CL-HCoTNav: Closed-Loop Hierarchical Chain-of-Thought for Zero-Shot Object-Goal Navigation with Vision-Language Models

Yuxin Cai*†, Haoruo Zhang*, Wei-Yun Yau†, Chen Lyu*
*Nanyang Technological University, Singapore
†Institute for Infocomm Research, A*STAR, Singapore

***Abstract*—Visual Object Goal Navigation (ObjectNav) requires a robot to locate and navigate to a target object using egocentric observations. However, generalizing policy behavior to new settings—unseen environments and novel target objects—remains a significant challenge. Traditional end-to-end learning methods exacerbate this issue, relying on memorized latent patterns rather than structured reasoning, which limits their ability to generalize effectively. While some recent approaches leverage foundation models for enhanced reasoning, they often overlook the inherent uncertainty and potential errors in vision-language model (VLM) outputs, lacking mechanisms to detect and correct mistakes. In this work, we introduce Closed-Loop Hierarchical Chain-of-Thought Navigation (CL-HCoTNav), a VLM-driven ObjectNav framework that integrates structured reasoning and closed-loop feedback into navigation policy learning. We fine-tune VLM using multi-turn question-answering (QA) pairs derived from human demonstration trajectories. This structured dataset enables hierarchical Chain-of-Thought (H-CoT) prompting, extracting compositional knowledge following the human cognitive process of locating a target object in iterative reasoning steps. In addition, we propose a Closed-Loop H-CoT mechanism that incorporates quantifiable detection and reasoning confidence scores into training loop. Our adaptive weighting strategy guides the model to prioritize high-confidence data pairs during navigation, reducing noise from observations and improving robustness against hallucinated or incorrect reasoning. Extensive experiments in the AI Habitat demonstrate that CL-HCoTNav achieves superior generalization to unseen scenes and novel object categories, outperforming state-of-the-art approaches in ObjectNav success rate (SR) and success weighted by path length (SPL) by 22.4%.**

I. Introduction

Humans efficiently navigate unfamiliar environments to find target objects by reasoning about semantic relationships [29]—for instance, recognizing that kitchens are typically adjacent to living rooms, or that exit signs indicate pathways to an exit. This structured reasoning enables humans to infer the probable locations of objects without exhaustive exploration or explicit SLAM-based mapping. Replicating this capability in robots is fundamental to ObjectNav task, where a robot must locate a specified object category in an unseen environment using only egocentric observations [25]. However, achieving reasoning-driven navigation remains an open challenge, particularly under zero-shot generalization settings, where a robot must navigate to previously unseen object categories or adapt to novel scene layouts without retraining. This makes Zero-Shot Object Navigation (ZSON) a critical yet largely unsolved problem [18], as illustrated in Fig. 1.

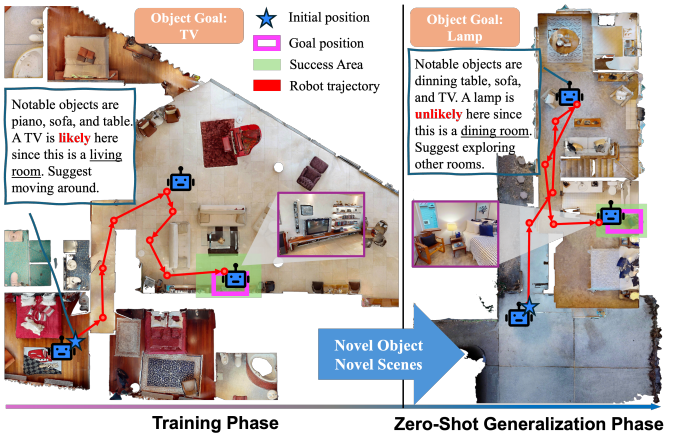

Fig. 1. Overview of the Zero-Shot Object Navigation (ZSON) problem. The left side illustrates the training phase, where a robot learns to navigate to target objects within a set of seen scenes and object categories. The right side depicts the zero-shot generalization phase, where the learned policy is evaluated in novel scenes and with unseen target objects without additional training.

The ZSON problem presents two primary challenges [34]. First, unseen object generalization requires a robot to infer the probable location of novel target objects based on learned semantic relationships, rather than memorizing specific object instances. For example, if a robot learned to find a "chair", it should infer the likely placement of a "stool" based on shared spatial and functional attributes, without requiring explicit training. Second, unseen scene generalization demands that a robot recognize high-level layout patterns rather than memorizing room arrangements. While real-world indoor spaces exhibit variations in layouts, they often share similar organizational structures. Effective generalization thus requires reasoning about spatial relationships—for instance, recognizing that a kitchen is typically adjacent to a dining area—rather than relying on rigid spatial distributions encountered during training. Addressing these challenges necessitates structured semantic reasoning beyond conventional feature-based learning. In this work, we explicitly model object-object and object-scene co-occurrence relationships to generalize across unseen objects and scenes while maintaining efficient navigation.

Existing approaches to ZSON based on end-to-end reinforcement learning (RL) directly optimize navigation policies

from egocentric visual inputs [21]. While capable of capturing complex visual-action associations, these methods require extensive interaction data for training and often struggle to generalize beyond the training distribution. Critically, they lack principled mechanisms for integrating semantic reasoning and uncertainty estimation into the navigation loop, limiting their ability to generalize to novel objects and environments. Recent advances in VLMs offer new opportunities to inject commonsense reasoning into navigation [23, 31]. These models excel at object recognition and relational reasoning, making them valuable for introducing semantic priors and inferring object-scene relationships [32, 27]. However, directly applying VLMs to ObjectNav remains challenging [12]. While language models can generate high-level narratives about navigation strategies, they often lack grounding in real-world sensory observations and may produce hallucinated or semantically incorrect reasoning. Moreover, their lack structured intermediate reasoning steps or fail to account for incorrect or uncertain reasoning, leading to unreliable performance.

