2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

CLUE: Adaptively Prioritized Contextual Cues by Leveraging a Unified Semantic Map for Effective Zero-Shot Object-Goal Navigation

Taeyun Kim[1], Alvin Jinsung Choi[1], Dasol Hong[1] and Hyun Myung[1*], *Senior Member, IEEE*

***Abstract*— Zero-shot object-goal navigation (ZSON) is a challenging problem in robotics that requires a comprehensive understanding of both language and visual observations. Contextual cues from rooms and objects are critical, but their relative importance depends on the target: some objects are strongly tied to specific room types, while others are better predicted by nearby co-located objects. Existing methods overlook this distinction, leading to inefficient and inaccurate exploration. We present CLUE, a novel navigation framework that adaptively balances the use of contextual rooms and objects by leveraging commonsense knowledge extracted from an offline large language model (LLM). By estimating a target's association with room types using LLM, the agent prioritizes room cues for predictable objects and object cues for those with weak room associations. Our framework constructs a unified semantic value map that integrates both types of contextual information, adaptively weighted by the target's ambiguity to guide exploration. Combined with multi-viewpoint verification and an exploration strategy informed by contextual cues, CLUE achieves robust and efficient navigation. Extensive experiments in simulation and real-world deployments show that our method consistently outperforms state-of-the-art baselines in both success rate (SR) and success weighted by path length (SPL), demonstrating its effectiveness and practicality for real-world navigation tasks.**

I. Introduction

Navigating to find a target object in an unseen environment, known as object-goal navigation (ObjectNav) [1], [2], is a fundamental challenge in robotics and embodied AI. Success in this task requires more than accurate object detection. The agent must reason about where objects are likely to appear by leveraging the environmental context. For example, the agent needs to recognize that toilets are usually located in bathrooms, while chairs are often found near tables. This requires linking object categories to their typical spatial and functional contexts, enabling the agent to infer potential object locations and bridge the gap between abstract commands and physical exploration.

Recent advances in foundation models such as large language models (LLMs) [3], [4] and vision-language models (VLMs) [5], [6] have enabled rapid progress in zero-shot ObjectNav (ZSON) [7]–[18]. ZSON offers a more general and efficient setting for real-world deployment, as it removes the need for task-specific training and allows an agent to adapt to new targets and environments on the fly. Some works [7], [8] employed CLIP [5] to align visual inputs with target categories, enabling zero-shot recognition of objects in novel scenes. Subsequent approaches [9]–[11] extended this idea by integrating similarity between the current visual observation and the target category into a physical semantic value map, supporting better spatial reasoning. Other works [13]–[18] leveraged LLMs for higher-level reasoning, using them to interpret accumulated observations and plan subsequent actions. Collectively, these efforts have established a strong foundation for ZSON by integrating powerful vision–language understanding with spatial reasoning.

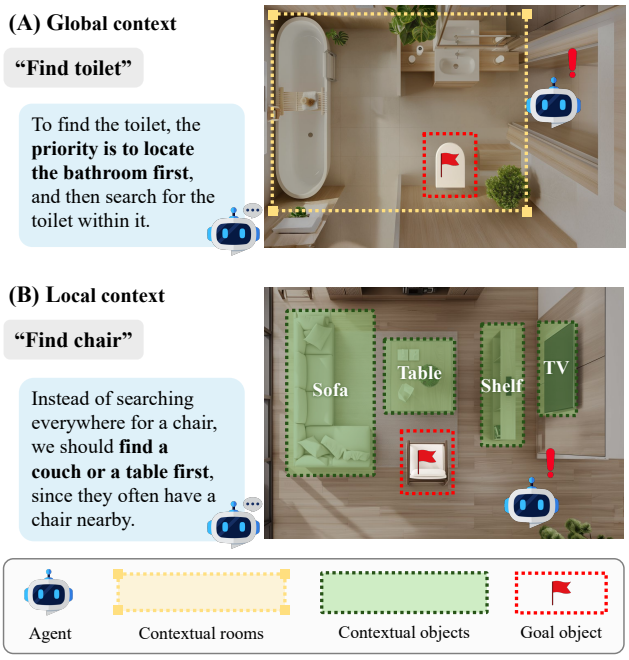

Fig. 1. Illustration of our adaptive strategy for target object search. For objects strongly correlated with particular room types, the agent emphasizes global context using contextual rooms where the target is likely to appear. For objects with low association to specific room types, the agent emphasizes local context using contextual objects that are likely to be co-located with the target.

Despite recent progress, current ZSON methods do not adaptively adjust the relative importance of contextual cues, limiting their effectiveness and accuracy. In many cases, contextual rooms where an object is typically found, and contextual objects it is often co-located with, provide strong semantic cues for the ObjectNav. Although both cues are valuable, their relative importance depends on the target. For example, when searching for a toilet that is strongly associated with a bathroom, prioritizing contextual rooms enables

*Corresponding author: Hyun Myung
All authors are with the School of Electrical Engineering, KAIST (Korea Advanced Institute of Science and Technology), Daejeon, 34141, Republic of Korea. {ktw1404, alvinjinsung, ds.hong, hmyung}@kaist.ac.kr
This work was supported by the Technology Commercialization support program, through the Korea Innovation Foundation funded by the Ministry of Science and ICT. The student is supported by BK21 FOUR.

more efficient navigation. Conversely, when searching for a chair that is not tied to a specific room type, emphasizing contextual objects such as a table leads to reliable navigation. In summary, contextual rooms provide a stronger signal for objects strongly tied to a specific room type, whereas contextual objects should be emphasized for objects whose locations can be inferred from co-located items.

