IMAGINENAV: PROMPTING VISION-LANGUAGE MODELS AS EMBODIED NAVIGATOR THROUGH SCENE IMAGINATION

Xinxin Zhao*, **Wenzhe Cai***, **Likun Tang** , **Teng Wang**†
School of Automation, Southeast University
{xinxin_zhao, wz_cai, lktang, wangteng}@seu.edu.cn

ABSTRACT

Visual navigation is an essential skill for home-assistance robots, providing the object-searching ability to accomplish long-horizon daily tasks. Many recent approaches use Large Language Models (LLMs) for commonsense inference to improve exploration efficiency. However, the planning process of LLMs is limited within texts and it is difficult to represent the spatial occupancy and geometry layout only by texts. Both are important for making rational navigation decisions. In this work, we seek to unleash the spatial perception and planning ability of Vision-Language Models (VLMs), and explore whether the VLM, with only on-board camera captured RGB/RGB-D stream inputs, can efficiently finish the visual navigation tasks in a mapless manner. We achieve this by developing the imagination-powered navigation framework **ImagineNav**, which imagines the future observation images at valuable robot views and translates the complex navigation planning process into a rather simple best-view image selection problem for VLM. To generate appropriate candidate robot views for imagination, we introduce the **Where2Imagine** module, which is distilled to align with human navigation habits. Finally, to reach the VLM preferred views, an off-the-shelf point-goal navigation policy is utilized. Empirical experiments on the challenging open-vocabulary object navigation benchmarks demonstrates the superiority of our proposed system.

1 INTRODUCTION

A useful home-assistant robot should be able to search for different kinds of objects without telling it the exact 3D object coordinates for completing our human instructions. As our human always buying and bringing new goods back home, the robot's object searching capability should not be limited in a closed-set of categories. Researchers often refer this problem as open-vocabulary object navigation task. Recently, as the emergence of the foundation models, including capable vision models (Radford et al., 2021; He et al., 2022; Zhou et al., 2022; Cheng et al., 2024; Kirillov et al., 2023; Wu et al., 2024a; Liu et al., 2023), large language models (LLMs) and vision-language models (VLMs) (Brown et al., 2020; Chowdhery et al., 2023; Zhang et al., 2022; Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023; Liu et al., 2024; Chen et al., 2024; Dai et al., 2023; Gao et al., 2023), building an agent that can accomplish the open-vocabulary object navigation becomes possible. A popular framework, as shown in Figure 1, uses modular approach to deal with this problem, which often composes of four components: A real-time mapping and segmentation module to perceive robot surrounding environments. A template-based translation module to compress the semantic map information into texts. A LLM-based module to understand the textual information from the previous step and make a plan in texts. Finally, a path-planning module which project the reasoning result from LLM back to the map and plan a collision-free path navigating towards it.

Although such a pipeline achieves great success in recent years (Zhou et al., 2023; Kuang et al., 2024; Wu et al., 2024b; Zhang et al., 2024a; Shah et al., 2023; Yu et al., 2023a; Loo et al., 2024), these complex cascaded systems have several limitations. Firstly, both the depth camera and the robot localization module can suffer from perception error, especially for long-range depth estimation, and this can make the mapping process inaccurate. Secondly, online object detection and segmentation are required to augment spatial maps with semantic labels and prepare for LLM's

*Equal Contribution, †Corresponding Author: Teng Wang.

reasoning input. This increases the computational burden for the robots. Thirdly, although the semantic information stored on the map can be easily expressed by text (e.g., list the categories of the observed objects), such pure text prompts have difficultly in explicitly describing the geometry information and object details in the map, making it difficult and ambiguous for LLMs to infer the best navigation plan.

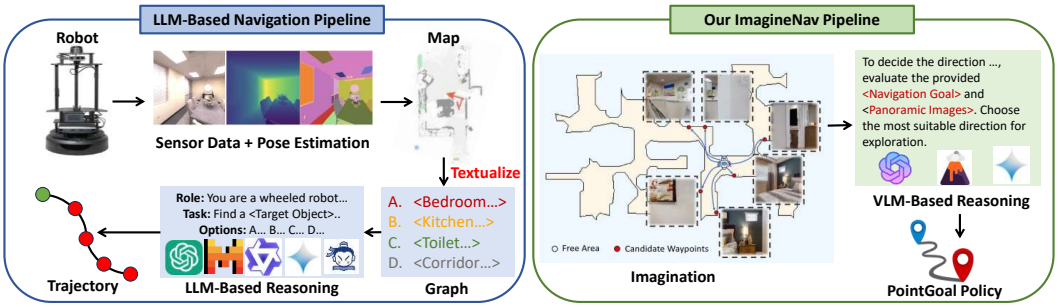

Figure 1: The comparison between the conventional LLM-based navigation pipeline and our ImagineNav pipeline. The traditional LLM-based navigation framework, illustrated on the left, relies on intricate sensor data processing and pose estimation for map creation, followed by LLM-driven reasoning to decide the exploration direction. Instead, our ImagineNav directly translates the long-horizon object goal navigation task into a sequence of best-view image selection tasks for VLM, which avoids the latency and compounding error in the traditional cascaded methods.

In this work, we try to explore whether it is possible to circumvent the complicated and fragile *mapping*$\rightarrow$ *translation*$\rightarrow$ *planning* framework, but develop a visual navigation approach with only raw RGB/RGB-D observations and pre-trained VLMs. Our proposed method - **ImagineNav** seeks to maximize the capabilities of VLMs in multimodal understanding and reasoning, and make the VLMs become an efficient embodied navigation agent. As most VLMs cannot understand the continuous physical world, it is infeasible to directly ask VLMs to generate navigable 3D waypoints. Instead, we translate the visual navigation problem into an imagination-powered best-view image selection task and let the VLMs select. To generate appropriate options for VLMs to choose from, we propose the **Where2Imagine** model which distills the human indoor navigation habits and generates future 3D navigation waypoints where a human might navigate based on the current visual observation. Such 3D navigation waypoints indicate relative poses with respect to the current frame and can be easily translated into new observation images using novel view synthesis (NVS) models. Afterwards, the VLMs only need to select the best imagined observation that is mostly related to the target object and drive the robot to follow the corresponding point-goal navigation trajectory. The above pose-aware imagination-and-selection capability allows the Object Goal Navigation (ObjectNav) task to be decomposed into a sequence of point-goal sub-tasks, facilitating the creation of collision-free navigation trajectories. Experimental results on standard benchmarks demonstrate the superiority of our ImagineNav over previous methods in open-vocabulary object navigation. In summary, our contributions are:

- We propose a mapless navigation approach ImagineNav. It leverages the imagination to generate image observations at potential future 3D waypoints as the VLMs' visual prompts, grounding the VLMs to become efficient navigation agents without any fine-tuning.
- We design a task-oriented model Where2Imagine to understand human navigation habits. This model is crucial to bridge the task-agnostic high-level VLM planners and the low-level navigation policies.
- Our ImagineNav increases success rate by a large margin of 15.1% and 10% respectively on HM3D (Ramakrishnan et al., 2021) and HSSD (Khanna et al., 2023). We also provide a detailed ablation analysis to help understand the important conclusions in our framework.