To address these limitations, we propose a VLM-driven navigation framework that introduces a hierarchical chain-of-thought (H-CoT) reasoning process. Rather than directly predicting actions, our method decomposes ObjectNav into structured multi-turn question-answering (QA) pairs that explicitly model human-like decision-making, inspired by how humans iteratively refine their understanding of an environment. To further mitigate the effects of noisy RGB observations and incorrect commonsense associations, we introduce a closed-loop mechanism that incorporates confidence-weighted adaptive learning, prioritizing high-confidence trajectories and reducing the influence of unreliable predictions. This refinement improves generalization to unseen objects and environments while enabling more robust and reliable navigation decisions.

Our key contributions are as follows:

- We introduce CL-HCoTNav, a VLM-driven ObjectNav framework that integrates structured hierarchical reasoning into navigation policy learning. We fine-tune a small-scale pre-trained VLM using multi-turn QA data derived from human demonstration trajectories, enabling H-CoT prompting to extract compositional knowledge and provide intermediate supervision for navigation.
- We further develop a closed-loop H-CoT mechanism that incorporates quantifiable detection and reasoning confidence scores, educing noise from observations and improving robustness against hallucinated reasoning.
- Extensive experiments in the AI Habitat demonstrate that our method outperforms state-of-the-art ObjectNav baselines in both success rate (SR) and success weighted by path length (SPL) by 22.4%.

II. Methodology

A. Problem formulation

In the ObjectNav task, consider a set of seen object goal classes $C_{\text{train}} = \{c_1, c_2, \ldots, c_n\}$ available during training, where n is the total number of seen classes, and a set of seen scenes X_{train}. During zero-shot testing, the agent navigates to objects from an unseen target class set $C_{\text{test}} = \{u_1, u_2, \ldots, u_m\}$, where m is the total number of unseen classes, within unseen scenes X_{test}. The training and testing sets are disjoint, i.e., $X_{\text{train}} \cap X_{\text{test}} = \emptyset$ and $C_{\text{train}} \cap C_{\text{test}} = \emptyset$. The robot is first trained to navigate to a target object c_i from C_{train} within a scene $X_{\text{train},i}$, given egocentric RGB observations, and is later evaluated in a zero-shot setting where it must navigate to a target object from C_{test} in an unseen scene $X_{\text{test},i}$. Each episode initializes the agent at a random position p_0^i with a random orientation in scene s_0^i, and the target object category c_i is provided. An episode is therefore characterized by $T_i = \{s_i, c_i, p_i, o_i, a_i\}$. At each time step t, the agent receives an observation o_t^i from its current viewpoint and selects an action a_t^i. The observation consists of an RGB image, the agent's location and orientation, and the target object category. The action space A consists of six discrete actions: *move_forward*, which moves the agent forward by 25 cm; *turn_left* and *turn_right*, which rotate the agent 30° left or right; *look_up* and *look_down*, which adjust the camera pitch by 30°; and *stop*, which signals that the target has been reached. An episode is considered successful if the agent executes the *stop* action and the target object is visible, as well as within 0.2 m of the target. Each episode is constrained to a maximum of 500 time steps.

B. Overview

The CL-HCoTNav framework is illustrated in Fig. 2. To model human-like reasoning in ObjectNav, our method transforms egocentric RGB observations and corresponding human demonstration trajectories into structured, multi-turn question-answering (QA) sequences. This is achieved using a pipeline of pre-trained (vision)-language models. Each QA sequence aligns intermediate reasoning steps with ground-truth actions, enabling supervision towards low-level behavior cloning.

At the core of the framework is the H-CoT prompting process, which decomposes ObjectNav into two reasoning phases. Perception and Planning Rounds. In the Perception Rounds, the system identifies salient objects from the current RGB observation, infers semantic co-occurrence relationships at both the room and object levels, and constructs a structured scene-level context. These steps allow the model to hypothesize where the target object is likely to be found, mimicking the cognitive strategies used by humans during navigation. The Planning Rounds then reason over the accumulated context, producing high-level navigation suggestions (e.g., "turn left" or "explore another room") that are discretized into executable control actions. This hierarchical prompting process acts as an interpretable bridge between perception and action, promoting generalization for out-of-distribution object goal and scenes.

To enhance robustness and learning efficiency, we further propose the Closed-Loop H-CoT mechanism. While H-CoT provides structured supervision, it remains vulnerable to label errors due to noisy visual inputs or incorrect semantic reasoning associations. To mitigate this, we introduce a confidence scoring system that evaluates the reliability of each QA pair