We propose **CLUE** (Adaptively Prioritized **C**ontextual Cues by **L**everaging a **U**nified Semantic Map for **E**ffective Zero-Shot Object-Goal Navigation), a novel navigation framework for ZSON that constructs a unified semantic value map with adaptively prioritized contextual cues. We first quantify the value of each cue. The target object score serves as a base for our semantic value map, which utilizes a vision-language model (VLM) to compute the similarity between the current observation and the target object category. The contextual room score is obtained using a similar technique, with the text prompt changed to predefined room types. The contextual object score is computed from the semantic correlation between the target and contextual objects and is modeled as a Gaussian distribution centered on their locations. These scores are then fused into a unified semantic value map, weighted by the normalized entropy of the target object's room association probability distribution, derived from offline LLM queries. This adaptive weighting allows the semantic value map to reflect the characteristics of the target object by balancing global room-level cues and local object-level cues, enabling more effective navigation. For targets with low entropy on room associations (Fig. 1(a)), the semantic value map prioritizes contextual room scores, highlighting global environmental cues. In contrast, for targets with high entropy (Fig. 1(b)), greater weight is assigned to contextual object scores, emphasizing local spatial relationships.

The agent navigates using this weighted semantic value map by sequentially moving toward the most promising frontier. To ensure robustness, CLUE performs multi-viewpoint observations of each target candidate for robust verification. A further advantage of our framework is its ability to perform higher-level reasoning without relying on costly online LLM queries. Instead, we use offline LLM queries to extract commonsense knowledge about target–room associations, contextual rooms, and co-located objects before execution, which significantly reduces computational overhead and avoids delays during real-world navigation.

We validate our approach through extensive experiments on the HM3D [19] dataset. CLUE achieves state-of-the-art performance, outperforming competitive baselines in both success rate (SR) and success weighted by path length (SPL). We further deploy CLUE on a Clearpath Jackal platform in real-world settings, demonstrating its practicality for onboard navigation.

In summary, our work makes three key contributions:

- We introduce **CLUE**, a framework that constructs a unified semantic value map by balancing contextual room and contextual object cues according to the target's characteristics, enabling more reliable exploration.
- We ensure real-time capability by leveraging offline LLM queries for commonsense knowledge, eliminating the latency and computational overhead of online LLM reasoning.
- We validate the effectiveness of our method through extensive experiments in both simulation benchmark and real-world deployment, achieving state-of-the-art performance compared with state-of-the-art baselines.

II. Related Works

A. Zero-Shot Object-Goal Navigation

With access to internet-scale training data, LLMs [3], [4] and VLMs [5], [6] have demonstrated strong generalization and reasoning capabilities. Recent works [8]–[11] have leveraged these models to tackle object-goal navigation in a zero-shot setting. COWs [8] employs CLIP [20] to align the current visual observation with the target object description, but relies on single-frame signals that lack spatial-temporal context. VLFM [9] constructs a semantic value map by computing CLIP-based similarity between the target object and current observations for better reasoning, but uses the same strategy for all targets without distinguishing whether they are better localized through room cues or co-located objects. Onemap [10] extends VLFM to sequential multi-object navigation by maintaining a memory of prior search locations, and ApexNav [11] builds on VLFM by adapting exploration based on the magnitude of semantic cues. However, both methods also employ an object-agnostic strategy, thereby preventing the full exploitation of contextual information.

While these methods demonstrate the potential for ZSON, they do not fully leverage contextual cues. Most rely only on local information or do not consider the relative importance of cues. In contrast, our approach constructs a unified semantic value map that adaptively balances global room-level and local object-level cues, weighting them according to the relative importance based on the target.

B. LLM-based Reasoning for ObjectNav

Recent works [12]–[15] leverage LLMs for higher-level reasoning by incorporating commonsense knowledge. These methods provide the LLM with structured representations of the scene, transforming raw observations into formats that capture object relationships and spatial context to improve decision-making. ESC [12] queries an LLM to infer which frontiers may lead to the target based on detected objects. SG-Nav [13] constructs a hierarchical 3D scene graph and queries the LLM to select navigation actions. VoroNav [14] encodes a topological map together with semantic information into textual descriptions of the environment and exploratory paths, which are then provided to the LLM to determine actions. OpenFMNav [15] builds a semantic score map from human instructions and employs LLM-based commonsense reasoning to guide exploration. However, these methods rely on online LLM queries, which often introduce computational overhead and redundancy.

Although LLMs provide strong reasoning capabilities through commonsense knowledge, continuously querying

them during execution hinders real-world applicability due to high latency. In contrast to other methods, we leverage LLMs offline to extract the most relevant knowledge about the target object prior to execution. This strategy avoids runtime delays while preserving high-level reasoning ability, enabling efficient navigation without redundant online queries.

III. CLUE: Unified Semantic Value Map for Scene Understanding

CLUE represents the scene with a unified semantic value map that integrates contextual room and object cues, weighted by their relative importance to the target for reliable navigation. Section III.*A* defines the problem, and Section III.*B* introduces the map representation that provides the spatial foundation for expressing contextual cues and guiding navigation. Section III.*C* details how target object cues are quantified as scores, which serve as the basis for the semantic map. Section III.*D* and Section III.*E* explain how contextual rooms and objects are identified and scored. Finally, Section III.*F* describes how these signals are fused into a unified semantic value map with adaptive weighting. The overall framework is illustrated in Fig. 2.

