplets $\mathcal{R}_t^{(v)}$, and view-associated room semantics $\hat{r}_t^{(v)}$.

Overall, SGG_t offers global semantics and a scene summary, while SGL_t provides per-direction traversability cues and coarse target localization, serving as structured inputs to memory maintenance and consensus decision-making.

B. Scene Graph Memory Maintenance

To support stable long-horizon navigation under strict mapless, RGB-only conditions, we maintain an online structured scene graph memory SGM_t. Rather than building a metric map, SGM_t merges revisited observations into homogeneous place nodes and stores place-level connectivity, revisit traces, and absence cues from not observing the target, so that exploration history becomes queryable and traceable for the decision module.

We parameterize the memory state as $SGM_t \triangleq \{\mathcal{P}_t, \mathcal{H}_t\}$, here $\mathcal{P}_t = \{P_i\}_{i=1}^{N_t}$ is the set of place nodes and $\mathcal{H}_t = \{R_{t-H+1}, \ldots, R_t\}$ is a sliding history window with maximum length H, with $H = 25$. Each node P_i is a schema-constrained record that maintains a prototype descriptor ϕ_i for place matching, a neighbor set encoding topological connectivity, and visit statistics used as revisit traces. Each record R_t stores the assigned place id at step t together with goal-relevance evidence, enabling downstream loop cues and stopping decisions.

At step t, given (SGG_t, SGL_t) from panoramic structured parsing, we build a compact place descriptor

$$\phi_t \triangleq (r_t, \mathcal{L}_t, \mathcal{T}_t, d_t, \mathbf{b}_t), \tag{2}$$

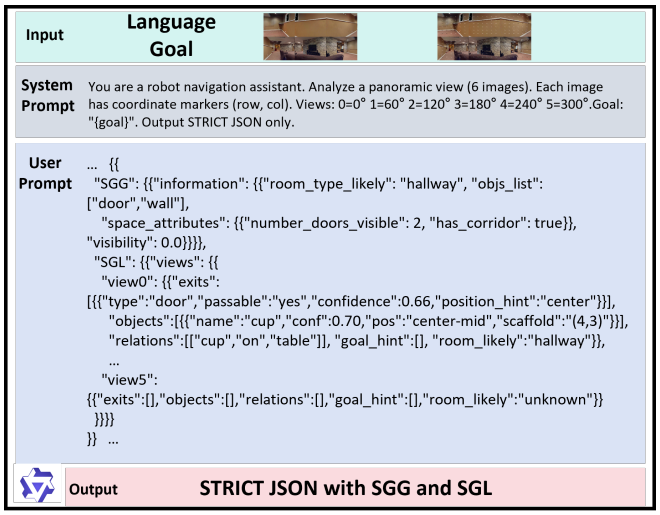

Fig. 3: Prompting template and STRICT-JSON output schema for VLM parsing. Enforcing a schema-constrained JSON format ensures that the VLM reliably produces SGG_t and SGL_t as queryable, updatable records.

where r_t and $\mathcal{L}_t$ come from SGG_t, d_t is the number of visible doors, and $\mathbf{b}_t$ collects coarse landmark attributes. To inject traversability cues into matching, we aggregate view-wise exits into an exit-signature set $\mathcal{T}_t = \bigcup_{v\in\mathcal{V}}\{\tau(e) \mid e \in \mathcal{E}_t^{(v)}\}$ with $\mathcal{V} = \{0,\dots,5\}$, where $\tau(\cdot)$ maps an exit entry to a discrete signature that is stable across revisits.

We then decide whether ϕ_t should be merged into an existing node or assigned to a new node:

$$\hat{s}_t = \arg\max_i \mathrm{Sim}(\phi_t, P_i), \qquad \kappa_t = \max_i \mathrm{Sim}(\phi_t, P_i). \quad (3)$$

If $\kappa_t \geq \vartheta$, we merge the current observation into node $\hat{s}_t$ and update its prototype; otherwise, we create a new node and assign it a new place id.