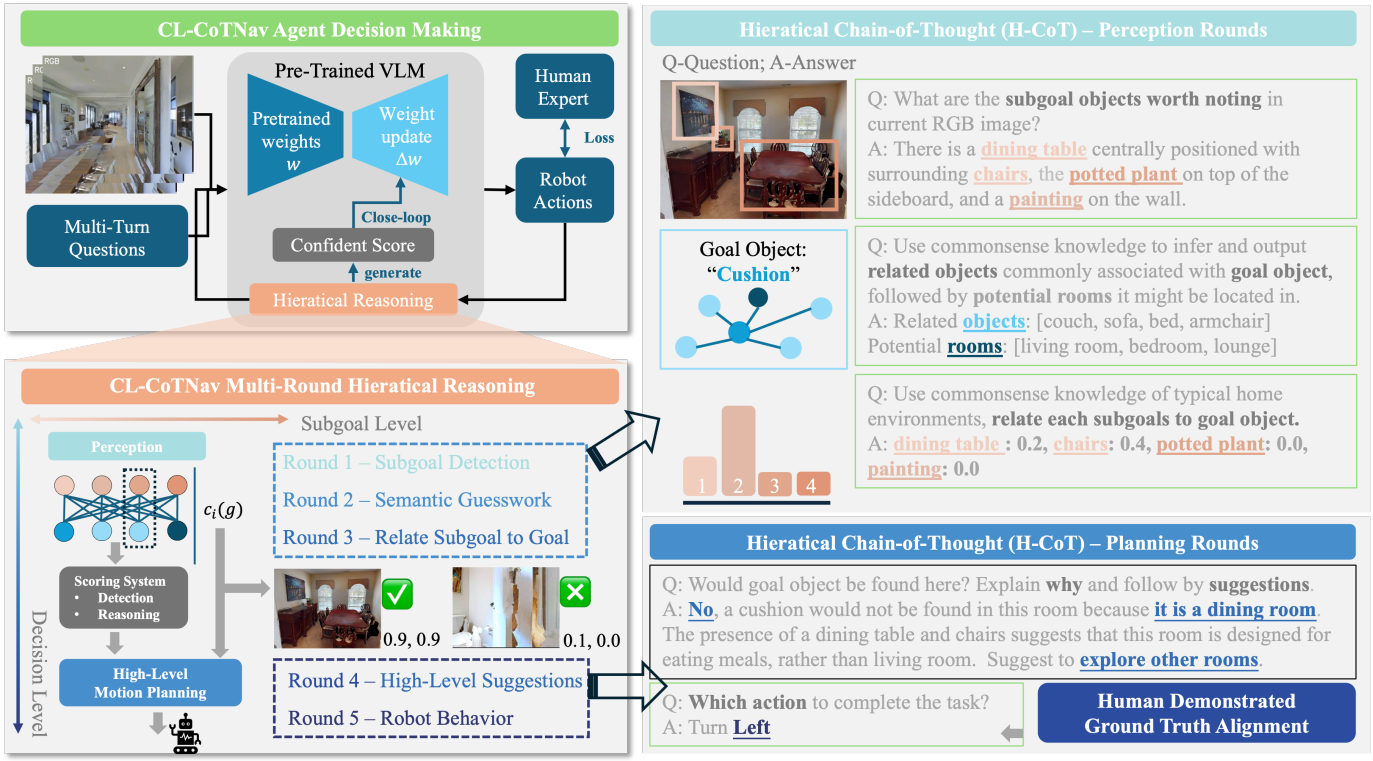

Fig. 2. Overview of proposed CL-HCoTNav. The structured dataset derived from human demonstration trajectories including two main turns: perception and planning, to iteratively extract compositional knowledge from egocentric RGB observations through a sequence of large pre-trained models and finally aligned with human demonstration actions.

based on reasoning consistency and input quality. During training, these confidence scores are integrated into an adaptive loss function that modulates the influence of each sample. High-confidence samples contribute more strongly, while uncertain or potentially hallucinated reasoning paths are down-weighted.

The complete QA dataset, annotated with hierarchical reasoning steps and associated confidence scores, is then used to fine-tune a smale-scale VLM via Low-Rank Adaptation (LoRA). By combining hierarchical prompting with closed-loop learning, CL-HCoTNav achieves robust generalization to unseen scenes and novel object categories. We next detail the two key components of our approach: the H-CoT reasoning process and the Closed-Loop confidence integration.

C. Hierarchical Chain-of-Thought (H-CoT)

To enrich the supervision beyond discrete human demonstrated actions, we introduce H-CoT—a structured reasoning process that mimics human cognitive strategies for navigation. H-CoT decomposes the ObjectNav task into two multi-turn reasoning stages: Perception Rounds for semantic scene understanding and Planning Rounds for informed decision-making. This design enables compositional reasoning over spatial and semantic cues, allowing the model to generalize beyond training distributions.

In the **Perception Rounds**, a sequence of QA rounds is applied to egocentric RGB observations to extract structured scene context. The first round identifies subgoal objects and their spatial arrangements (e.g. "a dining table is centrally positioned with surrounding chair"), capturing the visual layout of the environment. Subsequent rounds perform semantic guesswork through two levels of association. At the room level, the model infers the scene type based on typical object co-occurrence (e.g., "a TV and sofa suggest a living room"). At the object level, it relates detected subgoal objects to the target via commonsense priors (e.g., "a cushion is likely near a couch, but not near a stove"). These associations allow the model to hypothesize the likely presence of the target without direct observation. Each subgoal object is assigned a relevance score that reflects its semantic proximity to the target object, serving as a soft-attention mechanism. This scoring filters out spurious detections and prioritizes contextual cues that are semantically informative, forming a rich representation of the current scene to guide object navigation.

Building on this context, the **Planning Rounds** generate navigation suggestions grounded in the inferred scene semantics. The model evaluates whether the current room is a plausible location for the target; if not, it recommends exploration strategies such as "explore another room" or "turn around." These suggestions are then mapped into high-level textual decisions, which are further discretized into text-based executable control actions. Critically, these planning outputs are aligned with human demonstration actions, enabling abstract reasoning in behaviorally relevant supervision.

To generate this hierarchical supervision, we annotate human demonstration data using a pipeline of pre-trained (V)LLMs, resulting in this dataset of structured multi-turn QA pairs aligned with human actions. This dataset enables supervised training that incorporates both detection, reasoning, and control, bridging the gap between low-level action imitation and high-level semantic understanding. By explicitly modeling semantic co-occurrence relationships and decomposing decision-making into interpretable stages, H-CoT provides a strong inductive bias for zero-shot generalization and allows robot to compositionally reason about new objects and scenes.

D. Closed-Loop H-CoT Mechanism

While the H-CoT framework introduces structured reasoning for ObjectNav, it remains vulnerable to failures caused by noisy visual inputs, hallucinated associations, or unreliable outputs from the underlying vision-language models. These detection and reasoning inconsistencies can degrade navigation performance when treated equally during training. To address this, we introduce, a feedback-driven strategy that introduces reasoning confidence scores to modulate training dynamics.