A. Problem Definition

In ObjectNav, an agent is placed in an unknown environment with no prior map and tasked to locate a target object specified by a natural language query (e.g., 'bed' or 'chair'). The agent's perception is limited to an egocentric camera view. A trial is considered successful if the agent stops within the distance of d_s from the target object in at most T steps.

B. Map Representation

We construct a top-down 2D map to represent the environment. Similar to VLFM [9], we employ a 2D value map composed of a geometric map $\mathcal{M}_{\text{geo}}$ and a semantic value map $\mathcal{M}_{\text{sem}}$ to represent the environment.

The geometric map $\mathcal{M}_{\text{geo}}$ is generated by projecting the 3D point cloud from an RGB-D camera or LiDAR into a 2D plane, identifying traversable areas, obstacles, and unexplored regions. Local maps are continuously integrated into a global frame using the robot's pose, classifying regions as explored, unexplored, or occupied. Frontiers are defined as center points along boundaries between explored and unexplored regions to guide exploration.

The semantic value map $\mathcal{M}_{\text{sem}}$ provides a human-like interpretation of visual information, capturing the environment's semantic structure. Each cell encodes the strength of contextual cues, indicating the likelihood of the target's presence and guiding navigation. Values are computed by incorporating not only the target object cues but also contextual room and object cues.

C. Target Object Cues

CLUE aims to locate the target object in an unknown environment by quantifying target object cues as the basis of the semantic value map. Following VLFM [9], we leverage a VLM to compute target object scores. In particular, BLIP2 [6] is used to calculate v_{target} as the cosine similarity between the current view and the target category. This score represents the relevance of the view to the target and is stored in the corresponding pixel of the value map, which is continuously updated as new images are acquired.

Semantic confidence c controls how the target object value of a previously observed pixel is updated when revisited. It is defined as follows:

$$c = \cos^2\left(\frac{\theta}{\theta_{\text{FoV}}/2} \times \frac{\pi}{2}\right), \tag{1}$$

where θ is the angle between the pixel and the optical axis, and θ_{FoV} is the camera's horizontal field of view (FoV). The confidence approaches 1 for pixels near the image center and decreases toward 0 at the edges. This reflects that the cosine similarity is generally higher in central regions. Note that c affects only the update of previously observed pixels, not the assignment of values to newly observed ones.

The target object score is updated based on the confidence score, together with current and previous values. This update is performed through a weighted average of the current and previous values and confidence scores as follows:

$$v_{\text{target}}^{\text{new}} = \frac{c^{\text{cur}} v_{\text{target}}^{\text{cur}} + c^{\text{prev}} v_{\text{target}}^{\text{prev}}}{c^{\text{cur}} + c^{\text{prev}}}, \tag{2}$$

$$c^{\text{new}} = \frac{(c^{\text{cur}})^2 + (c^{\text{prev}})^2}{c^{\text{cur}} + c^{\text{prev}}}, \tag{3}$$

where $v_{\text{target}}^{\text{new}}$ and c^{new} denote the updated target object value and confidence score; $v_{\text{target}}^{\text{cur}}$ and c^{cur} come from the current observation; and $v_{\text{target}}^{\text{prev}}$ and c^{prev} are those stored from the previous step.

D. Contextual Rooms for Global Spatial Understanding

Contextual rooms strongly associated with the target object provide valuable cues for global spatial reasoning. Given a predefined set of common indoor room categories $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_n\}$, we query an LLM to identify the room in which the target object $\mathbf{O}_{\text{target}}$ is most likely to be found as contextual room $\mathbf{R}_i^{\text{C}}$. Our goal is to identify a contextual room and leverage it to guide navigation toward the target object.

To this end, we compute the contextual room score v_{room} in the same manner as the target object score v_{target}, but replace the VLM text prompt with the identified contextual room category. Thus, v_{room} reflects the relevance of the current view to the most relevant room, emphasizing global spatial cues. The confidence measure for each pixel is defined in the same way as for v_{target}, and the value map is updated using the same weighted update rule.

E. Contextual Objects for Local Spatial Understanding

Contextual objects that are often co-located with the target object serve as meaningful cues for local spatial reasoning. To identify them, we first query the LLM for a set of contextual objects $\mathcal{O}^{\text{C}} = \{\mathbf{O}_1^{\text{C}}, \ldots, \mathbf{O}_m^{\text{C}}\}$ that are semantically related to the target object $\mathbf{O}_{\text{target}}$. We then