The similarity function combines semantic overlap, traversability overlap, room consistency, and an attribute penalty:

$$\begin{aligned}\mathrm{Sim}(\phi_t, P_i) = \Pi_{[0,1]}\Big(&\alpha\, J(\mathcal{L}_t, \mathcal{L}_i) + \beta\, J(\mathcal{T}_t, \mathcal{T}_i) \\ &+ \gamma\, \mathbb{I}[r_t = r_i] - \Omega(\phi_t, P_i)\Big),\end{aligned} \quad (4)$$

where $\Pi_{[0,1]}(x) = \min\{1, \max\{0, x\}\}$ and $J(A,B) = |A \cap B|/(|A \cup B| + \varepsilon)$ with $\varepsilon > 0$. We set $\alpha = 0.45$, $\beta = 0.35$, and $\gamma = 0.20$. The penalty term measures differences in door count and landmark attributes:

$$\begin{aligned}\Omega(\phi_t, P_i) = \lambda_d\, \min\Big(1, \frac{|d_t - d_i|}{\kappa_d}\Big) \\ + \sum_{k\in\mathcal{K}} \lambda_k\, \mathbb{I}\big[b_t^{(k)} \neq b_i^{(k)}\big].\end{aligned} \quad (5)$$

where $\mathcal{K}$ is the attribute index set and we set $\kappa_d = 4$, $\lambda_d = 0.15$, and $\lambda_k = 0.08$.

After assignment, we update connectivity by inserting or reinforcing an edge between consecutive place ids $\hat{s}_{t-1}$ and $\hat{s}_t$ in their neighbor sets, and we update revisit traces by refreshing node-level visit statistics for $\hat{s}_t$. To explicitly encode absence cues and keep memory updates aligned with the target, we maintain a three-valued goal relevance state $\zeta_t \in \{P, U, A\}$, indicating Positive, Uncertain, and Absent evidence. We compute a goal-matching confidence $c_t = \max_{v\in\mathcal{V}} \max_{o\in\mathcal{O}_t^{(v)}} \mathrm{conf}_g(o)$ and set $c_t = 0$ if no match exists. The relevance state is then determined by

$$\zeta_t = \begin{cases} P, & c_t \geq \delta, \\ U, & 0 < c_t < \delta, \\ A, & c_t = 0, \end{cases} \qquad \delta = 0.65. \quad (6)$$

Finally, we write the place assignment and ζ_t into the trace record R_t and update the sliding window $\mathcal{H}_t$ online, yielding the updated SGM_t for loop-aware consensus decision-making and reliable stopping.

C. Decision Module

Given the current (SGG_t, SGL_t, SGM_t), the decision module outputs a high-level directive $u_t \in V \cup \{-1\}$, where $u_t = v \in V$ indicates navigating toward viewing direction v, and $u_t = -1$ indicates stopping. Overall, we implement decision-making as a unified closed-loop module that integrates a goal-conditioned Stop Judge, loop cues derived from SGM_t, multi-LLM consensus decision-making, and feasibility correction. Algorithm 1 summarizes the closed-loop decision process at step t.

Goal-conditioned Stop Judge. We reuse the goal confidence score c_t defined earlier, and derive a coarse proximity cue p_t from the scaffold coordinates of goal-matched objects in SGL_t, indicating whether the target lies within a stopping-appropriate interaction range. To improve robustness, we further introduce a cross-step consistency signal and two multi-view variables q_t and m_t, which measure the number of consecutive steps in which the target is observed stably and the number of views that exhibit a strongly goal-relevant state, respectively. Based on these key variables, we define the stopping score $S_t \in [0,1]$ as:

$$S_t = \Pi_{[0,1]}(w_c\, c_t + w_p\, p_t + w_q\, q_t + w_m\, m_t), \quad (7)$$

where $\Pi_{[0,1]}(x) \triangleq \min\{1, \max\{0, x\}\}$. A higher S_t indicates stronger confidence that the target is both visible and sufficiently close. Here w_c, w_p, w_q, w_m are weight parameters. The Stop Judge adopts a two-threshold, three-tier scheme. Let the hard-stop threshold be η_h and the permissive-stop threshold be η_s. If $S_t \geq \eta_h$ and a set of strong constraints is satisfied, a hard stop is triggered. If $\eta_s \leq S_t < \eta_h$, the system enters an allowed-to-stop state, meaning that the subsequent multi-LLM consensus stage is permitted to output the `Stop` command. If $S_t < \eta_s$, stopping is explicitly forbidden, and the decision is restricted to selecting a navigation direction only.