Instead of treating all training samples with equal importance regardless of their semantic clarity or visual quality, our approach attaches a confidence score to each sample during H-CoT generation, capturing the reliability of both detection (object grounding) and reasoning (semantic inference) at each turn. Specifically, for each multi-turn QA sequence labeled by pre-trained models, we parse out the final text-based action suggestion and compare it to the human-demonstrated ground truth action. The degree of semantic alignment—combined with visual detection certainty—forms a confidence score $c_i \in [0, 1]$ for each reasoning trajectory.

These confidence scores are then integrated into training via an adaptive loss weighting mechanism. ObjectNav is formulated as a multi-class classification task, where the model predicts a discrete navigation action (e.g., *forward*, *left*, *right*) conditioned on the RGB image and the corresponding structured QA prompt. The baseline training objective is defined using categorical cross-entropy:

$$L_{\text{CE}} = -\log \hat{y}_{i,y_i} \tag{1}$$

where $\hat{y}_{i,y_i}$ is the predicted probability assigned to the correct action label y_i. To prioritize learning from trustworthy trajectories and downweight unreliable supervision, we define a sigmoid-based adaptive loss function:

$$L_{\text{adaptive}} = \frac{1}{1 + \exp(-\alpha(c_i - \beta))} \cdot (-\log \hat{y}_{i,y_i}) \tag{2}$$

Here, c_i is the confidence score of the i-th sample, while α and β are hyperparameters controlling the sharpness and threshold of the weighting function.

This closed-loop design enables the model to selectively attend to high-quality detection and reasoning timesteps while minimizing the impact of noise or hallucinations from photo-realistic scene and language model. By integrating confidence-weighted fine-tuning into the training loop, the Closed-Loop H-CoT mechanism improves generalization to unseen environments and enhances the robustness of decision-making under real-world visual uncertainty.

III. Experiment

In this section, we evaluate our method and other ZSON methods through the commonly used Habitat platform and discuss the experimental results. More details of experiment settings, including the dataset split, simulation platform, and training parameter can be found in AppendixC.

A. Comparison Models

We compare our proposed method with representative baseline and state-of-the-art (SOTA) approaches in ObjectNav, spanning RL, imitation learning (IL), and VLM paradigms.

Baseline [36]: A RL approach trained from scratch using egocentric RGB inputs. It directly use a pre-trained ResNet to extract a 1-D visual feature from the RGB observation and concatenate with the semantic embedding of the target class as the input of the policy network. The policy is learned end-to-end using PPO without incorporating semantic priors.

Habitat-Web [21]: An IL method that trains ObjectNav agents directly from human demonstration trajectories. The model maps egocentric RGB observations to expert-labeled actions via an end-to-end MLP policy, similar to Baseline.

VLFM [30]: A VLM-based modular navigation framework that employs BLIP-2 [16] for semantic matching. It computes cosine similarity between the agent's current RGB view and the target object description, projecting scores onto a semantic map for goal selection. The model is frozen during deployment and not fine-tuned for ObjectNav tasks.

SSNet [35]: A RL-based zero-shot ObjectNav model that integrates object detection scores and word embedding similarity as anxillary input to policy learning.

DivScene [26]: An VLM approach that also introduces chain-of-thought (CoT) reasoning for decision-making. Unlike our method, which learns an end-to-end policy from human behavior using fine-tuned VLMs, DivScene employs CoT supervision based on shortest path trajectories.

B. Training Results

Table I presents the evaluation performance of CL-HCoTNav and other baseline models on the training splits after training. All methods achieve comparable success rates (SR) and success weighted by path length (SPL), indicating that ObjectNav is generally learnable across different paradigms.

CL-HCoTNav consistently achieves the highest SPL and competitive SR, demonstrating the effectiveness of structured hierarchical reasoning and confidence-aware fine-tuning. Compared to IL approaches such as Habitat-Web and DivScene, CL-HCoTNav exhibits notably higher SPL, suggesting that structured multi-round reasoning contributes to more efficient and purposeful trajectories. While VLFM leverages frozen vision-language embeddings and commonsense priors, it performs worse than fine-tuned approaches, particularly in SPL. This confirms that adaptation to navigation-specific tasks is

critical for fully exploiting the semantic reasoning capacity of large-scale VLMs. Although DivScene incorporates CoT supervision into its IL pipeline, its reliance on shortest-path ground truth limits its adaptability. In contrast, CL-HCoTNav further benefits from closed-loop mechanism, improving robustness by emphasizing high-confidence data pairs.

We also observe that increasing the number of human demonstration trajectories improves performance, though with diminishing returns. Moving from 35k to 50k demonstrations yields substantial gains in both SR and SPL, while the improvement from 50k to 70k is more modest. This suggests that while IL benefits from more data, its effectiveness saturates without structured supervision. In this context, our results highlight the importance of structured reasoning over pure data scaling for achieving high-quality navigation performance.

TABLE I
TRAINING RESULTS FOR OBJECTNAV. WE REPORT SUCCESS RATE (SR) AND SUCCESS WEIGHTED BY PATH LENGTH (SPL) FOR EACH SOTA.

