

FiLM-Nav: Efficient and Generalizable Navigation via VLM Fine-tuning

Naoki Yokoyama¹, Sehoon Ha¹

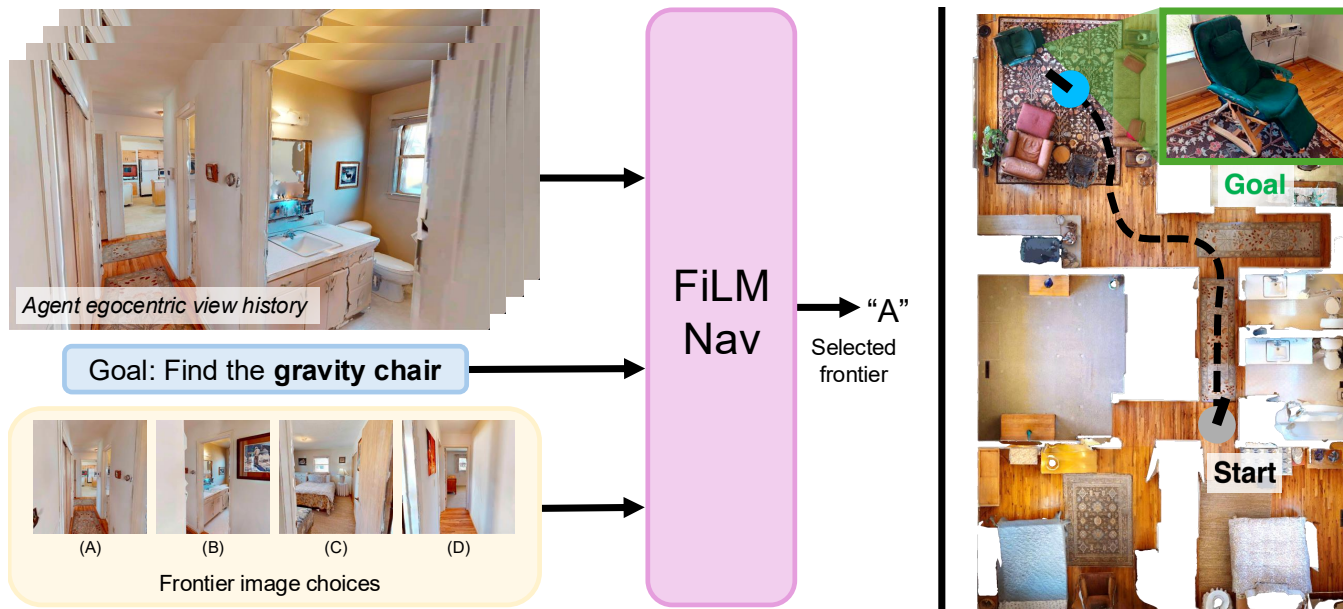


Fig. 1: FiLM-Nav takes the agent’s egocentric view history, a language goal (e.g., “Find the gravity chair”), and candidate frontier images shown on the left as input. The fine-tuned VLM policy then directly selects the next best frontier (“A”) to explore. This process leads to efficient navigation paths and successful localization of target objects, as demonstrated on the right, showcasing strong performance and generalization even for object categories unseen during fine-tuning by drawing upon the vast world knowledge embedded within the pre-trained foundation model.

Abstract—Enabling robotic assistants to navigate complex environments and locate objects described in free-form language is critical for real-world deployment. While foundation models, particularly Vision-Language Models (VLMs), offer powerful semantic understanding, effectively adapting their web-scale knowledge for embodied decision-making remains a key challenge. We present FiLM-Nav (Fine-tuned Language Model for Navigation), an approach that directly fine-tunes a pre-trained VLM as the navigation policy. In contrast to methods that use foundation models primarily in a zero-shot manner or for map annotation, FiLM-Nav learns to select the next best exploration frontier by conditioning directly on raw visual trajectory history and the navigation goal. Leveraging targeted simulated embodied experience allows the VLM to ground its powerful pre-trained representations in the specific dynamics and visual patterns relevant to goal-driven navigation. Critically, fine-tuning on a diverse data mixture combining ObjectNav, OVON, ImageNav, and an auxiliary spatial reasoning task proves essential for achieving robustness and broad generalization. FiLM-Nav sets a new state-of-the-art in both SPL and success rate on HM3D ObjectNav among open-vocabulary methods, and sets a state-of-the-art SPL on the challenging HM3D-OVON benchmark, demonstrating strong generalization to unseen object categories. Our work validates that directly fine-tuning VLMs on diverse simulated embodied data is a highly effective pathway towards generalizable and efficient semantic navigation capabilities.

¹NY and SH are with Georgia Institute of Technology
 nyokoyama@gatech.edu

I. INTRODUCTION

The development of intelligent embodied agents capable of operating autonomously in complex, human-centric environments remains a central challenge in robotics and artificial intelligence. A critical capability for such agents is semantic navigation: the ability to understand high-level goals, often specified in natural language, and navigate purposefully to find relevant objects or locations within previously unseen spaces. Object Goal Navigation (ObjectNav) [1] formalizes this challenge, tasking an agent with finding an instance of a specified object category (e.g., “find a chair”). The recent extension to Open-Vocabulary ObjectNav (OVON) [2] further raises the bar, requiring agents to generalize to potentially any object described using free-form language, significantly increasing the task’s complexity and real-world applicability.