Loop cues from structured memory. Under strict mapless, RGB-only conditions, direction selection based solely on instantaneous observations can easily lead to local deadlocks, repeated revisits, and loop/oscillation. We therefore leverage the exploration history stored in SGM_t to derive a

Algorithm 1 Decision Module of SGC-Nav at Step t

Require: $\mathrm{SGG}_t, \mathrm{SGL}_t, \mathrm{SGM}_t$, g, $\mathcal{M}$, *prev_state*
Ensure: high-level directive $u_t \in \mathcal{V} \cup \{-1\}$

$\ell_t \leftarrow \mathrm{LoopCue}(\mathrm{SGM}_t)$
tier $\leftarrow \mathrm{StopJudge}(\mathrm{SGL}_t, g, \textit{prev_state})$
if *tier* $= \mathrm{HARD}$ **then**
  return $u_t \leftarrow -1$
end if
allow_stop $\leftarrow$ (*tier* $= \mathrm{ALLOW}$)
ctx $\leftarrow \mathrm{BuildContext}(\mathrm{SGG}_t, \mathrm{SGL}_t, \mathrm{SGM}_t, \ell_t, \textit{prev_state})$
prev_votes $\leftarrow \emptyset$
for $r \leftarrow 1$ **to** R **do**
  votes $\leftarrow \emptyset$
  for all $m \in \mathcal{M}$ **do**
    $v_m \leftarrow \mathrm{Vote}(m, \textit{ctx}, \textit{prev_votes}, \textit{allow_stop})$
    votes $\leftarrow$ *votes* $\cup \{v_m\}$
  end for
  if $\mathrm{Consensus}(\textit{votes})$ **then**
    break
  end if
  prev_votes $\leftarrow$ *votes*
end for
$\hat{u} \leftarrow \mathrm{Aggregate}(\textit{votes})$
$u_t \leftarrow \mathrm{FeasibilityCorrection}(\hat{u}, \mathrm{SGL}_t, \textit{allow_stop})$
return u_t

loop state $\ell_t \in \{0, 1\}$. Specifically, the system is flagged as being in a potential loop when the same place node appears with high frequency within the recent history window, or when the global object sets over several consecutive steps exhibit substantial overlap. This signal provides the LLMs with additional long-horizon context, helping suppress redundant exploration and avoid repeatedly selecting directions that return to the same place. In addition, to prevent the robot from oscillating between two rooms—leading to a short-period pattern door $\rightarrow$ non-door $\rightarrow$ door $\rightarrow$ non-door—we impose a *consistency preference* during consensus decision-making. Specifically, when the current place node has been visited multiple times and there exists strong evidence of a highly traversable exit in a particular direction, we temporarily bias the decision to maintain that direction, until the place node changes or the target becomes visible again.

Multi-LLM consensus decision-making. When a hard stop is not triggered, direction selection is performed by a multi-LLM consensus procedure. Let $\mathcal{M} = \{m_1, \ldots, m_M\}$ denote the set of LLMs. Each LLM receives the same conditional inputs: (SGG_t, SGL_t, SGM_t), the loop state ℓ_t, and the recent history of chosen directions. In the r-th round, where $r \in \{1, \ldots, R\}$, output of the m-th LLM is

$$(v_t^{(m,r)}, \pi_t^{(m,r)}, \rho_t^{(m,r)}), \tag{8}$$

where $v_t^{(m,r)} \in V$, $\pi_t^{(m,r)} \in [0, 1]$ is the confidence score associated with its selected action, and $\rho_t^{(m,r)}$ summarizes the key rationale. Starting from round 2, we additionally provide every LLM with the outputs from all LLMs in the previous round as voting evidence, enabling evidence-level mutual evaluation and aggregation under the same structured context and promoting gradual convergence. After R rounds, if the LLMs still disagree, we resolve the decision by voting and confidence aggregation: we first select the candidate with the highest vote count; if there is a tie, we choose the one with the higher average confidence among the tied candidates.

Constraint-based correction. To prevent the decision module from producing an infeasible directive, we apply a feasibility correction before execution. If the selected direction corresponds to a view whose structured exit attributes indicate non-traversability, we override it with an alternative direction chosen based on exit traversability. Moreover, if the multi-LLM consensus proposes a stop action while the system is in the stop-forbidden state, we ignore the stop command and execute the selected navigation direction instead. Finally, we use PixNav [13] as the low-level controller to execute a finite-length action sequence to carry out navigation.