Method	SR (%)	SPL (%)	Training
Object Goals			
Baseline [36]	63.3	0.21	Yes
SSNet [35]	65.4	0.23	Yes
Habitat-Web [21]	69.1	0.26	Yes
VLFM [30]	70.1	0.28	No
DivScene [26]	73.8	0.30	Yes
CL-HCoTNav (MP3D-HD-35k-C16)	**74.1**	**0.31**	Yes
Scenes			
Baseline [36]	64.1	0.22	Yes
SSNet [35]	67.5	0.25	Yes
Habitat-Web [21]	70.8	0.27	Yes
VLFM [30]	71.3	0.28	No
DivScene [26]	75.6	0.32	Yes
CL-HCoTNav (MP3D-HD-35k)	73.5	0.32	Yes
CL-HCoTNav (MP3D-HD-50k)	74.8	0.35	Yes
CL-HCoTNav (MP3D-HD-70k)	**76.2**	**0.38**	Yes
Humans	93.7	42.5	

C. Zero-Shot Generalization Test

Table II reports performance of various models on ZSON under two settings: novel object categories and novel scenes. CL-HCoTNav achieves the highest success rate and SPL in both settings, demonstrating its strong generalization ability through hierarchical reasoning and confidence-aware learning.

For novel object generalization, models that incorporate semantic priors—such as VLFM, SSNet, DivScene, and CL-HCoTNav—outperform traditional RL (Baseline) and IL (Habitat-Web). Although trained on expert demonstrations, Habitat-Web fails to generalize well to unseen objects, likely due to overfitting to category-specific patterns without broader semantic reasoning. VLFM benefits from frozen vision-language embeddings but lacks fine-tuning, resulting in low SPL due to inefficient trajectory planning. SSNet improves upon RL by modeling object-object associations but remains limited by its static representations. DivScene introduces chain-of-thought reasoning during VLM finetuning, leading to improved generalization, but is constrained by reliance on shortest-path supervision. CL-HCoTNav outperforms all baselines by explicitly modeling both object-level and room-level semantic relationships and refining decision-making through confidence-based weighting. This structured reasoning enables the agent to infer the likely location of unseen targets and navigate efficiently without memorized spatial priors.

In the novel scene generalization setting, the focus shifts to layout adaptation. RL-based methods struggle due to overfitting to seen environments. While SSNet introduces object-aware exploration cues, its performance remains limited in unfamiliar layouts. Imitation learning models, including Habitat-Web, experience significant performance drops, suggesting limited adaptability to new spatial arrangements. VLFM achieves better generalization due to its strong semantic priors, but its frozen architecture limits efficiency. DivScene benefits from intermediate reasoning but remains less effective than CL-HCoTNav in navigating complex spatial layouts. CL-HCoTNav maintains a relatively small performance gap between training and zero-shot testing, indicating superior generalization. Its hierarchical planning allows flexible adaptation to novel configurations, while the closed-loop mechanism suppresses unreliable supervision, improving trajectory quality.

Overall, these results highlight the importance of structured multi-turn reasoning and adaptive learning in achieving robust generalization. CL-HCoTNav bridges the limitations of prior approaches by integrating semantic reasoning with confidence-aware training, offering a scalable solution for ZSON.

TABLE II
ZERO-SHOT OBJECTNAV RESULTS ON MP3D VAL SPLIT. WE REPORT SUCCESS RATE (SR) AND SUCCESS WEIGHTED BY PATH LENGTH (SPL) UNDER TWO GENERALIZATION SETTINGS: NOVEL OBJECT GOALS AND NOVEL SCENES. ALL MODELS ARE TRAINED ON 35K TRAJECTORIES.

Method	Novel Object Goals		Novel Scenes	
	SR (%)	SPL (%)	SR (%)	SPL (%)
Baseline [36]	22.7	5.1	25.3	6.1
Habitat-Web [21]	26.5	7.4	28.2	9.0
VLFM [30]	34.2	14.8	35.8	16.5
SSNet [35]	30.2	10.8	31.1	12.1
DivScene [26]	44.1	19.1	46.7	21.3
CL-HCoTNav (35k)	**55.2**	**25.7**	**58.5**	**27.4**

D. Ablation Study

This section isolates the contributions of hierarchical reasoning and closed-loop learning within the CL-HCoTNav framework, analyzing their impact on zero-shot generalization.

As shown in Table III, proposed H-CoT plays a central role in boosting generalization. Compared to a baseline using pure human annotations—where the QA format is limited to querying the target object and returning the human-demonstrated action—standard CoT prompting [8] improves performance by introducing intermediate reasoning. However, H-CoT further amplifies these gains by introducing a two-stage structure that separately models subgoal identification and semantic reasoning across room- and object-level contexts. This design leads to better-informed decision-making, significantly reducing failure cases arising from reactive or shallow policies.

Building on H-CoT, the Closed-Loop CoT mechanism provides an additional performance boost. By incorporat-