Large-scale foundation models, particularly Vision-Language Models (VLMs) and Large Language Models (LLMs), have catalyzed significant progress in this domain. These models possess remarkable zero-shot reasoning and world knowledge capabilities [3], which researchers have leveraged for embodied navigation. Existing approaches often use foundation models for high-level planning, instruction interpretation, reasoning about object locations based on semantic cues, or enhancing environmental representations [4]–

[11]. However, these models, typically trained on disembodied web data, exhibit limitations in spatial-temporal reasoning and physical grounding crucial for efficient robotic task execution [12], [13]. Many methods maintain a separation between the foundation model and the low-level navigation policy, often relying on intermediate abstractions that may introduce bottlenecks. Effectively adapting these powerful models to the specific constraints of embodied decision-making remains a key challenge.

We introduce **Fine-tuned Language Model for Navigation (FiLM-Nav)**, an approach that directly fine-tunes a pre-trained VLM [14] to serve as the core navigation policy. Instead of employing foundation models primarily for high-level planning, we formulate navigation as a sequential decision-making problem where the VLM selects the next best exploration frontier. Critically, FiLM-Nav processes sequences of raw egocentric visual observations representing the agent’s trajectory history, bypassing the need for explicit semantic maps containing object labels or pre-computed features. This allows the VLM to implicitly learn relevant spatio-temporal representations and ground its powerful pre-trained knowledge in the specific dynamics relevant to goal-driven navigation. By fine-tuning on a diverse mixture of ObjectNav, OVON, and ImageNav [15] tasks, and conditioning the policy on past visual experience, FiLM-Nav effectively adapts the VLM’s pre-trained knowledge for embodied control, achieving richer grounding and avoiding potential information bottlenecks associated with intermediate representations.

Our primary contributions are: (1) the development of FiLM-Nav, demonstrating the effectiveness of directly fine-tuning a VLM as a frontier-based navigation policy using raw visual history; (2) state-of-the-art SPL and success performance on HM3D ObjectNav [16] among open-vocabulary methods, and state-of-the-art SPL on the challenging HM3D-OVON benchmark [2], demonstrating strong generalization to unseen object categories; and (3) an ablation study validating the crucial role of diverse task training for robust and efficient navigation performance.

II. RELATED WORK

Foundation models (VLMs and LLMs) are becoming increasingly leveraged for ObjectNav [1] and OVON [2]. Prior research integrating these models can be broadly categorized based on their role in the navigation pipeline. One major line of work utilizes foundation models primarily for high-level reasoning and planning. Methods like ESC [4], L3MVN [5], InstructNav [9], and WMNav [6] typically employ LLMs or VLMs to interpret complex instructions, infer likely object locations based on semantic knowledge (*e.g.*, “sofas are usually in living rooms”), or generate sequences of sub-goals or waypoints. These high-level plans are then often used to guide more traditional exploration or navigation policies (such as frontier-based exploration) operating on geometric or semantic maps. In these approaches, the foundation model generally remains separate from the low-level control policy and often relies on converting rich visual information into more abstracted representations (textual descriptions,

symbolic states, map annotations) suitable for the language model’s input space.

Another line of work focuses on enhancing the environmental representation itself using foundation models. VLFM, [7], TANGO [10], and VLN-Game [11] generate language-grounded value maps, with GAMap [8] also incorporating geometric parts and affordances. These methods seek richer representations than simple categorical maps that can be used to plan more efficient paths. Some mapless approaches also exist, such as ImagineNav [17], which uses imagined future views to guide a VLM.

In contrast to prior work, FiLM-Nav directly fine-tunes a VLM [14] as the navigation policy to learn to select frontiers using raw visual history collected from diverse embodied data (ObjectNav, OVON, ImageNav [15]). Directly learning the navigation policy by fine-tuning the VLM on embodied experience, rather than using the foundation model solely for zero-shot high-level reasoning or as a separate map annotation module, allows for better grounding of world knowledge in the specific embodied task context compared to zero-shot inference or map annotation approaches. And unlike UniNaVid [18], FiLM-Nav does so without learning an end-to-end policy that is conditioned on a particular embodiment or low-level action space.

III. PROBLEM FORMULATION

Our paper addresses ObjectNav, where an embodied agent, placed in a previously unseen environment, must navigate to find an instance of a specific object category.

Task Setup. We follow the standard ObjectNav and OVON setup within HM3D environments [16]. The agent uses egocentric RGB-D and an odometry sensor providing the agent’s pose relative to its starting pose in the episode. At each step, it selects from a discrete set of actions: (`move_forward(0.25m)`, `turn_left(30°)`, `turn_right(30°)`, `stop`). Given an object category specified via language (*e.g.*, “chair”, “stove”), the agent must call `stop` within 1m of a target instance before exceeding 500 steps.