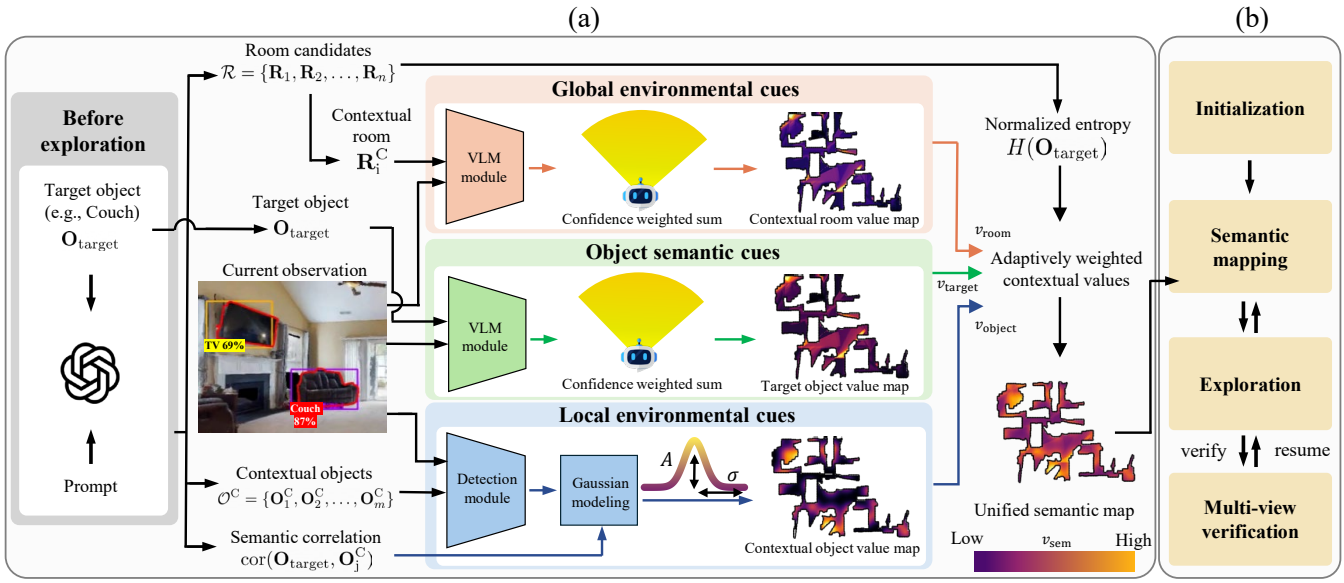

Fig. 2. Overview of CLUE. (a) Construction of a unified semantic value map by adaptively balancing contextual cues according to the target's characteristics. Prior to execution, commonsense knowledge is extracted from an LLM, including contextual rooms with their associated probabilities and contextual objects with their semantic relevance to the target. This information is used to build contextual room and contextual object value maps, which are then fused adaptively using entropy-based weighting to capture global and local environmental cues. (b) Utilization of the unified semantic value map for exploration through a coarse-to-fine strategy, where semantically relevant regions are prioritized and multi-view verification ensures efficient and accurate navigation.

compute the semantic correlation $\text{cor}(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C)$ between the target object and each contextual object, normalizing the scale so that the target's self-correlation is equal to 1.0. The correlations with contextual objects are precomputed by the LLM before the task begins. During exploration, these values are used by the agent to assign contextual object scores when such objects are detected by the detection module.

We employ a detection module to recognize contextual objects and compute the contextual object score v_{object}, modeled as a Gaussian distribution. This score reflects the likelihood of the target object's presence, being higher near the detected contextual objects and decaying smoothly with distance. For each detection, the module records the image pose, the center of the object point cloud cluster (p_x, p_y), the object class, and the detector's confidence score as an object node. The object cluster center is then projected using the point cloud to obtain $(\bar{x}, \bar{y})$ coordinates on the 2D semantic value map $\mathcal{M}_{\text{sem}}$. At this location, the precomputed correlation score, $\text{cor}(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C)$, is utilized to score the pixel and the surrounding area. The contextual object score v_{object} at 2D map location (x, y) is defined as follows:

$$v_{\text{object}}(x, y) = A(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C) \cdot \exp\left(-\frac{(x-\bar{x})^2 + (y-\bar{y})^2}{2\sigma(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C)^2}\right), \quad (4)$$

where $A(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C)$ is the amplitude that reflects the importance of the contextual object, and $\sigma(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C)$ controls the spatial spread of the contextual influence around the detected object's position $(\bar{x}, \bar{y})$. Formally, it is defined as follows:

$$A(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C) = A_0 \cdot \text{cor}(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C), \quad (5)$$

$$\sigma(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C) = \sigma_0 + \text{cor}(\mathbf{O}_{\text{target}}, \mathbf{O}_j^C), \quad (6)$$

where A_0 and σ_0, denote the base amplitude and the base standard deviation, respectively. In this design, stronger semantic correlation not only increases the weight of contextual objects through a larger amplitude but also broadens the spatial influence via a wider spread, enabling the agent to leverage co-located objects more effectively.

F. Fusing Contextual Cues for Unified Semantic Value Map

We fuse target object scores along with contextual room and object scores into a unified semantic value map, adaptively weighting cues according to the target's characteristics. The relative importance is determined by the association of the target object with specific room types. To estimate this association, we query an LLM to obtain the probability $P(\mathbf{R}_k|\mathbf{O}_{\text{target}})$ that the target object is likely to exist in the room type $\mathbf{R}_k$. From this probability distribution, we then compute the information entropy as a measure of the uncertainty in the object's candidate room location as follows:

$$H(\mathbf{O}_{\text{target}}) = \frac{-\sum_{k=1}^{n} P(\mathbf{R}_k|\mathbf{O}_{\text{target}}) \log(P(\mathbf{R}_k|\mathbf{O}_{\text{target}}))}{\log(n)}. \quad (7)$$