2 Related Works

2.1 Large models for robotic planning

Large-scale models pre-trained on extensive internet data have demonstrated formidable zero-shot reasoning capabilities in tasks such as planning (Huang et al., 2022), code generation (Liang et al.,

2023; Huang et al., 2023b), and solving science questions (Lewkowycz et al., 2022). The in-context learning capability of LLMs allows them to be applied to robotic task planning. Some methods (Liang et al., 2023; Huang et al., 2023b; Ahn et al., 2022) leverage LLMs to decompose tasks into subtasks, enhancing execution efficiency. Cap (Liang et al., 2023) generates robotic policy code directly from example language commands, enabling autonomous control and task execution based on natural language instructions. Instruct2Act (Huang et al., 2023b) combines LLM with foundational models (e.g., SAM and CLIP), reducing error rates in complex task execution, while SayCan (Ahn et al., 2022) combines LLM task planning with the feasibility of physical skills using pre-trained value functions, generating actionable plans for robots. However, one limitation of LLMs is their difficulty in embedding the robot's state directly into the planning process. To address this, many studies have turned to VLMs as alternatives. For instance, ViLA (Lin et al., 2024) significantly improves performance on multimodal tasks without compromising text capabilities by systematically exploring VLM pretraining design choices. CoPa (Huang et al., 2024) incorporates commonsense knowledge from VLMs, proposing a coarse-to-fine task-oriented grasping and task-aware motion planning approach. PIVOT (Nasiriany et al., 2024) transforms tasks into iteratively optimized visual question-answering problems via a refinement process. Socratic model (Zeng et al., 2022) integrates multiple pretrained large models (e.g., VLMs, LLMs, and audio models) in a modular fashion to enable reasoning and task execution through language-based interaction. These methods employ a set-of-examples (SOE) approach to guide VLM selection. We propose a new decision-making paradigm based on imagined imagery, wherein decisions are made on imaginations, enabling more nuanced, context-aware interactions that better harness VLMs' spatial perception capabilities.

2.2 Open-vocabulary object navigation

ObjectNav requires the robot to navigate toward a specific target category in unseen environments. Although previous works (Yadav et al., 2022; Ramrakhya et al., 2023; Chaplot et al., 2020; Ramakrishnan et al., 2022) can achieve high success rate in widely accepted benchmarks (e.g., habitat-challenge (Yadav et al., 2023)), most approaches are limited within a pre-defined object list, which is contradictory to the open-vocabulary real world. Therefore, many researchers start to discuss the open-vocabulary object navigation problem. End-to-end methods try to make use of compact multimodal features space (e.g., CLIP (Radford et al., 2021)) for grounding text knowledge into visual navigation problem but achieved limited performance (Khandelwal et al., 2022; Gadre et al., 2023; Majumdar et al., 2022). Instead, modular-based approaches (Huang et al., 2023a; Zhou et al., 2023; Achiam et al., 2023) typically necessitate the use of sensors for localization and mapping, high-level planning, and low-level control. These methods rely heavily on high-precision sensors for accurate self-localization and real-time map construction. Our approach introduces an imagination-based, mapless navigation framework. This framework circumvents the need for extensive training by transforming the complex process of navigation planning into a selection problem based solely on RGB inputs.

2.3 Imagination-based navigation

Recent methods (Zhai & Wang, 2022; Ramakrishnan et al., 2022; Zhu et al., 2022) have employed supervised learning to learn target-related functions in order to address the subtask of 'Where to look?' in navigation, specifically focusing on predicting the localization of target. These methods predict the absolute coordinates (Zhai & Wang, 2022) of the target, the shortest distance to target (Ramakrishnan et al., 2022), or the nearest boundary (Zhu et al., 2022) based on local maps. Instead, several studies (Ramakrishnan et al., 2020; Georgakis et al., 2022; Liang et al., 2021; Zhang et al., 2024b) have proposed various approaches to enhance the prediction of unobserved regions. For instance, (Ramakrishnan et al., 2020) introduced occupancy anticipation, where the agent infers an occupancy map based on RGB-D inputs. The L2M framework was introduced (Georgakis et al., 2022), consisting of a two-stage segmentation model that generates a semantic map beyond the agent's field of view and selects long-term goals based on the uncertainty of predictions. SSCNav algorithm (Liang et al., 2021) leverages semantic scene completion and confidence maps to infer the environment and guide navigation decisions. A self-supervised generative map (SGM) is proposed (Zhang et al., 2024b), which employs self-supervised approach to continually generate unobserved regions in the local map and predict the target's location. These methods primarily predict unobserved regions in top-down maps derived from egocentric RGB-D projections. In contrast,

Our method operates in the RGB space to perform guided-imagination. We utilize a compact model aligned with human navigation habits to generate new viewpoints, followed by the imagination process to produce corresponding visual representations.

3 METHODOLOGY

The open-vocabulary object navigation task requires the agent to locate an untrained target object instance in an unknown environment. At the start of each episode, the agent is spawned at a random position in an unfamiliar environment without any prior knowledge of the layout. The task is to find a target object g_i, which can belong to any category in an open-vocabulary setting. At each time step t, the agent receives an egocentric panorama view I_t, divided into 6 separate views, each represented by an RGB image $I_{t,i}$ accompanied by its depth map $D_{t,i}$. The discrete action space consists of the following commands: {Stop, MoveAhead, TurnLeft, TurnRight, LookUp, LookDown}. The 'MoveAhead' action moves the agent forward by 0.25m, while the rotational actions 'TurnLeft' and 'TurnRight' rotate the agent by 30 degrees. The task is considered successful if the agent reaches the target object with a geodesic distance smaller than a defined threshold (e.g., 1m) and executes the 'Stop' command within a fixed number of steps. Each episode has a maximum limit of 500 steps.

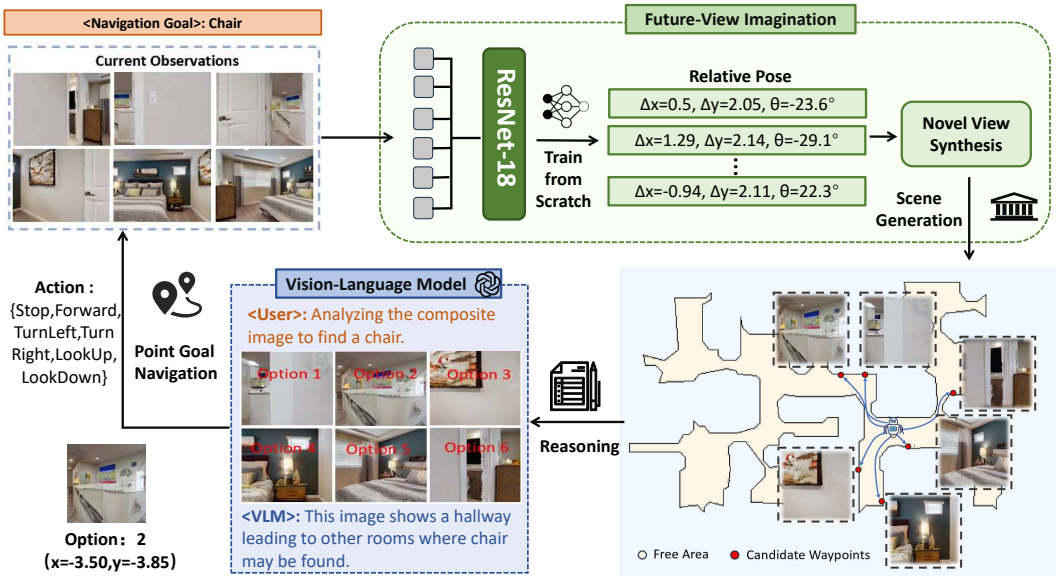

Figure 2: The overall pipeline of our mapless, open-vocabulary navigation framework. At each iteration, the agent captures a panoramic view of its surroundings. In the Imagination Module, the trained Where2Imagine module couples with novel view synthesis model to generate novel scene views. Guided by prompt templates, the VLM engages in target-oriented inference. Subsequently, the system executes the PointNav policy to determine the next navigational action. The above imagination, reasoning and planning procedure iterates until the target is reached.