Evaluation Metrics. We evaluate navigation performance using two standard metrics [19]: Success Rate (SR) and Success weighted by Path Length (SPL). SPL extends SR by incorporating path efficiency in addition to success, defined as:

$$\text{SPL} = \frac{1}{N} \sum_{i=1}^N S_i \frac{L_i}{\max(P_i, L_i)} \quad (1)$$

where S_i is the success indicator for episode i (0 or 1), L_i is the shortest path length, and P_i is the agent’s actual path length. By penalizing unnecessarily long trajectories, SPL provides a more comprehensive assessment of navigation quality that balances task completion with efficiency, which is critical for practical robotic deployment where time, energy, and computational resources are constrained. Consequently, SPL serves as the primary ranking metric in Habitat Challenge competitions [20]. Higher values indicate better performance for both metrics.

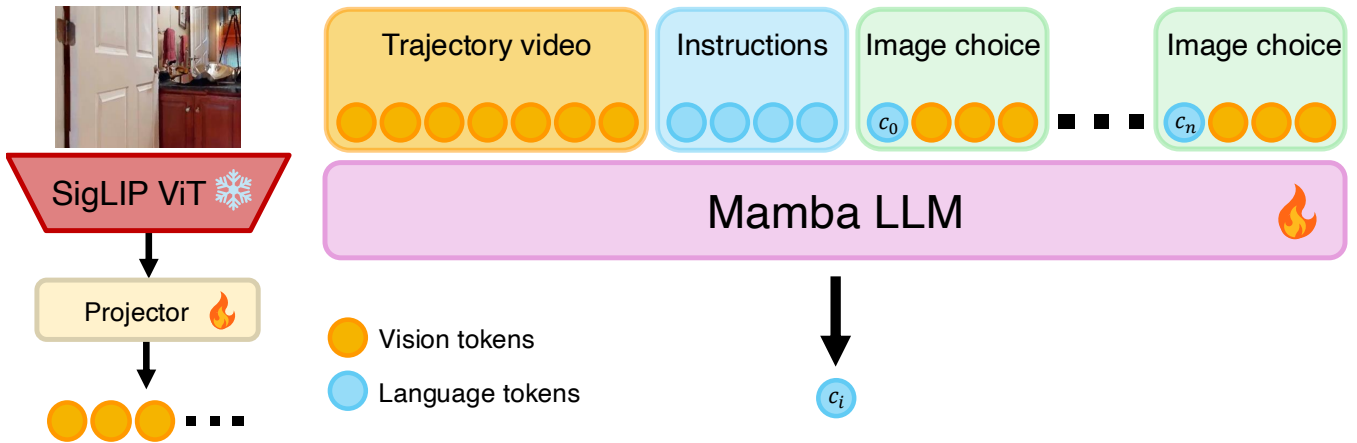


Fig. 2: FiLM-Nav architecture. *Left*: Input images are processed into vision tokens using a frozen SigLIP ViT and a trainable projector. *Right*: The LLM processes a sequence containing trajectory video tokens, language instructions, and image choices, each with vision tokens and a unique language token c_i . The LLM predicts the language token corresponding to the selected choice.

IV. METHOD: FINE-TUNED LANGUAGE MODEL FOR NAVIGATION (FiLM-NAV)

We introduce **Fine-tuned Language Model for Navigation** (FiLM-Nav), a novel approach that directly fine-tunes a pre-trained Vision-Language Model (VLM) to act as a high-level navigation policy for ObjectNav and OVON tasks. Instead of relying on zero-shot prompting or intermediate map representations, FiLM-Nav learns to select promising exploration frontiers based directly on raw visual history and language goals.

System Overview. At each timestep during navigation, FiLM-Nav takes the agent’s recent visual trajectory history, the natural language goal description, and visual representations of the current frontiers (boundaries between areas that have been explored and areas that have not been explored yet) as input. The core VLM processes this multi-modal sequence and outputs a choice corresponding to the frontier the agent should navigate towards. This high-level decision (frontier selection) passes a 2D waypoint to a low-level controller, which computes the next discrete action the robot should execute.

VLM Architecture and Input Processing. We utilize an efficient Mamba-based [21] VLM architecture, specifically the 2.8 billion parameter Cobra model pre-trained on SlimPajamas [22] and fine-tuned on web-scale image-text data [14]. Mamba’s State Space Model (SSM) architecture offers computational advantages over traditional transformers, particularly linear scaling of memory and computation with sequence length, making it well-suited for processing long sequences of vision tokens inherent in visual navigation tasks. In our empirical measurements, Mamba provides 60% faster inference (0.2s vs. 0.32s average) and 24% lower memory usage (12.5GB vs. 15.5GB for inference) compared to similarly-sized transformer models. Unlike transformers with growing KV-caches, Mamba maintains constant memory usage throughout inference, enabling more frequent replanning and extended operational periods critical for robotic deployment. Additionally, Cobra demonstrated

superior spatial reasoning capabilities on vision-language benchmarks [14], outperforming larger 7B+ transformer models despite having only 2.8B parameters, which is particularly relevant for navigation tasks requiring understanding of spatial relationships.

For multi-modal processing, we employ a pre-trained SigLIP ViT [23] as the image encoder. Visual features extracted by the frozen SigLIP encoder are projected into the LLM’s input embedding space using a trainable linear projection layer.

In order to handle long sequences of egocentric visual observations (trajectory history), we implement specific processing steps (Figure 2, left). Each frame is encoded into a grid of visual feature tokens. This grid undergoes average pooling across height, width, and temporal dimensions to create a compact representation before projection. For static images (e.g., frontier views, ImageNav goals), only height/width average pooling is applied before projection.