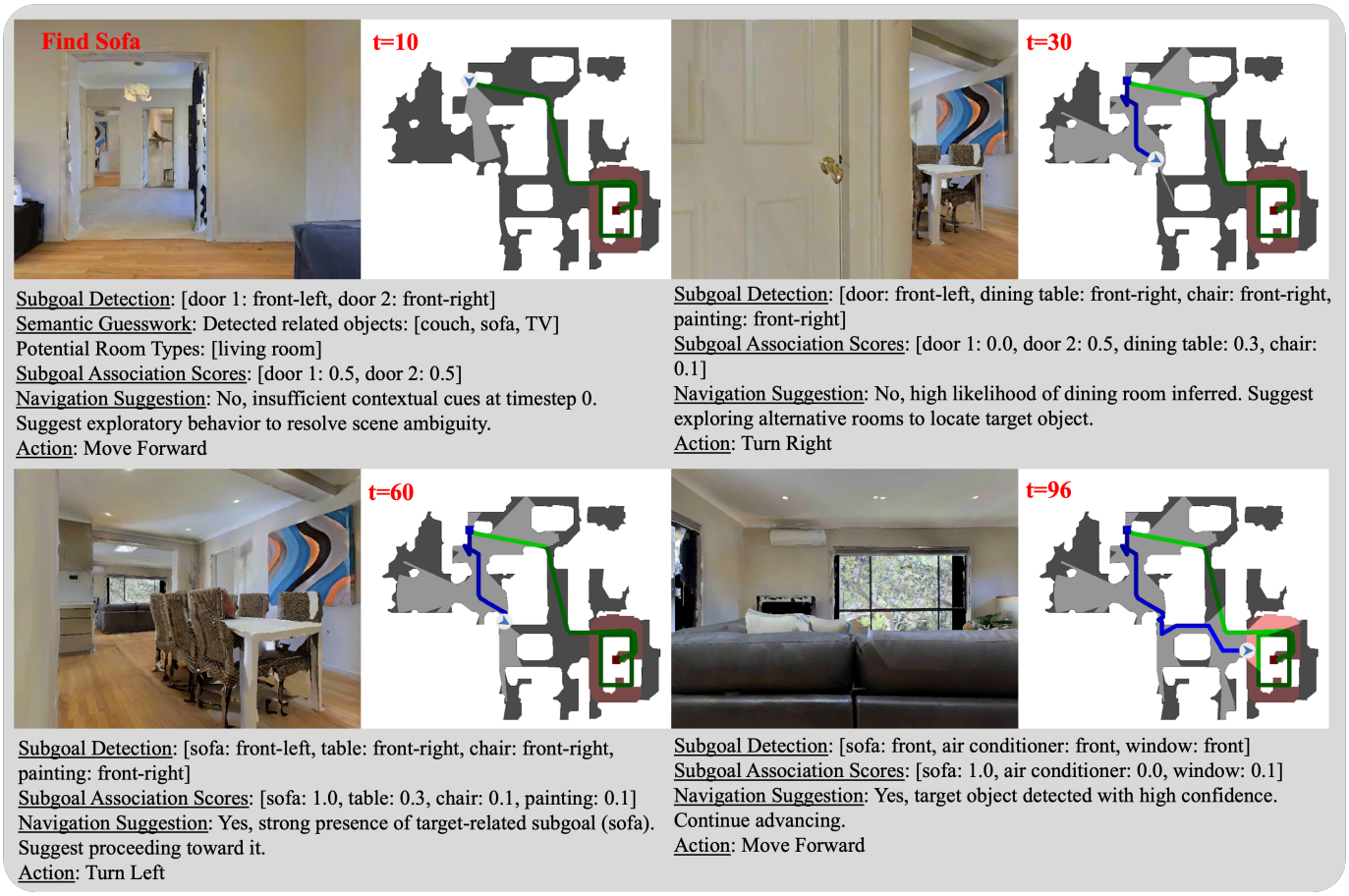

Fig. 3. Zero-shot generalization results on MP3D Val. The figure illustrates how CL-HCoTNav navigates in unseen scene layouts. The predicted navigation path is shown in blue and shortest path is shown in green. SPL = 0.71, ep_length = 97.

TABLE III
ABLATION STUDY RESULTS ON MP3D VAL (UNSEEN SCENES). THIS STUDY EVALUATES THE IMPACT OF HIERARCHICAL REASONING (H-COT) AND ADAPTIVE LEARNING (CLOSED-LOOP COT) ON GENERALIZATION.

Method	Success (↑)	SPL (↑)
Pure Text (Human Annotations)	24.3%	6.5%
Standard CoT [8]	36.5%	15.8%
H-CoT (Hierarchical CoT Only)	52.9%	23.1%
CL-HCoTNav (H-CoT + Closed-Loop)	55.2%	25.7%

ing reasoning confidence scores into the training loop, the model learns to prioritize high-quality supervision while down-weighting noisy or semantically ambiguous samples. This adaptive loss weighting refines decision-making and enhances robustness under distribution shifts, particularly in unseen environments. The improvement from H-CoT to CL-HCoTNav highlights the value of confidence-aware learning in mitigating the effect of hallucinated associations and unreliable intermediate predictions. Taken together, these results validate that CL-HCoTNav's superior generalization stems not only from its structured reasoning process, but also from its ability to selectively learn from trustworthy examples.

IV. CONCLUSION AND FUTURE WORK

In this work, we introduce CL-HCoTNav, a VLM-driven ObjectNav framework that integrates structured reasoning and closed-loop feedback into navigation decision-making. We fine-tune VLM using multi-turn QA pairs derived from human demonstration trajectories. This structured dataset enables hierarchical Chain-of-Thought prompting, iteratively extract compositional knowledge and provide auxiliary navigation guidance. Additionally, we propose a Closed-Loop H-CoT mechanism that incorporates confidence scores into training to prioritize high-confidence data pairs, enhancing robustness against hallucinated or incorrect reasoning. Extensive experiments in the AI Habitat environment demonstrated that our method achieves superior generalization to unseen scenes and target objects, outperforming state-of-the-art approaches.

Despite these advancements, our approach still relies on imitation learning, which is inherently limited by the quality and coverage of the dataset. To overcome this, future work will explore finetuning VLM using RL to enhance online policy learning beyond supervised data constraints. Further, we plan to conduct physical experiments to validate the approach in realistic settings.

REFERENCES

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] 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.

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[6] 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.

[7] Bolei Chen, Haina Zhu, Shengkang Yao, Siyi Lu, Ping Zhong, Yu Sheng, and Jianxin Wang. Socially aware object goal navigation with heterogeneous scene representation learning. *IEEE Robotics and Automation Letters*, 2024.

[8] William Chen, Oier Mees, Aviral Kumar, and Sergey Levine. Vision-language models provide promptable representations for reinforcement learning. *arXiv preprint arXiv:2402.02651*, 2024.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[10] Samyak Datta, Oleksandr Maksymets, Judy Hoffman, Stefan Lee, Dhruv Batra, and Devi Parikh. Integrating egocentric localization for more realistic point-goal navigation agents. In *Conference on Robot Learning*, pages 313–328. PMLR, 2021.

[11] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 8(79): eadf6991, 2023.