3.1 IMAGINENAV FRAMEWORK

This section provides an overview of the imagination-based open-vocabulary object navigation framework (ImagineNav). As depicted in Figure 2, the agent initially employs the Where2Imagine module to generate candidate locations for imagination based on the current observations. Subsequently, the visual observations at these locations are imagined by a NVS model. Utilizing the generated images, which are annotated with option labels, the agent leverages a multimodal large-scale model to assess both the spatial and semantic information of each scene, enabling the selection of a more efficient exploration direction. Specifically, the VLM is employed to reason over the imagined images from six different views using prompts, selecting the optimal waypoint. Finally, the agent executes low-level point navigation strategy to reach the corresponding sub-goal. This process iterates, where each new observation serves as input for further imagination, reasoning, and

navigation, until the agent successfully identifies an instance of the target object. Consequently, the ObjectNav task is reduced to a sequence of simpler point-to-point navigation subtasks. Since our ImagineNav does not require any training on object-oriented data for reasoning and planning, it is open-vocabulary and can zero-shot generalize to novel semantic targets. Sec 3.2 introduces the imagination module, Sec 3.3 explains the use of the advanced planner VLM, and Sec 3.4 describes our underlying point navigation strategy.

3.2 FUTURE-VIEW IMAGINATION

To better leverage the spatial perception and reasoning capabilities of VLMs for open-vocabulary object navigation in unknown environments, we propose an future-view imagination model, which is composed by Where2Imagine module and a NVS model. As shown in Figure 2, Where2Imagine predicts the relative pose $(\Delta x, \Delta y, \theta)$ of the potential next navigation waypoint based on the current RGB observation, where Δx denotes lateral displacement, Δy represents longitudinal displacement, and θ refers to changes in the camera's viewing angle. Subsequently, a novel view image is generated based on the predicted relative pose. There are several advanced methods to achieve this, such as framing the task as a few-shot 3D rendering problem (Sargent et al., 2024; Wimbauer et al., 2023; Cao & de Charette, 2023) or utilizing generative models (e.g., diffusion models) for image synthesis (Yu et al., 2024; Tseng et al., 2023; Yu et al., 2023b). In this work, we employ a pretrained diffusion model (Yu et al., 2024), which generates new view images by taking the current image and the predicted relative pose as inputs, enabling high-quality viewpoint transformation.

Specifically, we trained a ResNet-18 from scratch for the relative waypoint prediction task, with training data sourced from the Habitat-Web project (Ramrakhya et al., 2022). Habitat-Web collects remote user demonstrations of virtual robot operations through a virtual teleoperation system on the Amazon Mechanical Turk platform, including 80k ObjectNav and 12k Pick&Place demonstrations. Humans in these demonstrations typically opt for directions towards important semantic cues (e.g., doors) that facilitate exploration. These navigation preferences serve as a valuable basis for determining potential waypoints, thereby enhancing the rationality and safety of the navigation strategy. We transformed the human demonstration trajectories into a paired dataset (I_t, P_{t+T}), where I_t represents the RGB image at frame t, and P_{t+T} denotes the relative pose $(\Delta x, \Delta y, \theta)$ at frame $t + T$ with respect to frame t. To facilitate the learning of the network, we filtered out images with a depth threshold of less than 0.3, as these images generally lack rich semantic information (e.g., a plain wall). Due to the limitations of the NVS model in generating high-quality images for large perspective shifts (e.g., 120°,180°,240°) based solely on a single input image, we restricted the training data to instances where the angular deviation falls within ±30°. Through the Where2Imagine module, our imagination model aligns with human navigation habits.

3.3 HIGH-LEVEL PLANNING

High-level planning module leverages the spatial awareness and common-sense reasoning capabilities of the VLM to select the direction most likely to locate the navigation target. We prefer GPT-4o-mini as the high-level planner because it offers a balance between the reasoning capabilities and practical efficiency. Compared to larger models (e.g., GPT-4o), GPT-4o-mini is lightweight and cost-effective. Its smaller size ensures faster inference, allowing the system to make timely decisions in dynamic environments. To assist GPT-4o-mini in decision making, we designed a simple prompt template, requiring the VLM to summarize its choice in a JSON format containing { 'Reason', 'Choice' }. This format allows for a clear understanding of the VLM's reasoning process. As illustrated in Figure 3, the VLM receives the synthesized observations at future navigation waypoints and the navigation goal as inputs. Based on the prompt, 'Your choice should first be based on discovering navigation targets, followed by the potential of unexplored areas...' , the VLM analyzes each image's semantic information, selects the optimal exploration direction, and returns the answer in the specified format. By providing the imagined observations as visual prompts to VLM, our ImagineNav offers significant advantages in spatial reasoning and decision-making processes. Firstly, the VLMs are more skilled at handling multiple-choice decision tasks compared to 3D geometry question answering (i.e., directly inferring the 3D coordinates of next waypoints). Moreover, the introduction of imagination enhances the decision-making capability of the VLM by supplementing detailed information about distant or visually unclear objects. The advanced planning module proposes new sub-goals after the low-level controller completes navigation to the designated target.

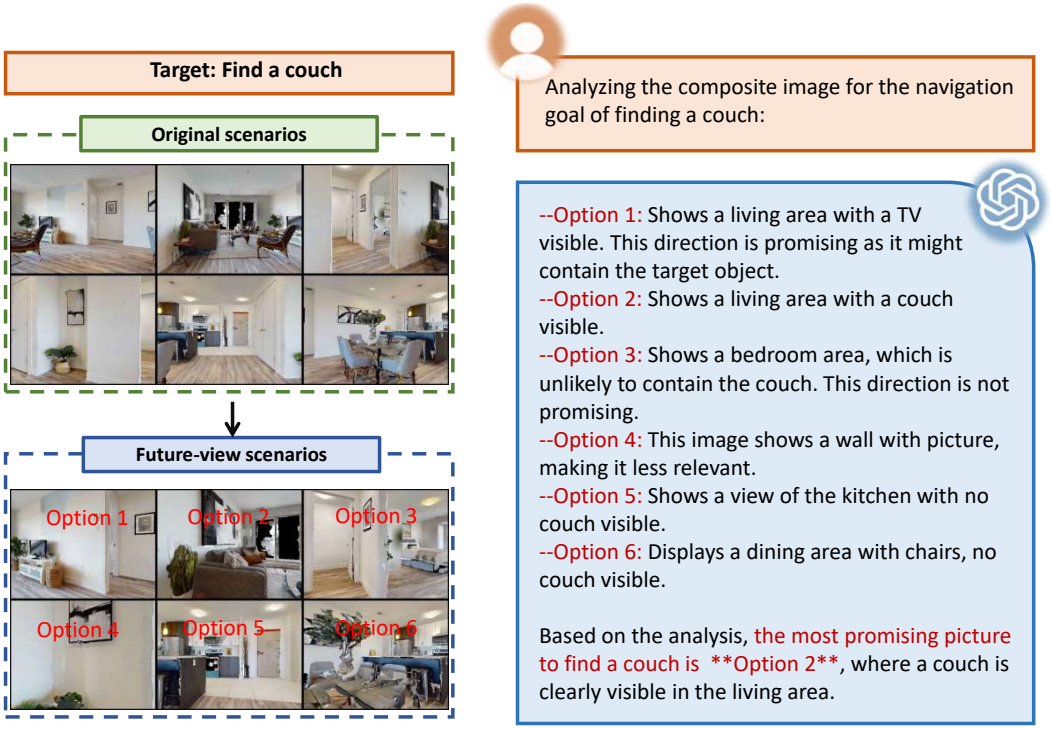

Figure 3: An example of the VLM analysis. By examining different future-view scenarios, the VLM pinpoints the direction most likely to incorporate the target object couch.