Input Sequence Formulation. At each timestep t , the FiLM-Nav model receives a structured input sequence composed of the following token types: (1) **Trajectory History Video Tokens**: Spatio-temporally pooled tokens from recent egocentric RGB history (up to 20 frames). (2) **Task Instruction Tokens**: Language tokens specifying the navigation goal. (3) **ImageNav Goal Image Tokens (Conditional)**: For ImageNav tasks only, height/width pooled vision tokens representing the goal image. (4) **Frontier Choice Tokens**: For each available frontier $k \in \{A, B, \dots, N\}$, a pair comprising: (a) A unique language token for the choice label (e.g., ‘A’), and (b) height/width pooled vision tokens for the representative RGB image (detailed next) of frontier k .

A conceptual representation of the input sequence is:

```
[<hist_video_tokens>] [<instr_tokens>]
[<opt_imgnav_goal>] [A] [<frontier_A.img>]
[B] [<frontier_B.img>] ... [N]
[<frontier_N.img>]
```

The model’s task is to predict the next token, which corresponds to the single language token (e.g., ‘A’) identifying the chosen frontier.

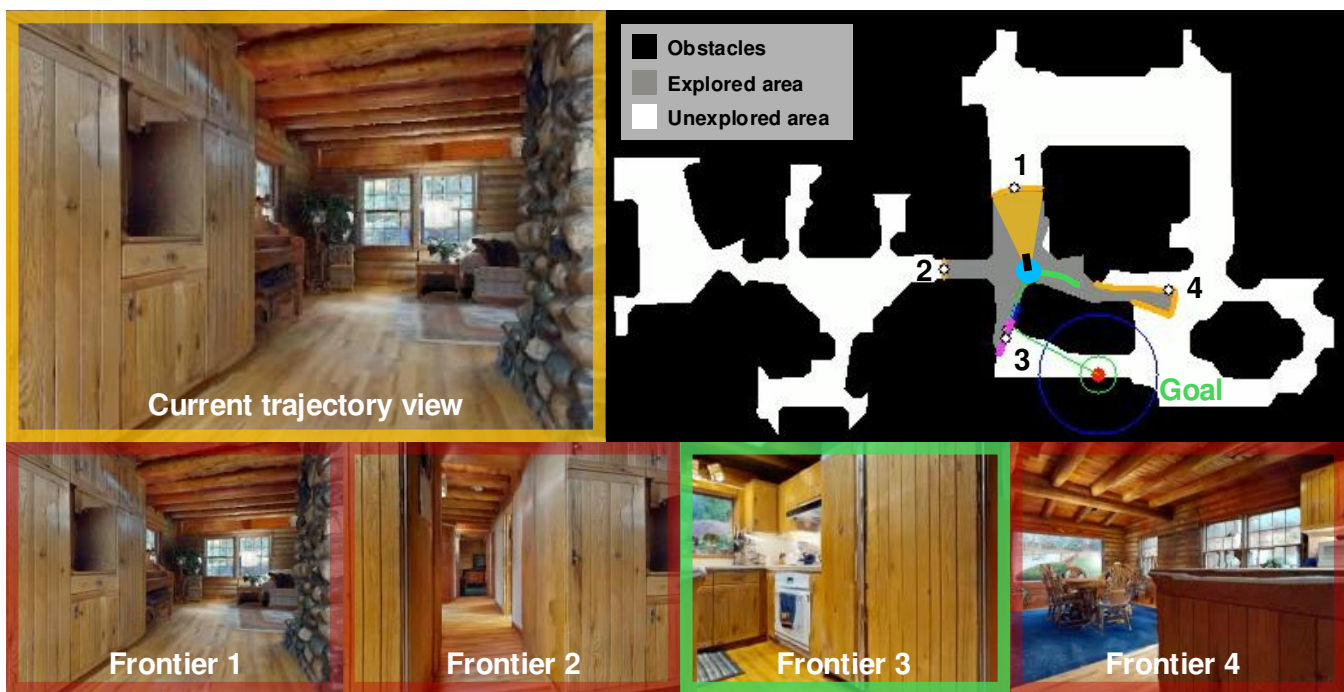


Fig. 3: Training data is generated via greedy frontier-based exploration. Each frontier is represented by a past RGB observation, selected from a pose offering optimal visibility towards the unexplored area. The frontier on the shortest path to the goal is recorded as the correct choice. FiLM-Nav uses these frontier images as goal-conditioned navigation options.

Frontier-Based Policy Learning. We frame the navigation problem as learning a policy $\pi(f_t|H_t, g)$ that selects the next frontier f_t to explore, conditioned on the trajectory history H_t (represented by pooled vision tokens) and the goal g (represented by language tokens, or potentially image tokens for ImageNav). The VLM directly embodies this policy, learning to output the token corresponding to the optimal frontier.

Navigation Task Training Data Generation. To fine-tune FiLM-Nav, we generate a large-scale dataset of supervised navigation examples using the Habitat simulator [24]. We collect trajectories for ObjectNav, OVON, and Image Goal Navigation (ImageNav) [15] within their respective training scene splits. In ImageNav, the agent must navigate to a location specified by a goal image rather than a language description. Incorporating ImageNav significantly broadens the environmental diversity experienced during fine-tuning; since object annotations are not required to generate image goals, the entire training set of 3D scenes from the HM3D dataset can be utilized (800 training scenes for ImageNav versus 145 for ObjectNav/OVON).