IV. EXPERIMENTS

A. Simulation Experiments

Experimental Setup. We conduct simulation experiments in the Habitat [14] simulator using the HM3D-v0.2 [15] dataset, and evaluate ObjectNav on its validation split comprising 36 building-scale scenes. We report Success Rate (SR) and Success weighted by Path Length (SPL). SR is the fraction of successful episodes, while SPL measures navigation efficiency as the average ratio between the shortest-path length and the robot's executed path length, computed over successful episodes. Success is evaluated following the standard Habitat ObjectNav protocol: when the robot issues the `Stop` action, the Euclidean distance between the robot's position and any instance of the target object must be within the 1.0 m success distance, and the evaluator's visibility criterion must also be satisfied. An episode is marked as a failure if the success condition is not met within 500 actions.

Implementation Details. Experiments are run on a workstation with Ubuntu 20.04 and a single NVIDIA RTX 4090 (24GB) GPU. The robot uses Habitat-style discrete action primitives, with a step size of 0.25 m for `MoveAhead` and 30° for `TurnLeft`/`TurnRight`. For evaluation, we evaluate on 250 randomly sampled episodes from the validation set and report SR and SPL. In our pipeline, we use Qwen3-VL-plus [22] as the multimodal model for panoramic structured parsing to generate SGG and SGL. The multi-LLM consensus in the decision module is formed by DeepSeek-V3.2 [23], Qwen-Max [22], and Qwen-Plus [22].

Comparison Study. We compare SGC-Nav with representative methods on the HM3D-v0.2 validation set. According to each method's reliance on sensing modalities and environment priors (e.g., explicit mapping, depth/pose cues), we group the compared methods into three categories in Tab. I: ID=I includes non-open-vocabulary methods that rely on stronger priors such as mapping and geometric cues; ID=II includes open-vocabulary methods that typically require mapping and/or depth-enhanced sensing; and ID=III includes

TABLE I: Comparison results on the HM3D-v0.2 validation set. In the table, ML/RO/OV indicate whether a method is Mapless, RGB-only, and Open-vocabulary, respectively.

ID	#	Method	ML	RO	OV	SR(%)	SPL(%)
I	1	VLFM [4]	×	×	×	63.6	32.5
II	2	L3MVN [6]	×	×	✓	36.3	15.7
	3	SG-Nav [9]	×	×	✓	49.6	25.5
	4	VLFM* [4]	×	×	✓	56.9	27.5
	5	InstructNav [25]	×	×	✓	58.0	20.9
III	6	ZSON [19]	✓	✓	✓	17.2	7.4
	7	PanoNav [11]	✓	✓	✓	28.0	10.3
	8	PixNav [13]	✓	✓	✓	35.2	13.5
	9	**SGC-Nav (Ours)**	✓	✓	✓	**54.4**	**24.6**

methods that strictly match our setting, i.e., *strict mapless, RGB-only, and open-vocabulary*. Results in ID=I/II are taken from a recent systematic reproduction study on HM3D-v0.2 [24], where VLFM* denotes their open-vocabulary re-deployment of VLFM under the same benchmark setting, while all ID=III baselines are reproduced by us under the same HM3D-v0.2 protocol. ZSON and PixNav are deployed directly from their open-source implementations. PanoNav is re-implemented following the paper description, with the only unspecified dynamic storage horizon set to $n = 10$.

Within Group ID=III, which strictly matches our setting, SGC-Nav achieves 54.4% SR / 24.6% SPL, substantially outperforming prior RGB-only, mapless open-vocabulary baselines: compared to PixNav, it improves SR by 19.2% and SPL by 11.1%; compared to PanoNav, it improves SR by 26.4% and SPL by 14.3%; compared to ZSON, it improves SR by 37.2% and SPL by 17.2%. These results indicate that under strict mapless, RGB-only constraints, relying on implicit history (e.g., text-based summaries) and single-model, observation-driven decision-making can still lead to unstable long-horizon behavior, including repeated revisits, local detours, and unreliable stopping. In contrast, SGC-Nav makes exploration history queryable and traceable via the structured scene graph memory SGM_t that explicitly encodes connectivity, revisit traces, and absence cues, and further stabilizes direction selection and stopping through loop-aware multi-model consensus, yielding consistent gains in both success and efficiency.