3.4 Controller

After the high-level planner provides the navigation points, the low-level controller executes Point Goal Navigation (PointNav) strategy to achieve these targets. Unlike ObjectNav, PointNav (turn to Δx, Δy) does not rely on semantic information from the environment but is instead driven solely by spatial perception. Currently, there are many methods available for achieving PointNav (Yang et al., 2023; Roth et al., 2024; Wijmans et al., 2022; Liang et al., 2024). To determine the execution actions at each step of the PointNav process, we use Variable Experience Rollout (VER) (Wijmans et al., 2022) as our underlying goal navigation strategy. VER combines the advantages of synchronous and asynchronous reinforcement learning, improving training efficiency and sample utilization in PointNav tasks, thereby enabling the agent to demonstrate stronger adaptability and generalization capabilities in complex environments.

4 Experiments

4.1 Experimental setup and metrics

We evaluate the effectiveness and navigation efficiency of our proposed method using the Habitat v3.0 simulator (Puig et al., 2023) on two standard ObjectNav datasets: HM3D (Ramakrishnan et al., 2021) and HSSD (Khanna et al., 2023). The HM3D dataset offers high-fidelity reconstructions of 20 entire buildings, including 80 training scenes and 20 validation scenes. The HSSD dataset provides 40 high-quality synthetic scenes, comprising 110 training scenes and 40 validation scenes. The experimental setup follows the ObjectNav-challenge-2023 (Yadav et al., 2023). For the data collection of the Where2Imagine module, we leveraged human demonstration trajectories from the MP3D (Chang et al., 2017) dataset within the habitat-web project with the camera height 0.88m and horizontal field of view (HFOV) of 79°. We report the performance in terms of Success Rate (SR), defined as the proportion of episodes where the agent's distance to the target object is less than 1m after executing the STOP action, and *SPL* (Anderson et al., 2018), Success weighted by path length,

$SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \left(\frac{\ell_i}{\max(p_i,\ell_i)}\right)$, where S_i be a binary success indicator in episode i, p_i is the agent path length and ℓ_i is the GT path length.

4.2 BASELINES

We conducted a comparative analysis of non-zero-shot and zero-shot ObjectNav methods to substantiate our proposed approach. Yamauchi (Topiwala et al., 2018) pioneered a frontier-based exploration strategy, emphasizing the boundaries between explored and unexplored regions. SemExp (Chaplot et al., 2020) advanced this concept by implementing goal-directed semantic exploration through the construction of semantic maps. In addition, we examined non-mapping closed-set object navigation baselines, including those based on imitation learning (Ramrakhya et al., 2022) and visual representation learning (Yadav et al., 2022).

For zero-shot object navigation, we consider several mapping-based baselines (Zhou et al., 2023; Wu et al., 2024b; Yokoyama et al., 2023) that integrate commonsense knowledge and semantic information to facilitate direct navigation toward target objects, leveraging semantic comprehension from pre-trained large models to aid in navigation. Furthermore, we explored RGB-based non-mapping navigation baselines: ZSON (Majumdar et al., 2022), which employs the CLIP (Radford et al., 2021) model to embed target images and object goals within a unified semantic space, thereby training a semantic goal-navigation agent, and PixNav (Cai et al., 2023), which utilizes pixel-level goal guidance, enabling pixel navigation through the use of VLMs and LLMs.

4.3 COMPARISON WITH PRIOR WORK

Table 1 presents a comparative analysis of the proposed ImagineNav against prior research efforts. On the HM3D dataset, ImagineNav achieves a success rate of 53.0% and a SPL of 23.8%, significantly outperforming most of the methods especially at success rate. Moreover, ImagineNav achieves the highest success rate and SPL on the HSSD dataset. Particularly, in open-vocabulary navigation tasks, our mapless ImagineNav even outperforms the best-performing map-based method VLFM (Yokoyama et al., 2023) by 0.5% at success rate. The above observations indicate that our ImagineNav demonstrates outstanding navigation performance across various settings, while maintaining low storage and computational complexities. Furthermore, since the pretrained NVS is directly employed without finetunned on the HM3D and HSSD datasets, we see a disparity between the quality of images generated by the NVS model and real images, limiting the capability of our model to some extent. To explore the upper limits of our framework, we instead use real panoramic images—specifically, the observation at the pose predicted by the Where2Imagine module—as visual prompts for the VLM model. Notably, both the success rate and SPL exhibit obvious improvements, obtaining 62.0% and 59.0% at success rate respectively on H3MD and HSSD benchmarks, which further demonstrates the superiority of our imagination-based navigation framework.

Table 1: **ImagineNav: Comparison with previous work.** The Where2Imagine model with T=11, utilizing ResNet-18 trained from scratch and GPT-4o-mini as the VLM, was evaluated over 200 epochs on the HM3D and HSSD datasets. ImagineNav uses NVS model to generate novel view images, while ImagineNav-Oracle uses real images of the candidate points.

Method	Open-Vocabulary	Mapless	HM3D		HSSD	
			Success Rate	SPL	Success Rate	SPL
FBE (Topiwala et al., 2018)	✗	✗	33.7	15.3	36.0	17.7
SemExp (Chaplot et al., 2020)	✗	✗	37.9	18.8	-	-
Habitat-Web (Ramrakhya et al., 2022)	✗	✓	41.5	16.0	-	-
OVRL (Yadav et al., 2022)	✗	✓	62.0	26.8	-	-
ESC (Zhou et al., 2023)	✓	✗	39.2	22.3	-	-
VoroNav (Wu et al., 2024b)	✓	✗	42.0	26.0	41.0	23.2
VLFM (Yokoyama et al., 2023)	✓	✗	52.5	30.4	-	-
ZSON (Majumdar et al., 2022)	✓	✓	25.5	12.6	-	-
PixNav (Cai et al., 2023)	✓	✓	37.9	20.5	-	-
ImagineNav	✓	✓	**53.0**	**23.8**	**51.0**	**24.9**
ImagineNav-Oracle	✓	✓	**62.0**	**31.1**	**59.0**	**27.0**

4.4 Ablation Study on Main Components

We conducted an ablation study on main components of the imagination module to demonstrate their effectiveness, with each variant evaluated for 100 epoches. As shown in Table 2, we increase success rate from 43.0 to 55.0 by utilizing future imaginations as visual prompt of the VLM for deciding exploration direction. Please note that at row #2 without Where2Imagine, we simply generate future views at the six locations, which are two meters away from current observations along their respective observation orientations. This improvement suggests the superiority of future imagination in facilitating VLM's reasoning. Such improvements can be attributed to the greater semantic disparity between different imaginations, as illustrated in Figure 4. Further incorporating Where2Image improve success rate from 55.0 to 64.0, and from 49.0 to 56.0 under settings of 'NVS' and 'w/o NVS', respectively. As mentioned above, the capability of our ImagineNav is limited by the performance of the off-the-shelf NVS model (Yu et al., 2024) to some extent as evidenced by comparing rows #2 with #4 and rows #3 with #5. Nevertheless, the incorporation of Where2Imagine partially mitigates the adverse effects of the NVS model.