For all tasks, instead of using simple shortest paths, trajectories are generated using a greedy frontier-based exploration oracle [25] combined with goal-seeking behavior. This approach prevents the model from overfitting to overly direct shortest paths, encouraging it to learn exploration behaviors needed to find targets that are not initially visible, a known limitation when training solely on optimal paths [2], [26]. An occupancy map is built incrementally using depth and odometry, and frontiers (boundaries between explored and unexplored space) are identified. We design the greedy

oracle to normally navigate towards the closest frontier, but to then switch to navigating towards the frontier leading to the goal once a goal instance (or the ImageNav goal location) is nearby (3.5 meters).

At each timestep during trajectory generation, we record the agent’s egocentric RGB-D observation, the set of all currently identified frontiers, and the ‘correct’ frontier label (defined as the frontier lying on the shortest path to the closest goal instance).

Each frontier is visually represented by an RGB image selected from the agent’s history (Figure 3). Importantly, frontier observations are collected from the robot’s past trajectory history, *not from the frontier locations themselves*. The image chosen corresponds to the camera pose that maximizes perpendicularity between the camera’s view direction and the boundary constituting the frontier, providing the best view *toward* the unexplored area. This ensures no training-inference gap, as the agent only uses observations it has already captured during the current episode, with the same frontier representation process operating identically during both training and evaluation.

Auxiliary Spatial Reasoning Task. To explicitly enhance the model’s spatial reasoning, crucial for efficient navigation, we introduce an auxiliary training task. During trajectory generation within training environments, we collect the agent’s recent visual history (last 20 frames) and capture several candidate RGB images from nearby locations visited during the segment. At the trajectory’s end, the model receives: (1) its egocentric visual history as video tokens, (2) multiple candidate images with unique letter identifiers (A, B, C, ...), and (3) a multiple-choice prompt formatted as “Which

part of the environment is located at (X, Y) ?" where (X, Y) denotes a 2D location relative to the robot’s local frame at the current timestep. The model must predict the letter token corresponding to the correct candidate image, which is the image that taken from the given relative coordinates. This task forces the model to learn correspondences between first-person visual experience over time, and infer how different spaces are arranged in typical real-world indoor environments.

Fine-tuning Procedure. We fine-tune the pre-trained VLM using the combined dataset from ObjectNav, OVON, ImageNav, and the auxiliary spatial task. The training objective is the standard next-token prediction loss (cross-entropy) commonly used for LLMs. The model is trained to predict the single token corresponding to the correct frontier choice (or the correct image choice for the auxiliary task). During fine-tuning, the weights of the SigLIP ViT image encoder are kept frozen while we train the linear projection layer and the LLM weights.

We use the AdamW optimizer with learning rate 2×10^{-5} , global batch size of 576 (achieved through distributed training and gradient accumulation), and train for 3 epochs on 8 A40 GPUs. The combined dataset comprises approximately 26k samples each for ObjectNav and OVON, 40k for ImageNav, and 30k for the auxiliary spatial task. Training takes around 15 hours total.

Online Deployment within Habitat. During evaluation, FiLM-Nav operates in a closed loop with the following components: (1) *Mapping and Frontier Extraction*: An occupancy map is updated using depth and odometry; frontiers are extracted as boundaries between explored/unexplored areas. (2) *Frontier Representation*: Representative RGB images for each frontier are selected from the observation history by choosing past camera poses that maximize perpendicularity with frontier boundaries. (3) *VLM Policy Inference*: The trajectory history, language goal, and frontier images are formatted into the input sequence; the fine-tuned VLM predicts the optimal frontier token. (4) *Low-Level Control*: The coordinates of the chosen frontier are passed to a pre-trained PointNav policy [27] to predict the next discrete action to take. (5) *Object Detection*: similar to [7], we use OWLViTv2 [28] to detect the goal object, and upon detection, MobileSAM [29] is used to segment the object’s point cloud from the depth image. The PointNav policy takes as input the point on the cloud that is closest to the agent, and `stop` is called once the agent is within the success distance.

V. EXPERIMENTS

In this section, we detail the experimental setup designed to evaluate the performance of FiLM-Nav and validate our contributions. We aim to answer the following key research questions:

- 1) How does FiLM-Nav perform on standard HM3D ObjectNav benchmarks compared to other state-of-the-art methods capable of open-vocabulary navigation?
- 2) How effectively does FiLM-Nav generalize to the challenges of Open-Vocabulary ObjectNav (OVON),

TABLE I: Performance on the HM3D-v0.1 ObjectNav validation set. Methods that can only support the six object categories present in the HM3D ObjectNav benchmark are grayed out. FiLM-Nav achieves the highest SPL and the highest SR among methods that are open-vocabulary.

Method	Open-vocab?	SR (\uparrow)	SPL (\uparrow)
OVRLv2 [30]	\times	64.7%	28.1%
PIRLNav [26]	\times	70.4%	34.1%
GAMap [8]	\checkmark	53.1%	26.0%
VLN-Game [11]	\checkmark	61.3%	26.8%
VLFM [7]	\checkmark	52.5%	30.4%
WMNav [6]	\checkmark	58.1%	31.2%
FiLM-Nav (Ours)	\checkmark	61.7%	37.3%

particularly regarding unseen object categories and synonyms?