[12] Xu Guo and Han Yu. On the domain adaptation and generalization of pretrained language models: A survey. *arXiv preprint arXiv:2211.03154*, 2022.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[14] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022.

[15] Yuxuan Kuang, Hai Lin, and Meng Jiang. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. *arXiv preprint arXiv:2402.10670*, 2024.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[17] Shiwei Lian and Feitian Zhang. Tdanet: Target-directed attention network for object-goal visual navigation with zero-shot ability. *arXiv preprint arXiv:2404.08353*, 2024.

[18] 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.

[19] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[21] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5173–5183, 2022.

[22] James A Sethian. A fast marching level set method for monotonically advancing fronts. *proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[23] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

[24] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023.

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

[26] Zhaowei Wang, Hongming Zhang, Tianqing Fang, Ye Tian, Yue Yang, Kaixin Ma, Xiaoman Pan, Yangqiu Song, and Dong Yu. Divscene: Benchmarking lvlms for object navigation with diverse scenes and objects. *arXiv preprint arXiv:2410.02730*, 2024.

[27] 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.

[28] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6750–6759, 2019.

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

[30] 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.

[31] 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.

[32] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[33] Liang Zhang, Leqi Wei, Peiyi Shen, Wei Wei, Guangming Zhu, and Juan Song. Semantic slam based on object detection and improved octomap. *IEEE Access*, 6:75545–75559, 2018.

[34] Qianfan Zhao, Lu Zhang, Bin He, and Zhiyong Liu. Semantic policy network for zero-shot object goal visual navigation. *IEEE Robotics and Automation Letters*, 2023.

[35] Qianfan Zhao, Lu Zhang, Bin He, Hong Qiao, and Zhiyong Liu. Zero-shot object goal visual navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2025–2031. IEEE, 2023.

[36] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.

APPENDIX

In this appendix, we provide supplementary materials to further elaborate on CL-HCoTNav:

- Related Work
- Hyperparameters for VLM Fine-tuning
- Dataset and Evaluation Protocols
- Evaluation Metrics
- Implementation Details

RELATED WORK

A. Object Goal Visual Navigation

Visual Object Goal Navigation (ObjectNav) requires a robot to locate and navigate to a target object using egocentric observations[25]. Existing approaches can be categorized into modular learning and end-to-end learning [11]. Modular learning methods decompose the navigation task into semantic mapping, goal selection, and motion planning. These methods construct a top-down semantic map, select exploration goals based on learned or heuristic policies, and execute low-level actions through a local planner [11]. Representative pipelines [6, 33] incrementally build episodic semantic maps while employing goal-agnostic, frontier-based exploration. The semantic exploration policy determines navigation goals by leveraging learned priors on object spatial relationships, while the local planner [22] generates paths and executes actions. The exploration policy operates at a coarse time scale, whereas the planner continuously updates the map and refines the path at a finer scale. Although modular approaches are interpretable and transferable to real-world settings, their reliance on accurate localization and mapping limits practice in large-scale environments. End-to-end learning methods, in contrast, directly map raw sensor inputs and goal descriptions to navigation actions using deep neural networks, bypassing explicit mapping. It learns implicit representations of the observation before inputting it into the navigation policy, exploiting the object relationships or semantic contexts, aiming for a more robust navigation policy. [17] incorporate attention mechanisms to prioritize relevant observed objects. Meta-learning strategies [28] enable adaptive navigation in unseen environments without explicit supervision, and graph neural networks [7] have been employed to model object relationships and extract semantic interaction features. [4] introduce an image-level representation to bridge the gap between RGB inputs and control-space actions. While end-to-end approaches captures complex visual-action associations, they require large-scale interactive training data and often struggle with long-horizon dependencies and memory retention. In this work, we adopt this end-to-end learning paradigm, integrating semantic relationships reasoning among the target object, observed objects and scene via the H-CoT reasoning process.

B. Zero-Shot Object Goal Navigation

Traditional ObjectNav methods are trained on a fixed set of object categories and scenes [36, 29], limiting their ability to generalize beyond seen cases. In contrast, humans can effortlessly locate novel objects in unfamiliar environments without prior exposure. Achieving similar generalization in ObjectNav requires disentangling navigation ability from specific training scenes and target objects. To evaluate this, experiments typically partition object target classes and scenes into seen and unseen categories, assessing the model's capability to navigate to novel objects and scenes without additional training. End-to-end methods such as SSNet [35] incorporate object detection results and cosine similarity between word embeddings to prevent class-specific policy overfitting. SPNet [34] refines policy learning through object-goal embeddings that guide action selection based on semantic similarity. EmbCLIP [14] and ZSON [18] utilize pretrained vision encoders and text-based embeddings [20] to establish semantic relationships between target objects and observed scenes, improving generalization without requiring additional annotated training data.

Despite advances in semantic embeddings and target-guided exploration, ZSON remains challenging due to the variability of object and scene distributions and the absence of unified features and prior knowledge for efficient search. To address this, our method leverages human-demonstrated trajectories, interpreting them hierarchically to generate structured priors, semantic reasoning, and agnostic exploration, thereby achieving robust generalization without extensive retraining.

C. Foundation Models for Visual Navigation

Recent works integrate pre-trained multimodal models into ObjectNav, reducing the need for training from scratch while leveraging strong visual recognition or reasoning. These methods primarily use foundation models for (1) exploration guidance by relating observations to target objects or (2) policy learning for direct action prediction. For exploration guidance, VLFM [30] employs BLIP-2 [16] to compute cosine similarity between observations and target prompts, projecting scores onto a semantic map to guide frontier-based exploration. OpenFMNav [15] decomposes ObjectNav into sequential stages, utilizing vision-language models (VLMs) for perception, reasoning, and constructing a semantic score map for language-guided navigation. CoW [16] provides object grounding, similarly projecting relevance scores onto a top-down map. While these methods enable zero-shot navigation without additional training, frozen VLMs may produce incorrect associations due to biases or hallucinations, lacking mechanisms for real-time correction. For policy learning, ViNT [24] encodes vision-language features into a transformer-based policy, requiring an additional adapter to map from token space to action space. DivScene [26] directly fine-tunes VLMs using imitation learning, but its reliance on annotated shortest paths—assuming prior knowledge of the target object's location—limits the model's ability to learn efficient search strategies. To address these limitations, our method directly employs a compact VLM as the full navigation policy, introducing a structured reasoning framework that extracts compositional knowledge from human-demonstrated trajectories.