Table 2: **ImagineNav: ablation study on the imagination module**. 'Imagination' refers to whether the future imaginations are used as visual prompts of the VLM. When it is removed, we feed current observations into VLM for deciding the best exploration direction, and set the next waypoint 2 meters away from the current location along the direction. Here, the distance of 2 meters is considered as it is comparable to that generated by T=11. 'NVS' indicates whether the image is captured from a real environment or synthesized via the NVS model.

Imagination	Where2Imagine	NVS	HM3D	
			Success Rate	SPL
✗	✗	Oracle	43.0	24.7
✓	✗	Oracle	55.0	27.6
✓	✓	Oracle	64.0	28.3
✓	✗	PolyOculus	49.0	23.3
✓	✓	PolyOculus	56.0	24.3

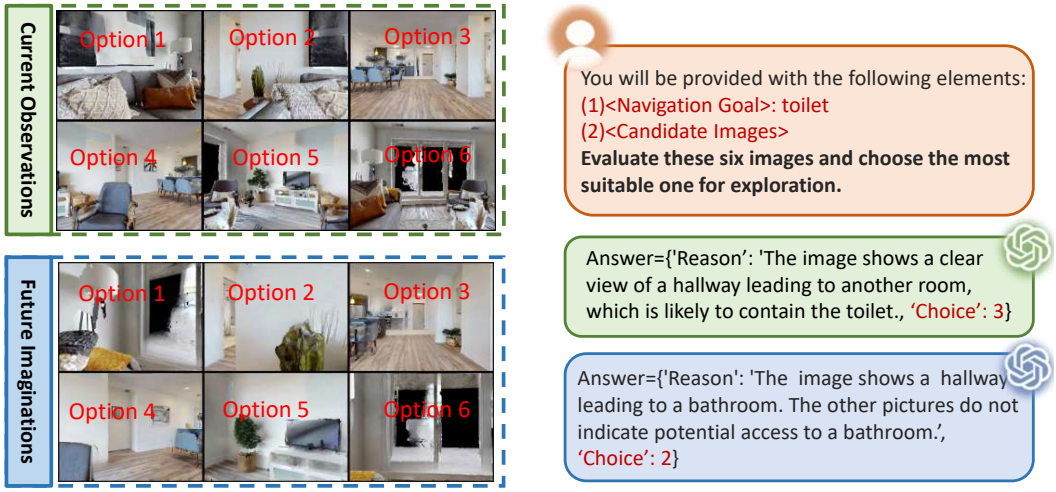

Figure 4: Visualization of the synthesized image observations at future navigation waypoints predicted by the imagination module. It can be seen that there exists drastic semantic disparity between different imaginations. In contrast, the semantic information is relatively consistent across different current observations. The varying semantics across different future views highlight the advantages of the imagination module in enhancing the VLM's decision-making capabilities.

4.5 Analysis of Successful and Failed Trajectories

Figure 5 illustrates that our method achieves efficient path planning and navigation across different targets. Especially, as shown in the top middle of Figure 5, the agent needs to navigate through multiple rooms. The complex environment increases the risk of losing direction, but our ImagineNav successfully infers the optimal path and finds the target, demonstrating its ability to handle long-distance and multi-room scenarios. We also present some failure examples at the bottom of Figure

5. We identified three key factors contributing to these navigation failures. First, some object instances are neglected for marking by the simulator, and therefore a successfully trajectory is wrongly considered as a failure (a.k.a. false failure) as shown in the bottom left of Figure 5. Second, the synthesized image from the NVS does not align with the real observation, such as creating objects that are not present in the real scene as shown in the middle of Figure 5, which causes the VLM to make incorrect inference. Moreover, the lack of historical information make the agent easily trapped in local optima, thereby limiting its performance in long-term navigationas shown in the bottom right of Figure 5.

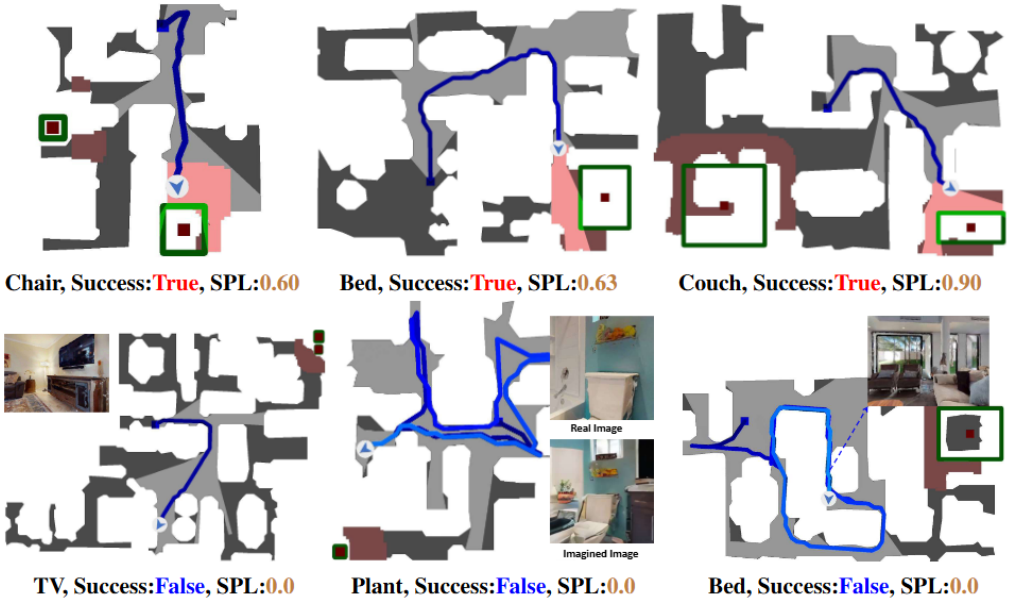

Figure 5: Visualization of the navigation trajectory. The top and bottom rows respectively show the complete top-down trajectories of successful and unsuccessful examples.

4.6 Analysis of Where2Imagine module

We explore the impact of the sampling step T on the final navigation performance by varying T from 8 to 15. For each T, we re-generate the labeled image data and re-train the ResNet-18 for relative pose prediction. Each variant was tested for 100 epochs under conditions where the agent had access to real panoramic observations. As shown in Table 3, the best success rate and SPL are obtained when T is set to 11. Furthermore, we visualize several navigational trajectories under different values of T in Figure 6 to facilitate explanation. As can be seen, when T is relatively small (i,e., 8), the agent is easily trapped as marked by red square, since it mainly resorts to local semantic information for inferring its exploration direction, making it susceptible to converging on suboptimal local solutions.