- 3) What are the contributions of the different components of our training strategy (diverse task mixture, auxiliary spatial task) to the overall performance?

Experimental Setup. All experiments are conducted using the Habitat simulator [24] with 3D scans from the HM3D dataset [16]. We follow the standard ObjectNav task setup as described in Section III.

Tasks and Benchmarks. We evaluate FiLM-Nav on standard benchmarks: **HM3D ObjectNav**, which uses a predefined set of six object categories, using the validation sets of HM3D-v0.1 [16] and the improved HM3D-v0.2; and **HM3D-OVON** [2], which tests generalization via its validation splits: VAL SEEN (categories seen during OVON training), VAL SEEN SYNONYMS (unseen synonyms/neighbors of seen categories), and VAL UNSEEN (semantically distinct unseen categories). Performance on the OVON unseen splits, particularly VAL SEEN SYNONYMS and VAL UNSEEN, directly measures generalization to novel language descriptions of objects.

Baselines. We compare FiLM-Nav against recent state-of-the-art methods. For **HM3D ObjectNav**, baselines include GAMap [8], VLN-Game [11], VLFM [7], WMNav [6], ImagineNav [17], InstructNav [9], and Uni-NaVid [18] (selecting strongest open-vocabulary methods available on respective benchmark versions v0.1/v0.2). For **HM3D-OVON**, baselines are VLFM [7], DAGRL+OD [2], and TANGO [10].

VI. RESULTS AND ANALYSIS

In this section, we present the empirical evaluation of FiLM-Nav (**F**ine-tuned **L**anguage **M**odel for **N**avigation).

Performance on HM3D ObjectNav. We first evaluate FiLM-Nav on the standard HM3D ObjectNav validation sets, comparing against state-of-the-art methods, with a focus on those that are also capable of open-vocabulary navigation (versus those that only support the six categories that appear in the HM3D ObjectNav dataset). Table I shows the results on the HM3D-v0.1 benchmark, and Table II presents the results on the improved HM3D-v0.2 benchmark.

TABLE II: Performance on the HM3D-v0.2 ObjectNav validation set. FiLM-Nav sets a new state-of-the-art in both SR and SPL.

Method	Open-vocab?	SR (\uparrow)	SPL (\uparrow)
ImagineNav [17]	✓	53.0%	23.8%
InstructNav [9]	✓	58.0%	20.9%
VLFM [7]	✓	62.6%	31.0%
WMNav [6]	✓	72.2%	33.3%
Uni-NaVid [18]	✓	73.7%	37.1%
FiLM-Nav (Ours)	✓	77.0%	41.3%

On the HM3D-v0.1 benchmark (Table I), FiLM-Nav achieves the highest Success weighted by Path Length (SPL) score of 37.3%. This significantly surpasses the previous best SPL among compared open-vocabulary methods (WMNav [6] with 31.2%), representing an absolute improvement of 6.1% and a relative improvement of 19.6%. Notably, it also surpasses PIRLNav [26] (34.1% SPL) by 3.2% absolute and 9.4% relative, despite PIRLNav being trained using nearly 2,400 hours of human demonstrations collected within Habitat specifically for the six standard ObjectNav categories. These results highlight the superior path efficiency achieved by directly fine-tuning the VLM as a policy. Our method also achieves a state-of-the-art success rate (SR) of 61.7% among open-vocabulary methods.

It is worth noting that our current frontier-based mapping implementation operates in 2D, and cannot handle stair traversal due to the lack of z-coordinate information in the sensor configuration. On the other hand, methods like PIRLNav and OVRL learn end-to-end policies capable of multi-floor navigation, as they are directly trained to output discrete actions to search for the six goal object categories used by the benchmark without constructing maps. This accounts for failures in 13.9% of HM3D-v0.1 episodes that require floor transitions to go from the start position to an instance of the goal category. Despite this limitation and our reliance on a pre-trained object detector (versus end-to-end object recognition training that adapts to HM3D-specific visual characteristics), FiLM-Nav’s efficient path planning enables it to achieve superior SPL scores.

Although FiLM-Nav has a lower SR than PIRLNav (61.7% vs. 70.4%), it achieves a higher SPL score (37.3% vs. 34.1%) due to its exceptional path efficiency. Crucially, SPL measures “efficient success”; since unsuccessful episodes contribute 0 to SPL, achieving *higher* SPL despite *lower* SR means *FiLM-Nav’s successful episodes are dramatically more efficient*. This demonstrates that when FiLM-Nav successfully reaches the goal, it does so via substantially shorter paths than competing methods.

On the HM3D-v0.2 benchmark (Table II), FiLM-Nav establishes new state-of-the-art performance in both metrics, achieving 77.0% SR and 41.3% SPL. It outperforms the prior leading method, Uni-NaVid [18] (73.7% SR / 37.1% SPL), by 3.3% absolute SR and 4.2% absolute SPL (an 11.3% relative SPL improvement). The strong performance on both benchmarks, particularly the substantial gains in SPL,

TABLE III: Performance comparison on the HM3D-OVON validation splits. FiLM-Nav achieves state-of-the-art performance in most metrics across the splits, with the highest SPL values throughout.