Compared with Group ID=II methods that leverage stronger priors (mapping or depth), SGC-Nav still attains a 54.4% success rate under no explicit mapping and RGB-only inputs, outperforming or remaining competitive with several mapping-based open-vocabulary methods (e.g., higher SR than SG-Nav at 49.6%). Meanwhile, a gap remains to methods with stronger geometric priors (e.g., InstructNav with 58.0% SR and VLFM* with 56.9% SR / 27.5% SPL), which is consistent with the advantages brought by depth-based mapping and globally consistent planning. Moreover, the ID=I method VLFM reaches 63.6% SR / 32.5% SPL under stronger priors in a closed-vocabulary setting, whereas SGC-Nav attains a competitive SR and comparable efficiency despite operating under strict RGB-only and mapless constraints. Overall, these cross-group comparisons suggest that structured memory and consensus-based decision-making can substantially narrow the performance gap to strong-prior methods even without explicit depth sensing or mapping.

TABLE II: Ablation on three-level scene graphs (SGG/SGL/SGM) on HM3D validation split. Each row corresponds to a configuration indicated by ✓/×.

SGG	SGL	SGM	SR(%)	SPL(%)
✓	×	✓	32.0	13.1
✓	✓	×	36.0	15.4
×	✓	✓	50.6	19.5
✓	✓	✓	**54.4**	**24.6**

TABLE III: Ablation on decision policy: multi-LLM consensus vs. single-LLM on HM3D validation split.

Decision policy	SR(%)	SPL(%)
Single DeepSeek-V3.2 [23]	44.0	18.7
Single Qwen-Max [22]	44.0	22.9
Multi-LLM (Ours)	**54.4**	**24.6**

Ablation Study on Scene Graphs. To verify the role of the three-level scene graphs in decision-making, we keep the decision module unchanged and ablate the scene graph inputs. Tab. II reports the SR and SPL results after removing the SGG, SGL, and SGM components, respectively.

Removing any level of scene graph leads to a performance drop compared to the full model, indicating that each structured representation contributes to navigation. Among them, removing SGL causes the most pronounced degradation (SR drops to 32.0%, SPL drops to 13.1%), suggesting that the per-view action-centric cues in SGL_t (especially traversable exit evidence and coarse goal localization) are critical for stable direction selection and downstream stopping decisions. Removing SGM also yields a substantial drop (SR 36.0%, SPL 15.4%), showing that the queryable long-horizon memory—place merging/assignment, connectivity, revisit traces, and absence cues—plays an important role in suppressing repeated revisits and stabilizing long-horizon exploration. In contrast, removing SGG results in a smaller decrease (SR 50.6%, SPL 19.5%), implying that when SGL_t and SGM_t are preserved, the robot can rely primarily on local action-relevant structure and topological memory for decision-making, while panoramic semantics still provide additional global anchoring that improves robustness and efficiency.

Overall, the ablation results validate the complementarity of the three-level scene graphs for stable long-horizon exploration: SGL_t and SGM_t are key to mitigating deadlocks and improving reliability, while SGG_t provides global semantic support that further improves overall robustness and navigation efficiency.

Ablation Study on Decision Module. To validate the effectiveness of the multi-LLM consensus mechanism in the decision module, we keep all other components unchanged