We provide the agent with a comprehensive semantic representation to guide the search for the target object $\mathbf{O}_{\text{target}}$. The target object score v_{room}, the contextual room score v_{room}, and the contextual object score v_{object} are integrated into the semantic value map $\mathcal{M}_{\text{sem}}$, with each cell value given by:

$$v_{\text{sem}} = v_{\text{target}} + \omega_{\text{room}} \cdot v_{\text{room}} + \omega_{\text{object}} \cdot v_{\text{object}}, \quad (8)$$

where the weights ω_{room} and ω_{object} are determined by the

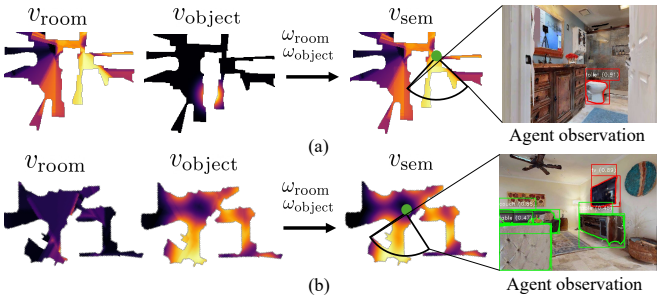

Fig. 3. Illustration of the agent's current observation and its corresponding representation on the individual value maps (v_{room}, v_{object}) and the final unified semantic value map v_{sem}. (a) An example of a low-entropy object (toilet), where contextual rooms provide distinctive guidance while contextual objects do not. (b) An example of a high-entropy object (TV), where the unified map is more strongly influenced by local contextual objects due to the lack of distinctive spatial evidence from contextual rooms.

entropy $H(\mathbf{O}_{\text{target}})$. It is formally defined as follows:

$$\begin{aligned} \omega_{\text{room}} &= 1 - H(\mathbf{O}_{\text{target}}) \\ \omega_{\text{object}} &= H(\mathbf{O}_{\text{target}}) \ . \end{aligned} \tag{9}$$

Through this adaptive modeling, low-entropy target objects that are strongly associated with a specific room category prioritize contextual room cues over contextual object cues, emphasizing global environmental signals. In contrast, for high-entropy target objects that may appear across diverse room types, we emphasize contextual object cues as they provide the most reliable and discriminative information for guiding navigation, highlighting local environmental signals. As shown in Fig. 3, the semantic value map captures the spatial properties most relevant to the target object and effectively guides the agent's navigation.

IV. CLUE: Exploration Strategy

A. Initialization

At the start of navigation, an agent with a limited FoV has insufficient semantic cues to reliably guide its movement. To mitigate this, the agent first performs an in-place rotation to capture the maximum possible observations from its initial position. Once it has accumulated sufficient semantic and geometric understanding of the surroundings, the agent selects the most promising frontier as its destination and begins moving toward it.

B. Exploration Using Fused Semantic Value Map

We adopt frontier-based exploration as our navigation policy. The agent selects the frontier it considers the most likely to lead to the target object. To improve efficiency, we incorporate the strategy proposed in ApexNav [11]. We first determine if sufficient semantic cues have been accumulated, indicated by checking whether any frontier score on the unified semantic value map $\mathcal{M}_{\text{sem}}$ exceeds a predefined threshold. Before sufficient semantic information is provided, the agent performs geometry-based exploration by solving a traveling salesman problem (TSP) to plan its path. After the condition is met, the agent switches to selecting the frontier with the highest semantic score as its navigation goal. This results in a conditional exploration strategy in which the agent performs rapid geometry-driven exploration when semantic information is sparse, then transitions to semantic-driven exploration once reliable cues become available.

C. Multi-View Verification

Once a potential target is identified, the agent enters a verification phase. Standard single-image detection can be prone to occlusion and false positives, so our method navigates to multiple viewpoints around the candidate. The detection results from these viewpoints are aggregated, taking into account both the number of consistent detections and their class agreement. Based on this evidence, the system determines whether the candidate is a true target or a false positive. If deemed a false positive, the agent discards the candidate and resumes the search.

V. Experimental Setting for Evaluation

A. Dataset

We evaluate CLUE by comparing our proposed method with state-of-the-art methods on the ObjectNav task in the Habitat simulator [21] using the HM3D dataset [19]. The HM3D dataset contains 2,000 episodes across 20 scenes with 6 types of target objects. It features diverse, building-scale environments, making it suitable for demonstrating the effectiveness and robustness of the proposed framework.

B. Baselines

We categorize the baselines based on how they enable the agent to interpret unknown environments. End-to-end methods jointly learn environment representations and ObjectNav policies, including ZSON [7], SemExp [22], and PONI [23]. Although ZSON is often regarded as zero-shot, it still requires ImageNav training. VLM-based methods [8], [9], [11] compare the target with the current observation, with VLFM [9] and ApexNav [11] notable for its value map formulation similar to ours. LLM-based methods such as SG-Nav [13] and VoroNav [14] build structured scene representations and query an LLM for action decisions. We also include other LLM-driven approaches [12], [15]–[18] that model complex environmental relationships. All baselines are evaluated in the same environments using both success rate and efficiency.

C. Metrics

We use two standard metrics [24] to evaluate ObjectNav performance: success rate (SR) and success weighted by path length (SPL). SR is the proportion of episodes in which the agent successfully reaches the target object. SPL measures navigation efficiency by comparing the agent's actual path length to the shortest possible path length to the target.