DATASET

We evaluate our method using the Matterport3D (MP3D) scenes [5] in the Habitat simulator [19], which provides high-resolution, photo-realistic indoor scenes with 21 object goal categories. Our experiments follow the standard ZSON setting, which evaluates generalization across both novel object categories and unseen scenes [34]. Training is conducted on the MP3D-HD-70k dataset [21], which contains over 70,000 human demonstration trajectories collected across 56 scenes. To ensure data quality, we remove failed trajectories, filter non-navigable starting positions, and cap all episodes to a maximum length of 500 steps. After preprocessing, we obtain cleaned subsets of 35k, 50k, and 70k demonstrations, balanced between object classes and scene types. We design two evaluation protocols to measure generalization, as shown in Table IV: (1) object generalization, where the target categories differ between training and test, and (2) scene generalization, where the environments differ.

TABLE IV
BREAKDOWN OF TRAIN AND TEST DATASETS FOR SCENE AND OBJECT GENERALIZATION EXPERIMENTS. INSIDE () INDICATE EPISODE OR TARGET NUMBERS AFTER CLEANING.

Split	Dataset	Scenes	Episodes	Targets
Scene Generalization				
Train	MP3D-HD-70k	56	70,176 (53,827)	28 (21)
Train	MP3D-HD-50k	40	49,778 (37,925)	28 (21)
Train	MP3D-HD-35k	28	34,641 (26,517)	28 (21)
Test	MP3D-Val	11	2,195 (1,148)	21
Object Generalization				
Train	MP3D-HD-35k-C16	28	20,595	16
Test	MP3D-HD-35k-C05	28	5,922	5

Object Generalization. For this split, we adopt the setting from [34], where the 21 object categories in MP3D are divided into 16 seen and 5 unseen classes. The training set (MP3D-HD-35k-C16) includes trajectories involving only the seen classes, while the test set (MP3D-HD-35k-C05) uses the same 28 training scenes but targets the five unseen categories: *counter, bed, toilet, chest_of_drawers, plant*. This setting evaluates the model's ability to reason over novel object semantics not encountered during training. The remaining 16 categories—e.g., *chair, table, sofa, tv_monitor, sink*—are exclusively used for training.

Scene Generalization. To assess generalization to unseen spatial layouts, we train on the full MP3D-HD dataset using subsets of 28, 40, or 56 scenes (i.e., MP3D-HD-35k/50k/70k), each containing trajectories across all 21 object categories. Evaluation is performed on the MP3D-Val set, which includes 2,195 episodes across 11 held-out scenes that do not overlap with any training environments. This setting focuses on the model's ability to adapt to novel layouts and scene compositions, even when the object categories remain the same. Table IV summarizes the number of scenes, episodes, and target categories used across all training and evaluation configurations.

METRICS

We follow [1] to evaluate our method using Success Rate (SR), Success Weighted by Path Length (SPL), and Soft SPL for object-goal navigation tasks. SR is defined as: $\frac{1}{N}\sum_{i=1}^{N} S_i$ where $S_i = 1$ if the robot successfully reaches the target; otherwise, the episode is considered a failure. Success Weighted by Path Length (SPL) is defined as: $SPL = \frac{1}{N}\sum_{i=1}^{N} \frac{l_i}{\max(l_i, p_i)}$ where l_i is the shortest path from the start position to a successful stop position, and p_i is the robot's actual trajectory length in episode i. Finally, Soft SPL [10] accounts for navigation efficiency while incorporating partial progress toward the goal.

IMPLEMENTATION DETAILS

In the ZSON task, the robot is required to search for an instance of a specified object category (e.g., *bed*) within an unseen environment using only egocentric perception. The robot is equipped with an RGB-D camera and an odometry sensor that provides its pose relative to the episode's starting position. The simulated robot is 0.88 meters tall with a radius of 0.18 meters. It captures 480×640 RGB-D observations through a forward-facing camera mounted at a height of 0.88 meters, with a horizontal field of view (HFoV) of 79 degrees. All experiments are carried out using the Habitat Lab simulation platform [19].

To generate multi-round QA annotations for our dataset, we employ different pre-trained models for specific submodules: Qwen-VL-Chat [3] for subgoal detection, Qwen-7B [2] for semantic guesswork and object-target association, and ChatGPT-3.5-turbo for high-level action suggestion and scene-level reasoning. An example of the conversation template used in annotation generation is shown in Fig. 2. For navigation policy learning, we fine-tune a 2B-parameter VLM based on the InternVL2 [9] framework, which uses a ViT-based vision encoder and InternVL2-LM as the language model. Fine-tuning is performed using the LoRA technique [13] on a compute node equipped with 4 NVIDIA V100 GPUs. We use a batch size of 16 and train for 3 epochs, which takes approximately 19 hours on the MP3D-HD-50k dataset. The LoRA-specific fine-tuning hyperparameters are summarized below:

TABLE V
HYPERPARAMETERS FOR LORA FINE-TUNING

Parameter	Value
LoRA rank (r)	8
LoRA scaling factor (α)	16
LoRA dropout	0.05
Learning rate	3×10^{-4}
Batch size	16
Gradient accumulation steps	4
Weight decay	0.006
Warmup steps	500