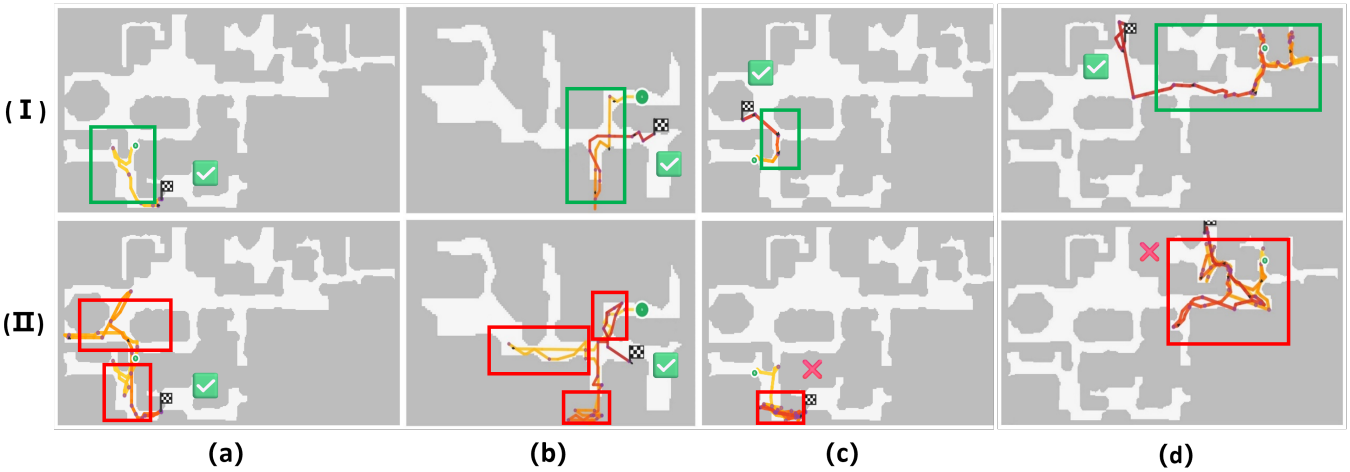

Fig. 4: Qualitative trajectory comparisons on HM3D ObjectNav under strict mapless, RGB-only navigation. (I) Full SGC-Nav with structured SGM yields more direct exploration and reliable goal reaching; (II) replacing SGM with text-only history incurs repeated revisits/oscillation and may fail within the action budget. (a–d) show four representative episodes for comparison between (I) and (II).

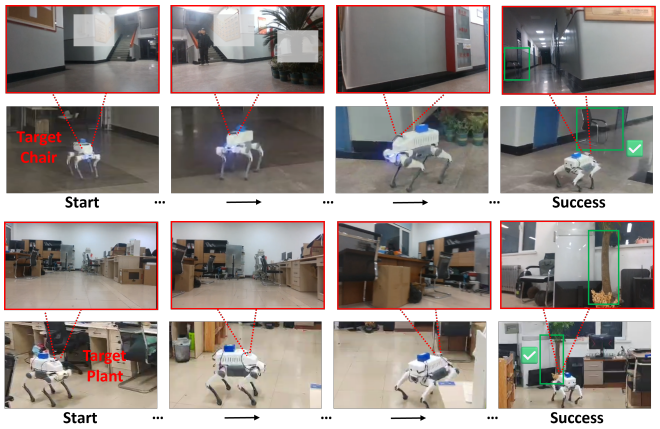

Fig. 5: Real-robot qualitative results of SGC-Nav in indoor environments. The top row shows a corridor scene and the bottom row shows an office scene. The red boxed insets present the corresponding egocentric views.

and replace the three-model consensus with single-model decision-making. Tab. III reports the resulting SR and SPL. The results show that multi-LLM consensus substantially improves navigation reliability: compared to single-model decision-making (SR 44.0%), the three-model consensus increases SR to 54.4%, and improves SPL to 24.6% (from 18.7% with DeepSeek and 22.9% with Qwen-Max). Single-model decision-making is also sensitive to the choice of model: although DeepSeek and Qwen-Max achieve the same SR, their SPL differs noticeably, indicating that a single model is more prone to local misjudgments and bias-driven detours that degrade efficiency. In contrast, multi-LLM consensus treats cross-model disagreement as an explicit uncertainty signal and performs a limited number of rounds of evidence-level mutual critique and aggregation under the same structured context, which suppresses occasional misjudgments and stabilizes both direction selection and stopping, yielding consistent gains in both SR and SPL.

To more intuitively illustrate how SGM affects loop detection in decision-making, and to highlight the structural advantages of SGM itself, we replace SGM with a purely text-based history that describes previously observed scenes. This replacement disables the SGM-derived loop cue and the associated loop-aware pruning behavior. As shown in Fig. 4, (I) corresponds to the full system and (II) corresponds to the ablated variant, and we present representative episodes for visualization.

From the qualitative analysis of the trajectories, the text-memory variant repeatedly revisits previously explored corridors and probes side branches, resulting in a clear increase in redundant steps; it can also exhibit pronounced back-and-forth circling within a single room, indicating an inability to consistently commit to an exit direction (shown as the red boxes in Fig. 4 (a) and (b)). In harder cases, the text-memory variant may remain trapped in a local region for an extended period, repeatedly exploring without expanding its coverage, and ultimately fails to reach the target within the action budget (shown as the red boxes in Fig. 4 (c) and (d)). In contrast, the full system continues to progress into new areas by leveraging queryable topological states, revisit traces, and a consistency preference derived from structured SGM, which suppresses repeated revisits and enables the robot to escape local loops more quickly (shown as the green boxes in Fig. 4). These observations suggest that injecting history as natural-language alone cannot explicitly represent revisit frequency, absence cues, or topological connectivity, and therefore fails to reliably trigger loop suppression and pruning over explored regions.