D. Implementation Details

For each episode, the maximum number of allowed steps is set to $T = 500$, with a success distance threshold of $d_s = 0.2$m. The agent is equipped with an RGB-D sensor providing images at a resolution of 640×480, with depth values ranging from 0.5m to 5.0m. The camera is mounted

TABLE I. Performance comparison of diverse methods on the HM3D benchmark. Bold indicates the best performance, and gray highlighting indicates the second-best. (a) End-to-end methods. (b) Zero-shot ObjectNav methods.

	Method	LLM query	HM3D	
			SR↑	SPL↑
(a)	ZSON [7]	-	25.5	12.6
(b)	VLFM [9]	-	52.5	30.4
	ApexNav [11]	Offline	59.6	33.0
	ESC [12]	Online	39.2	22.3
	VoroNav [14]	Online	45.0	26.0
	TopV-Nav [16]	Online	45.9	28.0
	L3MVN [17]	Online	50.4	23.1
	OpenFMNav [15]	Online	54.9	24.4
	SG-Nav [13]	Online	54.0	24.9
	TriHelper [18]	Online	56.5	25.3
	CLUE (Ours)	Offline	**61.7**	**34.3**

TABLE II. Ablation study of CLUE modules on the HM3D dataset. We analyze the effect of incorporating contextual rooms and contextual objects, along with adaptive weighting, measured in terms of SR and SPL. Bold indicates the best performance.

Contextual rooms	Contextual objects	Adaptive weighting	SR↑	SPL↑
			59.6	33.0
✓			59.8	33.2
	✓		60.2	31.8
✓	✓		61.2	33.8
✓	✓	✓	**61.7**	**34.3**

TABLE III. Ablation study of CLUE adaptive weighting strategy on the HM3D dataset. We analyze the effect of varying weights ω_{room} and ω_{object} on objects with different entropy levels $H(\mathbf{O}_{\text{target}})$, measured in terms of SR. Bold indicates the best performance.

$\mathbf{O}_{\text{target}}$ (Episode)	$H(\mathbf{O}_{\text{target}})$	SR by (ω_{room}, ω_{object})			
		CLUE	(1.0,0.0)	(0.5,0.5)	(0.0,1.0)
Toilet (398)	0.043	**72.9**	72.6	63.8	72.4
Bed (433)	0.124	57.7	57.2	**59.1**	57.7
Couch (376)	0.203	**61.7**	60.6	**61.7**	61.4
TV (281)	0.462	39.8	37.4	**40.2**	38.4
Chair (428)	0.883	**73.3**	69.0	73.1	72.7
Potted Plant (84)	0.915	**42.9**	38.1	41.7	**42.9**
Total (2000)	-	**61.7**	59.8	60.2	61.2

0.88m above the ground. The agent performs discrete actions consisting of moving forward by 0.2m or rotating by 30°. For the VLM used to compute the contextual room score, we employ the BLIP2 [6]. The Gemini-Pro 2.5 model [25] is used to obtain the semantic entropy $H(\mathbf{O}_{\text{target}})$, identify contextual objects, and compute $\text{cor}(\mathbf{O}_{\text{target}}, \mathbf{O}_j^{\text{C}})$. For object detection, we use YOLOv7 [26] for COCO [27] classes and GroundingDINO [28] for non-COCO classes. For segmentation, we use MobileSAM [29]. All experiments are conducted on RTX 3090 GPUs.

For real-world experiments, we deploy CLUE on the Clearpath Jackal platform with a Velodyne VLP-16 LiDAR and a Ricoh Theta Z1. To process the omnidirectional sensor data within our framework, the observation area is limited to 90° quadrants. Due to the sparsity of the point cloud, TRIP [30] and TRG-planner [31] were used to extract ground planes and plan path for safe navigation.

VI. Experiment Results

A. Overall Performance Evaluation

Table I reports the overall performance on the HM3D validation set. CLUE achieves state-of-the-art results in both SR and SPL, demonstrating strong accuracy and efficiency. In particular, it surpasses ApexNav, a top-performing method that does not rely on online LLM queries, on both metrics and shows especially large gains of 3.52% increase in SR. ApexNav neither incorporates contextual room cues nor adaptively emphasizes different types of cues based on the characteristics of the target, causing the agent to overlook critical semantic cues and make inaccurate and inefficient movements during exploration. In contrast, our approach adaptively weights different types of contextual cues, prioritizing the most informative semantic signals for each target and enabling accurate, robust navigation. CLUE also outperforms methods that employ online LLM queries, most notably in SPL, demonstrating robust reasoning while maintaining efficient navigation. By extracting the most relevant contextual information about the target using commonsense knowledge prior to execution, our method enables the agent to perform high-level reasoning and actions efficiently and consistently.

These results confirm that our approach effectively unifies contextual global-level and local-level cues with offline commonsense knowledge, enabling a state-of-the-art ZSON that is both accurate and practical for real-world deployment.

B. Ablation Study

We demonstrate the effectiveness of our proposed method through an ablation study on the HM3D dataset, as shown in Table II. Without contextual rooms and contextual objects, the agent struggles to gather sufficient semantic information and visual evidence, resulting in the lowest performance on both metrics. Each contextual information enables the agent to accumulate richer semantic information for more reliable guidance. Adaptive weighting further enhances the reliability of the resulting unified semantic map by prioritizing the most relevant semantic cues based on the target object. These results confirm that our proposed method improves the agent's ability to locate the target object.