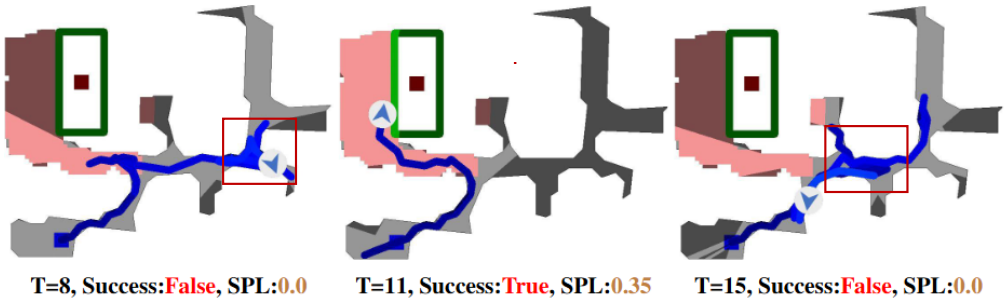

Figure 6: Comparison of trajectories at different sampling steps T. This image presents a top-down view of the entire trajectory as the agent searches for the target (a chair). The red box highlights the situation where the agent encounters a local trap during navigation.

Conversely, when T is excessively large, although the agent has access to more distant information, it is prone to miss some critical intermediate semantics which are closely related to target and are worth exploring, leading to erroneous long-range decisions, particularly in intricate environments. However, an optimally calibrated T can strike a delicate balance between exploration and perception, thereby facilitating to obtain impressive navigation performance.

We compared different backbones to evaluate their impacts on both the relative waypoint prediction and final navigation performances. Specifically, we modified the final output layers of the ResNet-18 and ViT to fit our dataset, allowing parameter updates during training. In contrast, DINOv2 and MAE models were connected to a five-layer MLP, with only the MLP parameters trained while freezing the backbone. The experiments were conducted on real RGB observations without resorting to the NVS model. Each variant was tested for 100 epochs. The results in Table 4 show that the ResNet-18, when trained from scratch, achieves the best performances in both relative wayopint prediction and ObjectNav while featuring a more lightweight architecture. Furthermore, the large performance gap between different backnones suggests the importance of the Where2Imagine module, indicating the usefulness of learning from human demonstrations. Please note that we observe a slight inconsistency between waypoint prediction loss and navigation success rate under backnones of DINOv2 and MAE. This might be because the waypoint prediction is evaluated on the testing data from MP3D dataset while the navigation is performed on HM3D. The differences between the datasets result in a certain degree of variability.

Table 3: **Where2Imagine:** the influence of sampling step **T** on navigation performance.

T	HM3D	
	Success Rate	**SPL**
8	51.0	25.1
10	64.0	29.4
11	64.0	28.3
12	59.0	30.0
15	59.0	26.8

Table 4: **Where2Imagine:** the impact of different **backbones**. Loss refers to the test loss of Where2Imagine. TFS: training from scratch, FT: fine-tuning.

Backbone	Params	Flops	Loss	HM3D	
				Success Rate	**SPL**
ResNet-18 (TFS)	11.4M	1.8G	0.12	64.0	28.3
ResNet-18 (FT)	11.4M	1.8G	0.24	61.0	29.7
ViT (TFS)	86.0M	16.9G	0.22	61.0	29.7
ViT (FT)	86.0M	16.9G	0.23	58.0	31.0
DINOv2	22.6M	5.5G	0.22	58.0	27.9
MAE	87.1M	4.4G	0.20	57.0	26.5

4.7 Analysis of VLM planner

We conducted a comparative evaluation of the effects of different VLMs on navigation performance, as detailed in Table 5. The experiments used real RGB without NVS model. The results demonstrate that GPT-4o-mini and GPT-4-Turbo exhibit negligible differences in success rate and SPL metrics, while showing a marked advantage over LLaVa, underscoring the significant role of advanced model reasoning capabilities in influencing experimental outcomes. Moreover, for models of the same architecture, it is possible to opt for more cost-effective variants without compromising navigation performance, thus enabling more resource-efficient and time-saving.

Table 5: Effect of different VLM.

VLM	HM3D	
	Success Rate	**SPL**
LLaVa	44.0	21.6
GPT-4-Turbo	63.0	29.4
GPT-4o-mini	64.0	28.3

5 Conclusion

We propose the ImagineNav framework, a mapless, open-vocabulary object navigation approach leveraging VLM. By incorporating imagination mechanism, the system effectively predicts future waypoints, transforming traditional navigation planning into a visual selection task for the VLM. Ablation studies highlight ImagineNav's strong potential for long-horizon navigation. Moving forward, we aim to enhance the quality of viewpoint generation and optimize the use of historical memory to further improve navigation performance and robustness.

6 ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 62273093).

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022. URL https://arxiv.org/abs/2204.01691.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents, 2018. URL https://arxiv.org/abs/1807.06757.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

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, 2023. URL https://arxiv.org/abs/2309.10309.

Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields, 2023. URL https://arxiv.org/abs/2212.02501.

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. International Conference on 3D Vision (3DV), 2017.

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.

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, pp. 24185–24198, 2024.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240): 1–113, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500, 2023.

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

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010, 2023.

Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation, 2022. URL https://arxiv.org/abs/2106.15648.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009, 2022.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation, 2023a. URL https://arxiv.org/abs/2210.05714.

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models, 2024. URL https://arxiv.org/abs/2403.08248.

Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model, 2023b. URL https://arxiv.org/abs/2305.11176.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022. URL https://arxiv.org/abs/2201.07207.

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, pp. 14829–14838, 2022.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation, 2023. URL https://arxiv.org/abs/2306.11290.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026, 2023.

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.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023. URL https://arxiv.org/abs/2209.07753.

Jing Liang, Peng Gao, Xuesu Xiao, Adarsh Jagan Sathyamoorthy, Mohamed Elnoor, Ming C. Lin, and Dinesh Manocha. Mtg: Mapless trajectory generator with traversability coverage for outdoor navigation, 2024. URL https://arxiv.org/abs/2309.08214.

Yiqing Liang, Boyuan Chen, and Shuran Song. Sscnav: Confidence-aware semantic scene completion for visual semantic navigation. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 13194–13200, 2021. doi: 10.1109/ICRA48506.2021.9560925.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024. URL https://arxiv.org/abs/2312.07533.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023.

Joel Loo, Zhanxin Wu, and David Hsu. Open scene graphs for open world object-goal navigation. arXiv preprint arXiv:2407.02473, 2024.

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.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024. URL https://arxiv.org/abs/2402.07872.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots, 2023. URL https://arxiv.org/abs/2310.13724.

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, pp. 8748–8763. PMLR, 2021.

Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation, 2020. URL https://arxiv.org/abs/2008.09285.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

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

Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale, 2022. URL https://arxiv.org/abs/2204.03514.

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17896–17906, 2023.

Pascal Roth, Julian Nubert, Fan Yang, Mayank Mittal, and Marco Hutter. Viplanner: Visual semantic imperative learning for local navigation, 2024. URL https://arxiv.org/abs/2310.00982.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. Zeronvs: Zero-shot 360-degree view synthesis from a single image, 2024. URL https://arxiv.org/abs/2310.17994.

Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In Conference on Robot Learning, pp. 2683–2699. PMLR, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

Anirudh Topiwala, Pranav Inani, and Abhishek Kathpal. Frontier based exploration for autonomous robot, 2018. URL https://arxiv.org/abs/1806.03581.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models, 2023. URL https://arxiv.org/abs/2303.17598.

Erik Wijmans, Irfan Essa, and Dhruv Batra. Ver: Scaling on-policy rl leads to the emergence of navigation in embodied rearrangement, 2022. URL https://arxiv.org/abs/2210.05064.

Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction, 2023. URL https://arxiv.org/abs/2301.07668.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3783–3795, 2024a.

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

Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation, 2022. URL https://arxiv.org/abs/2204.13226.