Method	VAL SEEN		SYNONYMS		VAL UNSEEN	
	SR \uparrow	SPL \uparrow	SR \uparrow	SPL \uparrow	SR \uparrow	SPL \uparrow
VLFM [7]	38.1	20.8	37.7	21.2	38.5	22.2
DAgRL+OD [2]	38.5	21.1	39.0	21.4	37.1	19.9
TANGO [10]	-	-	-	-	35.5	19.5
Uni-NaVid [18]	41.3	21.1	43.9	21.8	39.5	19.8
FiLM-Nav (Ours)	44.9	24.5	40.1	23.1	40.8	24.4

demonstrates the effectiveness of our fine-tuning approach for generating efficient navigation behaviors in standard ObjectNav tasks.

Performance on Open-Vocabulary ObjectNav (OVON).

Next, we evaluate FiLM-Nav on the more challenging HM3D-OVON benchmark [2], designed to test generalization to a wide range of object categories beyond those seen during standard ObjectNav training. Table III presents the results across the three standard OVON validation splits: VAL SEEN (seen categories), VAL SEEN SYNONYMS (unseen synonyms/close categories), and VAL UNSEEN (unseen distinct categories).

Notably, FiLM-Nav achieves state-of-the-art SPL on all three splits of the OVON dataset, and state-of-the-art SR on two of the three. On the VAL SEEN split, FiLM-Nav obtains 44.9% SR and 24.5% SPL, outperforming the prior best (Uni-NaVid [18]) by a significant margin (+3.6% absolute SR, +3.4% absolute SPL). On the VAL SEEN SYNONYMS split, which tests robustness to linguistic variations for known object types, Uni-NaVid achieves the highest SR at 43.9%, while FiLM-Nav achieves the second highest SR at 40.1% and maintains the best SPL at 23.1% compared to Uni-NaVid’s 21.8% SPL. Crucially, on the VAL UNSEEN split, which directly measures generalization to object categories unseen and unrelated to those seen during navigation training, FiLM-Nav achieves 40.8% SR and 24.4% SPL. This substantially surpasses the previous best results on this challenging split (Uni-NaVid [18] with 39.5% SR / 19.8% SPL, and VLFM [7] with 38.5% SR / 22.2% SPL). This strong performance on unseen categories underscores the benefit of fine-tuning a large pre-trained VLM; the model effectively leverages its broad world knowledge acquired during pre-training and adapts it through our navigation-focused fine-tuning process to successfully locate objects it has never been explicitly trained to navigate towards, suggesting that direct grounding via fine-tuning allows for more effective transfer of pre-trained knowledge to novel embodied scenarios compared to zero-shot methods or those relying on intermediate representations. Overall, the results on OVON demonstrate that FiLM-Nav excels in standard ObjectNav and possesses strong generalization capabilities for open-vocabulary scenarios, achieving efficient and successful navigation for a diverse and extensive range of object goals described in language, particularly showing strength in SPL metrics across all splits.

Ablation Study. To understand the contributions of the different components of our training strategy, we perform

TABLE IV: Ablation study evaluating contribution of training data sources (ON=ObjectNav, IN=ImageNav, OV=OVON) and the auxiliary task (AX) on HM3D-v0.2 and HM3D-OVON validation sets. Full model (bottom row) performs best.

Training data				HM3D v0.2		VAL SEEN		SYNO-NYMS		VAL UNSEEN	
ON	IN	OV	AX	SR	SPL	SR	SPL	SR	SPL	SR	SPL
✓	-	-	-	74.3	38.8	41.7	22.0	38.8	21.6	39.4	22.5
✓	✓	-	-	74.9	39.1	43.1	22.8	39.3	21.7	41.5	23.8
✓	✓	✓	-	74.9	39.2	43.3	23.3	39.9	22.4	40.8	23.6
✓	✓	✓	✓	77.0	41.3	44.9	24.5	40.1	23.1	40.8	24.4

an ablation study investigating the impact of the training data mixture (ObjectNav, ImageNav, OVON datasets) and our proposed auxiliary spatial reasoning task. The results are presented in Table IV, evaluated on HM3D-v0.2 and HM3D-OVON.

We start with a baseline model trained solely on the standard ObjectNav dataset (Row 1). This model achieves strong performance on HM3D (74.3% SR / 38.8% SPL) and exhibits considerable zero-shot generalization to OVON splits (*e.g.*, 39.4% SR / 22.5% SPL on VAL UNSEEN), again highlighting the benefit of the VLM’s pre-trained knowledge in spite of the limited amount of categories encountered during navigation training.

Adding ImageNav training data (Row 2 vs. Row 1), which exposes the model to more diverse navigation scenarios across 800 scenes without object labels (vs. only 145 train scenes with labels), provides slight improvements, particularly boosting OVON generalization (*e.g.*, VAL UNSEEN SPL increases from 22.5% to 23.8%). This improvement likely stems from ImageNav exposing the model to a vastly more diverse set of goal representations during training, since each unique goal image is projected into a unique language embedding using the SigLIP encoder and the projection layer, compared to the small, fixed set of embeddings corresponding to the 6 standard ObjectNav categories, thus enhancing the model’s ability to generalize its goal-conditioning mechanism to the diverse set of possible goal object categories of OVON.