We further analyze the effectiveness of our adaptive weighting strategy. Specifically, we replace adaptive weighting based on target–room association entropy $H(\mathbf{O}_{\text{target}})$ with manually defined fixed weights for the contextual room and contextual object scores. Assigning a higher weight to the room score forces the agent to prioritize global semantic cues regardless of the target, whereas a higher weight on the object score biases it toward local semantic cues. As shown in Table III, our adaptive weighting framework achieves the highest overall success rate across different types of target objects, demonstrating its effectiveness. For low-entropy objects with strong spatial associations, such as a toilet

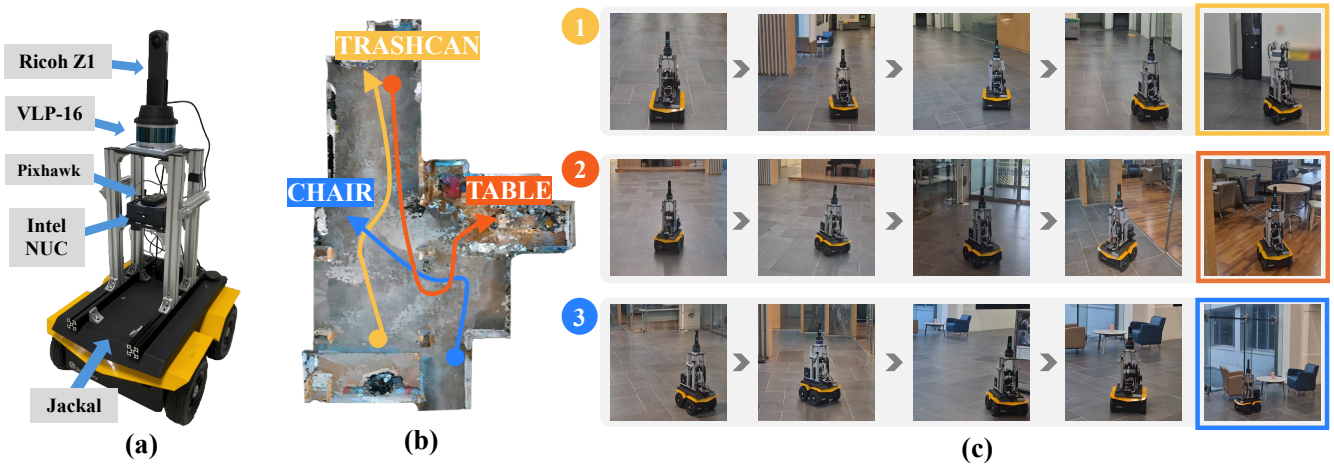

Fig. 4. Configuration for real-world experiments. (a) A customized UGV platform based on a Clearpath Jackal, equipped with an Intel NUC, a Velodyne VLP-16 LiDAR, and a Ricoh Theta Z1 camera. (b) Real-world environments along with the trajectories the robot traveled to search for various objects. (c) The robot leverages contextual cues to move to likely object locations, then verifies the findings using multi-view verification.

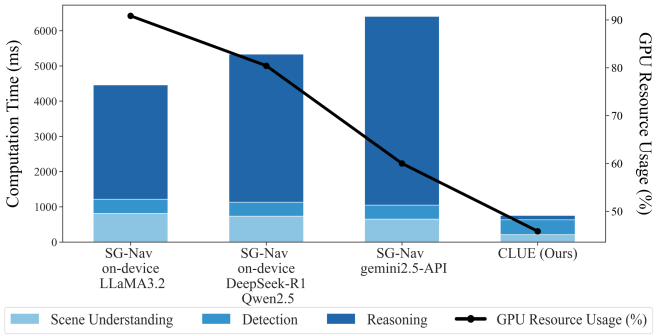

Fig. 5. Resource usage analysis: comparison of computation time (ms) and GPU resource usage (%) across different ObjectNav frameworks. These metrics are critical for evaluating performance on resource-constrained on-board systems.

($H(\mathbf{O}_{\text{target}}) = 0.043$), the room-only model performs well, while the object-only model suffers performance degradation. In contrast, for high-entropy objects that can appear in multiple locations, such as a chair ($H(\mathbf{O}_{\text{target}}) = 0.883$) and potted plant ($H(\mathbf{O}_{\text{target}}) = 0.915$), relying solely on room cues leads to inefficient exploration, whereas object cues provide stronger guidance. The model that combines both cues with equal weights still underperforms compared with CLUE, confirming that non-adaptive integration is suboptimal.

These results demonstrate that CLUE's adaptive strategy of dynamically adjusting the importance of global and local cues according to the object's characteristics is more effective than fixed weighting. This validates the core idea of our work, showing that optimizing the exploration strategy according to object entropy plays a decisive role.

C. Resource Analysis

We analyze the time consumption and GPU resource usage of CLUE in comparison with competing baselines. In particular, we evaluate SG-Nav, which relies on online LLM queries with chain-of-thought reasoning and incremental prompting, techniques proposed to mitigate the high time complexity of LLM queries. We consider three variants of SG-Nav, distinguished by their model configurations: one with a single on-device LLaMA3.2-vision, a second with separate on-device DeepSeek-R1(LLM) and Qwen2.5-VL(VLM) modules, and a third that leverages a gemini2.5-API LLM. To assess time consumption, we measure the per-step runtime for scene understanding, detection, and reasoning. For GPU usage, we report the peak memory required to execute each algorithm.