Karmesh Yadav, Jacob Krantz, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Jimmy Yang, Austin Wang, John Turner, Aaron Gokaslan, Vincent-Pierre Berges, Roozbeh Mootaghi, Oleksandr Maksymets, Angel X Chang, Manolis Savva, Alexander Clegg, Devendra Singh Chaplot, and Dhruv Batra. https://aihabitat.org/challenge/2023/, 2023.

Fan Yang, Chen Wang, Cesar Cadena, and Marco Hutter. iplanner: Imperative path planning, 2023. URL https://arxiv.org/abs/2302.11434.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation, 2023. URL https://arxiv.org/abs/2312.03275.

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

Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models, 2023b. URL https://arxiv.org/abs/2304.10700.

Jason J. Yu, Tristan Aumentado-Armstrong, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Polyoculus: Simultaneous multi-view image-based novel view synthesis, 2024. URL https://arxiv.org/abs/2402.17986.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language, 2022. URL https://arxiv.org/abs/2204.00598.

Albert J. Zhai and Shenlong Wang. Peanut: Predicting and navigating to unseen targets, 2022. URL https://arxiv.org/abs/2212.02497.

Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. Trihelper: Zero-shot object navigation with dynamic assistance. arXiv preprint arXiv:2403.15223, 2024a.

Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16414–16425, June 2024b.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation, 2023. URL https://arxiv.org/abs/2301.13166.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In European Conference on Computer Vision, pp. 350–368. Springer, 2022.

Minzhao Zhu, Binglei Zhao, and Tao Kong. Navigating to objects in unseen environments by distance prediction, 2022. URL https://arxiv.org/abs/2202.03735.

A APPENDIX

A.1 DECISION-MAKING DETAILS OF VISION-LANGUAGE MODELS

The process of inputting data into the VLM is as follows: First, the agent acquires an RGB observation based on its current pose and the Where2Imagine module predicts a relative pose based on the current observation. The RGB observation and the relative pose predicted by Where2Imagine are jointly input into the NVS to generate a new view image. This process is repeated at intervals of 60°, ultimately producing six novel view images, which are stitched together to serve as input to the VLM. In the ablation experiments, the VLM inputs for different variants differ in whether or not Where2Imagine is used for relative pose prediction and novel view generation, but at the start of each cycle, the initial input to the model is the RGB observation from the agent's current position at 60° intervals. Figure 7 shows the complete prompts and responses of the VLM.

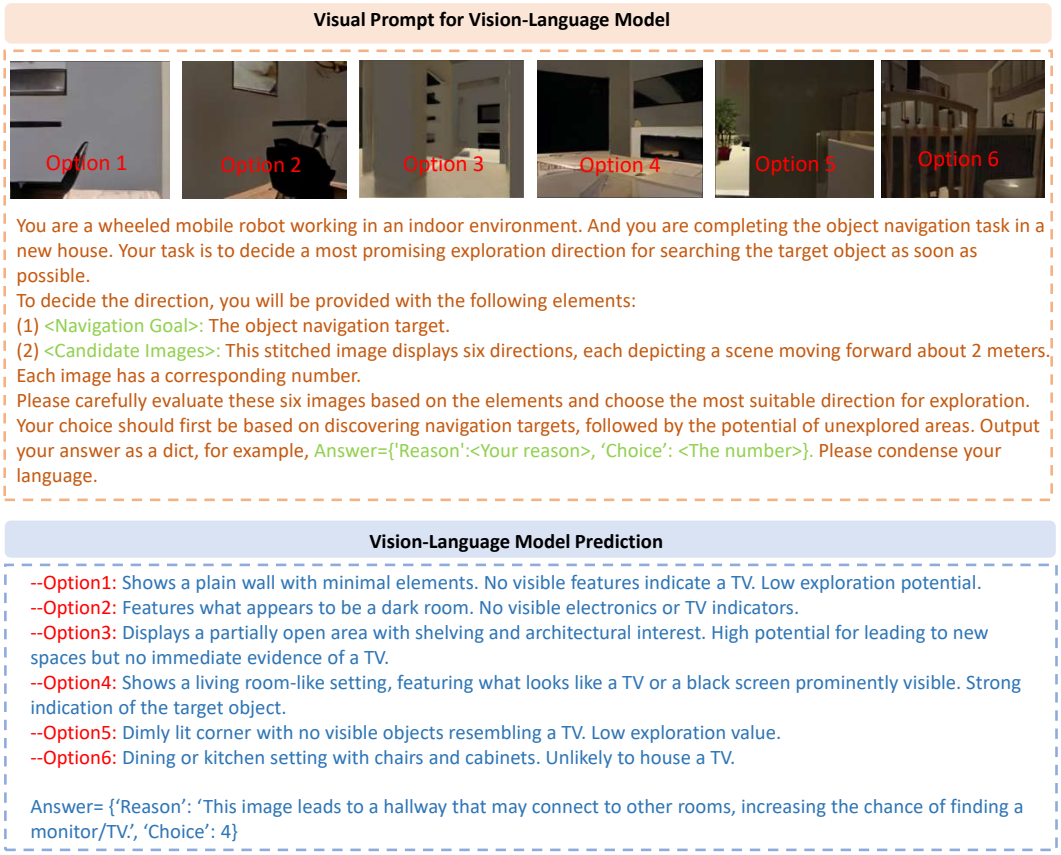

Figure 7: Complete prompt input and decision output of vision-language model.

A.2 COMPARISON BETWEEN THE WHERE2IMAGINE MODULE AND UNIFORM SAMPLING

As shown in Figure 8, the prediction points of the Where2Imagine module are more concentrated in walkable areas, and the predicted views are richer in information, covering more structural and semantic details such as furniture, doors, and windows within the room. In contrast, the r=2.0m sampling points are uniformly distributed based on a fixed radius, without considering obstacles or scene information. Where2Imagine's prediction tendency not only aligns better with realistic navigable paths but also effectively guides the agent toward areas with higher information density, improving its environmental perception efficiency.