B. Real-Robot Experiments

We further validate SGC-Nav on a real quadruped robot in indoor environments under strict mapless, RGB-only navigation. The robot executes a discrete action set {`Stop`, `MoveAhead`, `TurnLeft`, `TurnRight`}, where each forward motion advances 0.25 m and each left/right rotation

turns 30°. The robot is equipped with an Intel RealSense D435i camera to stream RGB images for online decision-making.

We conduct real-world goal-driven navigation trials in two typical indoor settings: an office scene and a corridor scene. In each trial, an open-vocabulary natural-language target category is given, and the robot is randomly initialized at a start location. We run 20 trials in total with four target categories: *monitor*, *chair*, *stairs*, and *plant*. A trial is considered successful if the robot stops at the target location and the target category is confirmed by human inspection. Fig. 5 shows two representative successful runs, including searching for a *plant* in the office scene and finding a *chair* in the corridor scene. Overall, SGC-Nav achieves a 45% success rate in real-world trials, demonstrating the feasibility of structured memory and consensus-based decision-making for RGB-only deployment on a physical robot.

V. CONCLUSIONS

In this paper, we study zero-shot open-vocabulary ObjectNav under strict mapless, RGB-only constraints and propose SGC-Nav. SGC-Nav introduces three-level scene graphs (SGG, SGL, SGM) to explicitly encode connectivity, revisit traces, and absence cues as a queryable and traceable long-term history, thereby providing stable support for long-horizon reasoning; it further introduces a loop-aware multi-LLM consensus mechanism that treats inter-model disagreement as an uncertainty signal and performs a limited number of evidence-level critique and aggregation rounds to improve both direction selection and stopping reliability. We validate SGC-Nav on HM3D and in real-robot experiments, demonstrating its effectiveness and stability under strict mapless, RGB-only constraints. Future work will explore more efficient memory update and correction, reduce the inference overhead of multi-model consensus, and conduct long-horizon evaluations in larger and more complex real-world environments.

REFERENCES

[1] Jingwen Sun, Jing Wu, Ze Ji, and Yu-Kun Lai. A survey of object goal navigation. *IEEE Transactions on Automation Science and Engineering*, 22:2292–2308, 2024.

[2] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024.

[3] Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu, Changwei Wang, Man Zhang, Li Guo, and Shibiao Xu. Advances in embodied navigation using large language models: A survey. *arXiv preprint arXiv:2311.00530*, 2023.

[4] 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.

[5] 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.

[6] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE, 2023.

[7] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*, 2024.

[8] Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation. *arXiv preprint arXiv:2506.06487*, 2025.

[9] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307, 2024.

[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] Qunchao Jin, Yilin Wu, and Changhao Chen. Panonav: Mapless zero-shot object navigation with panoramic scene parsing and dynamic memory. *arXiv preprint arXiv:2511.06840*, 2025.

[12] Congcong Wen, Yisiyuan Huang, Hao Huang, Yanjia Huang, Shuaihang Yuan, Yu Hao, Hui Lin, Yu-Shen Liu, and Yi Fang. Zero-shot object navigation with vision-language models reasoning. In *International Conference on Pattern Recognition*, pages 389–404. Springer, 2025.

[13] Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5228–5234. IEEE, 2024.

[14] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[15] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023.

[16] Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, Yi Fang, et al. Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. *Advances in Neural Information Processing Systems*, 37:39386–39408, 2024.

[17] Sixian Zhang, Xinyao Yu, Xinhang Song, Yiyao Wang, and Shuqiang Jiang. Function-centric bayesian network for zero-shot object goal navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19535–19545, 2025.

[18] Mobin Habibpour and Fatemeh Afghah. History-augmented vision-language models for frontier-based zero-shot object navigation. *arXiv preprint arXiv:2506.16623*, 2025.

[19] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2410.09874*, 2024.

[21] Teng Wang, Xinxin Zhao, Wenzhe Cai, and Changyin Sun. Imaginenav++: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2512.17435*, 2025.

[22] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[23] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[24] Mingjie Zhang, Yuheng Du, Chengkai Wu, Jinni Zhou, Zhenchao Qi, Jun Ma, and Boyu Zhou. Apexnav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion. *IEEE Robotics and Automation Letters*, 10(11):11530–11537, 2025.

[25] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment, 2024.