As shown in Fig. 5, the on-device LLaMA3.2 model variant of SG-Nav consumes the most GPU resources, whereas on-device variants requires the most time for both scene understanding and decision-making. In contrast, CLUE is considerably more efficient than SG-Nav, demanding fewer resources and delivering faster performance. It is also important to note that simulation benchmarks typically measure performance in terms of steps, often overlooking the actual runtime needed to complete an episode. This experiment demonstrates that CLUE is well suited for navigation tasks on resource-constrained on-board systems.

D. Real-World Experiment

To validate CLUE in real-world settings, we conducted physical robot experiments. The robot configuration is shown in Fig. 4(a), and the test environment containing various objects is illustrated in Fig. 4(b). The system was evaluated using only text-based queries for different target objects. As shown in Fig. 4(c), our method successfully and consistently locates the specified objects, confirming its effectiveness in real-world scenarios. Notably, the system successfully identified targets even when they were placed in atypical locations, enabled by our unified semantic value map's comprehensive modeling of contextual cues and diverse correlations. In addition, multi-view verification further enhanced the robustness and reliability of object localization.

VII. Conclusion

We presented CLUE, an adaptive framework for ZSON that prioritizes different types of contextual cues to enable effective and robust navigation. Unlike prior methods that apply uniform strategies or rely on costly online LLM queries, our approach adaptively prioritizes contextual cues and leverages offline commonsense knowledge to construct a unified semantic value map. This map integrates both contextual room and object cues, with their relative importance adaptively weighted by the target's characteristics. Through this design, the agent balances global and local cues to infer likely object locations. Extensive experiments on the HM3D benchmark, along with real-world deployment on a Clearpath Jackal, show that CLUE achieves superior accuracy and efficiency compared with state-of-the-art baselines. While effective, our framework assumes predefined room categories and relies on precomputed commonsense knowledge, which may limit generalization. Future work will extend the approach to an open-set setting and incorporate online adaptation for greater robustness in diverse scenarios.

References

[1] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "ObjectNav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[2] J. Sun, J. Wu, Z. Ji, and Y.-K. Lai, "A survey of object goal navigation," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 2292–2308, 2024.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learning*, 2021, pp. 8748–8763.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learning*, 2023, pp. 19 730–19 742.

[7] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "ZSON: Zero-shot object-goal navigation using multimodal goal embeddings," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 32 340–32 352, 2022.

[8] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "CoWs on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23 171–23 181.

[9] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "VLFM: Vision-language frontier maps for zero-shot semantic navigation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 42–48.

[10] F. L. Busch, T. Homberger, J. Ortega-Peimbert, Q. Yang, and O. Andersson, "One map to find them all: Real-time open-vo1cabulary mapping for zero-shot multi-object navigation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 14 835–14 842.

[11] M. Zhang, Y. Du, C. Wu, J. Zhou, Z. Qi, J. Ma, and B. Zhou, "ApexNav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion," *IEEE Robot. Automat. Lett.*, 2025.

[12] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, "ESC: Exploration with soft commonsense constraints for zero-shot object navigation," in *Proc. Int. Conf. Mach. Learning*, 2023, pp. 42 829–42 842.

[13] H. Yin, X. Xu, Z. Wu, J. Zhou, and J. Lu, "SG-Nav: Online 3D scene graph prompting for LLM-based zero-shot object navigation," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 5285–5307, 2024.

[14] P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang, and C. Liu, "VoroNav: Voronoi-based zero-shot object navigation with large language model," in *Proc. Int. Conf. Mach. Learning*, 2024, pp. 53 737–53 775.

[15] Y. Kuang, H. Lin, and M. Jiang, "OpenFMNav: Towards open-set zero-shot object navigation via vision-language foundation models," *arXiv preprint arXiv:2402.10670*, 2024.

[16] L. Zhong, C. Gao, Z. Ding, Y. Liao, H. Ma, S. Zhang, X. Zhou, and S. Liu, "TopV-Nav: Unlocking the top-view spatial reasoning potential of MLLM for zero-shot object navigation," *arXiv preprint arXiv:2411.16425*, 2024.

[17] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging large language models for visual target navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2023, pp. 3554–3560.

[18] L. Zhang, Q. Zhang, H. Wang, E. Xiao, Z. Jiang, H. Chen, and R. Xu, "TriHelper: Zero-shot object navigation with dynamic assistance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2024, pp. 10 035–10 042.

[19] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and B. Chruv, "Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI," in *Adv. Neural Inf. Process. Syst. Datasets and Benchmarks Track (Round 2)*, 2021.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learning*, 2021, pp. 8748–8763.

[21] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied AI research," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347.

[22] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 4247–4258, 2020.

[23] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "PONI: Potential functions for ObjectGoal navigation with interaction-free learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18 890–18 900.

[24] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.

[25] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[26] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[28] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 38–55.

[29] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight SAM for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.

[30] M. Oh, B. Yu, I. Nahrendra, S. Jang, H. Lee, D. Lee, S. Lee, Y. Kim, M. K. Christiansen, H. Lim, and H. Myung, "TRIP: Terrain traversability mapping with risk-aware prediction for enhanced online quadrupedal robot navigation," *arXiv preprint arXiv:2411.17134*, 2024.

[31] D. Lee, I. M. A. Nahrendra, M. Oh, B. Yu, and H. Myung, "TRG-Planner: Traversal risk graph-based path planning in unstructured environments for safe and efficient navigation," *IEEE Robot. Automat. Lett.*, vol. 10, no. 2, pp. 1736–1743, 2025.