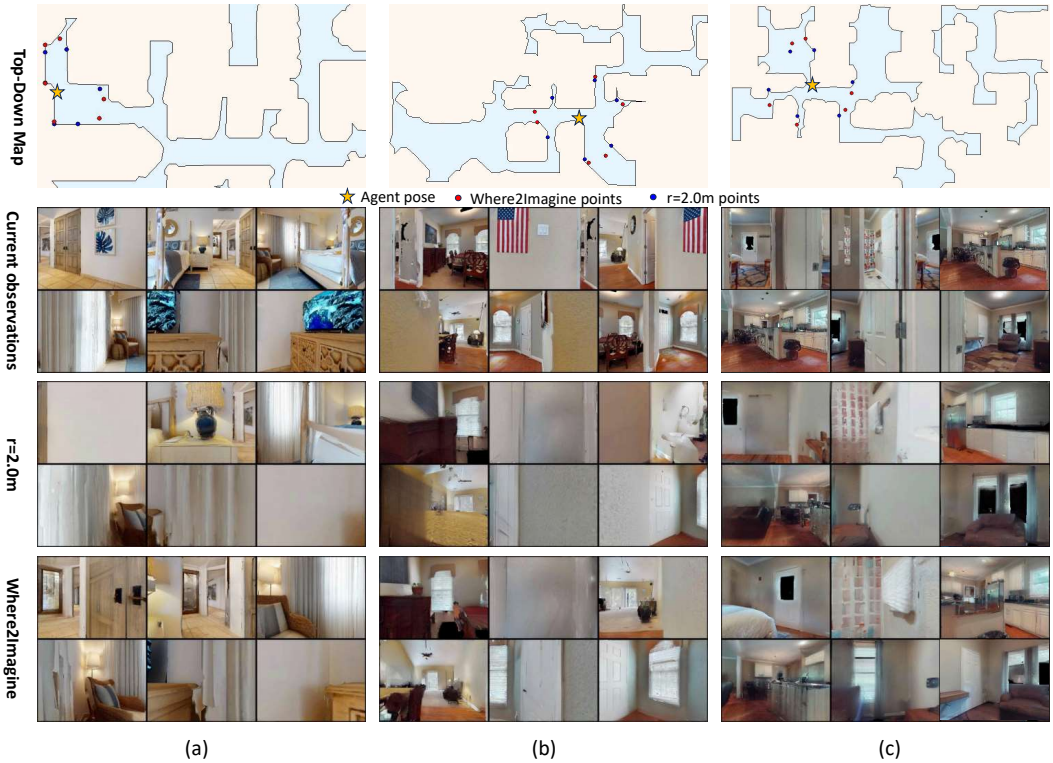

Figure 8: The visualization of the relative pose predicted by the Where2Imagine module and poses sampled at 60° intervals with 2.0m radius. The upper part shows the agent's current position (star marker) in different environments, as well as the distribution of the relative poses predicted by the Where2Imagine module (red dots) and the poses sampled at 60° intervals with a 2.0m radius (blue dots). The lower part shows the first-person views from different poses. Compared to uniform sampling at r=2.0m, the Where2Imagine module tends to predict more towards walkable areas and directions with higher information density.

A.3 Training Dataset for Where2Imagine

By replicating human demonstration trajectories, we collect first-person perspective images I_t from the trajectory and, after T frames, use the relative camera pose $P_{t+T} = (\Delta x, \Delta y, \theta)$ as image labels to generate training data for the Where2Imagine module, as shown in Figure 9. After training, the module is able to predict the next waypoint based on current observations, aligning with human navigation preferences.

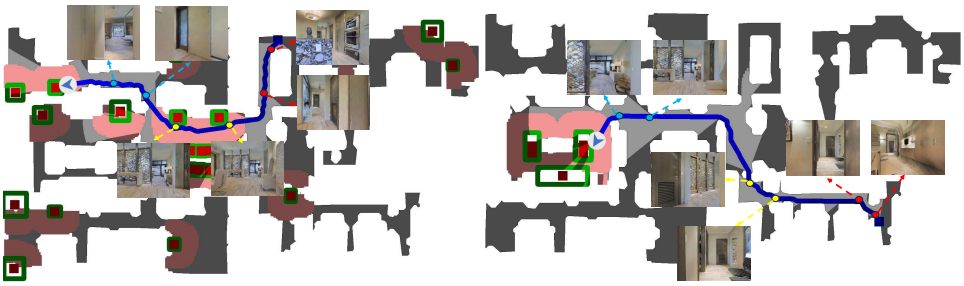

Figure 9: Visualization of human demonstration trajectories with different sampling intervals on the MP3D dataset in the Habitat-Web project. The left shows longer intervals, while the right shows relatively shorter intervals.

A.4 SELF-CENTERED VISUALIZATION OF WHERE2IMAGINE

Figure 10 presents the prediction results of the Where2Imagine model on two different datasets, HM3D and HSSD. Each subfigure consists of four parts: the input image, the predicted waypoint (red dot), the observation image and the corresponding NVS image. These visualizations demonstrate the model's adaptability across different datasets and its ability to guide the agent towards more exploratory positions while avoiding obstacles.

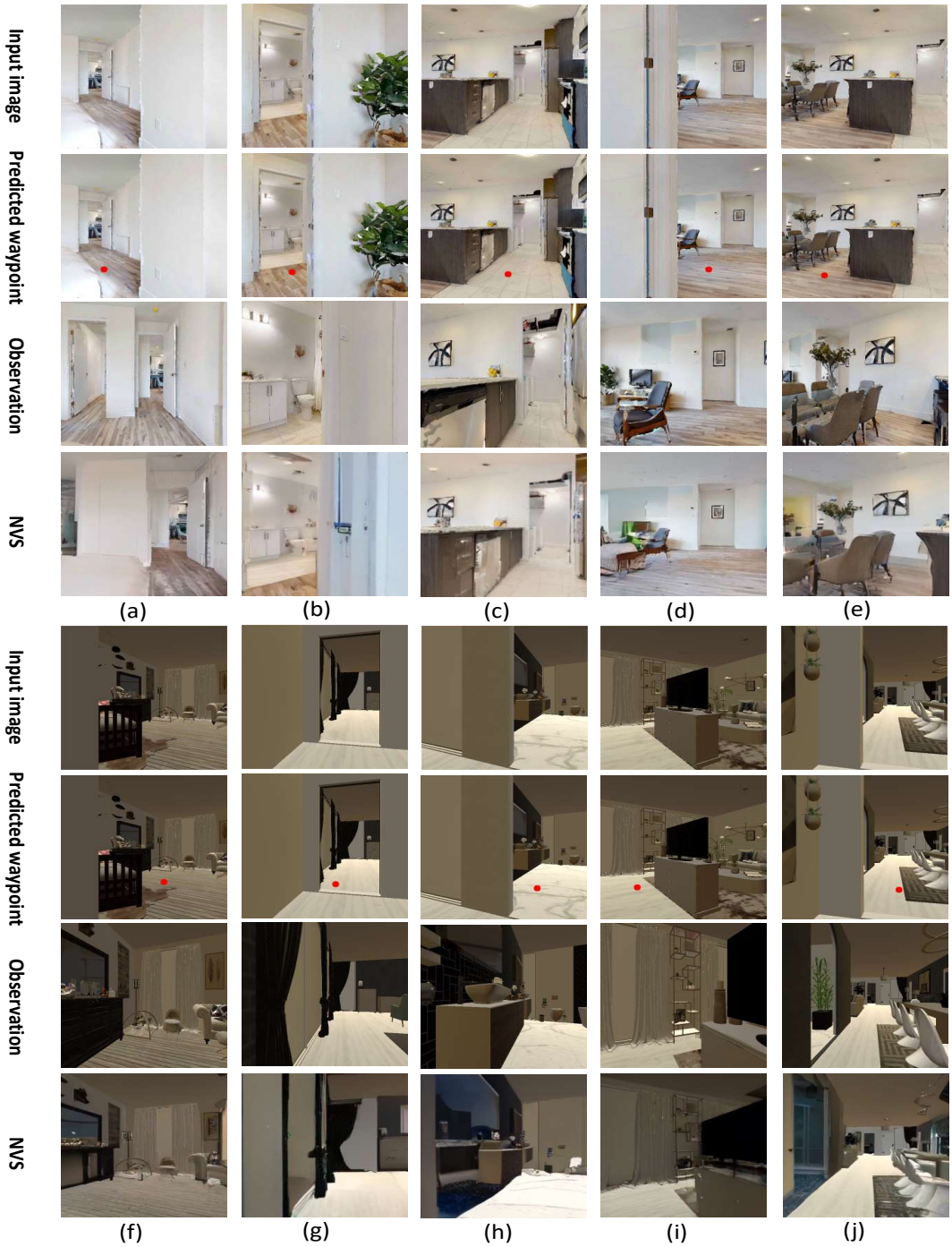

Figure 10: Self-centered visualization of the Where2Imagine model's prediction results. The top shows predictions on the HM3D dataset, while the bottom displays predictions on the HSSD dataset.