Incorporating OVON training data (Row 3 vs. Row 2), which includes explicit navigation examples for a wider variety of object categories, further improves performance on OVON splits (VAL SEEN SPL from 22.8% to 23.3%, VAL SEEN SYNONYMS SPL from 21.7% to 22.4%) while maintaining HM3D performance. This demonstrates the value of training on target distributions closer to the open-vocabulary evaluation.

Finally, incorporating our auxiliary spatial reasoning task (Row 4 vs. Row 3) provides further consistent gains, completing the full FiLM-Nav model. Performance improves across almost all metrics with a jump of up to 2.1% in absolute SPL and 2.1% in absolute SR (see Table IV). This suggests that enhancing the model’s ability to correlate visual history with spatial layout aids efficient frontier selection and contributes positively to overall navigation performance.

In summary, the ablation study confirms that leveraging a diverse mixture of navigation tasks (ObjectNav, ImageNav,

OVON) and incorporating the targeted auxiliary spatial reasoning task are crucial components that contribute to the state-of-the-art performance of FiLM-Nav.

Real-World Transferability. While our primary experiments utilize simulation for systematic evaluation, we provide evidence supporting transferability to real-world settings. First, multiple prior works have demonstrated successful *zero-shot* sim-to-real transfer using models trained on the same HM3D dataset in Habitat [18], [31], validating the photorealistic simulation’s effectiveness for real-world deployment. Second, FiLM-Nav operates at the frontier selection level of abstraction, avoiding low-level control details that typically cause sim-to-real failures in robotics. The model processes RGB images to select exploration directions, a capability that transfers more readily than precise motor control. Third, our use of a pre-trained VLM foundation provides inherent robustness to visual domain shift, as these models are trained on a web-scale dataset of diverse real-world imagery.

To empirically validate transferability, we conducted preliminary real-world evaluation: (1) We collected egocentric videos in a real apartment, with the camera trajectory capturing potential frontier directions. FiLM-Nav successfully selected correct frontiers toward goal objects when presented with these real-world observations. (2) We evaluated on randomly sampled POV tours from the RealEstate10K dataset [32], labeling frames as frontiers and annotating whether they lead toward goal objects. The model demonstrated effective frontier selection on this real-world residential imagery. While deployment and configuration on a real mobile robot hardware platform remains to be future work, these results indicate that our approach can process and make decisions on real-world visual input effectively. Additional videos and evaluation details are available at <https://filmnav.netlify.app>.

VII. CONCLUSION

This paper introduced FiLM-Nav, a novel approach that demonstrates the efficacy of directly fine-tuning a pre-trained VLM [14] as the core navigation policy for Object Goal Navigation (ObjectNav) and its open-vocabulary extension (OVON). Departing from methods that use foundation models primarily for zero-shot reasoning or map enhancement, FiLM-Nav formulates navigation as selecting the optimal exploration frontier conditioned directly on raw visual trajectory history and the language goal. Fine-tuning on a diverse mixture of simulated embodied tasks (ObjectNav, OVON, ImageNav) combined with an auxiliary spatial reasoning task effectively grounds the VLM’s powerful pre-trained representations in the spatio-temporal context required for efficient, goal-directed navigation. Our experiments validate this strategy, showing that FiLM-Nav achieves state-of-the-art navigation efficiency (SPL) on standard HM3D ObjectNav benchmarks and establishes a new state-of-the-art in SPL across all validation splits of the demanding HM3D-OVON benchmark, which highlights its strong generalization to novel object categories.

While FiLM-Nav demonstrates strong performance, several limitations present opportunities for future work. The current framework targets finding *any* instance of a specified category rather than specific instances with distinguishing properties, and focuses on single-object search tasks rather than complex multi-object instructions. Additionally, the system’s reliance on an external object detector for termination and explicit occupancy mapping for frontier generation introduces dependencies that could be addressed through end-to-end learning. Despite these limitations, our work validates that directly fine-tuning VLMs on diverse simulated embodied data is a highly effective pathway toward generalizable and efficient semantic navigation capabilities, providing a foundation for developing even more versatile embodied agents.

VIII. ACKNOWLEDGEMENTS

This work was supported by the Korea Evaluation Institute of Industrial Technology (KEIT) funded by the Korea Government (MOTIE) under Grant No.20018216, Development of Mobile Intelligence SW for autonomous navigation of legged robots in dynamic and atypical environments for real application.

REFERENCES

- [1] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijnmans, “Objectnav revisited: On evaluation of embodied agents navigating to objects,” 2020.
- [2] N. Yokoyama, R. Ramrakhya, A. Das, D. Batra, and S. Ha, “Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.
- [3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021.
- [4] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, “ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation,” *arXiv preprint arXiv:2301.13166*, 2023.
- [5] B. Yu, H. Kasaei, and M. Cao, “L3mvm: Leveraging large language models for visual target navigation,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.
- [6] D. Nie, X. Guo, Y. Duan, R. Zhang, and L. Chen, “Wmnav: Integrating vision-language models into world models for object goal navigation,” *arXiv preprint arXiv:2503.02247*, 2025.
- [7] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, “Vlfn: Vision-language frontier maps for zero-shot semantic navigation,” in *International Conference on Robotics and Automation (ICRA)*, 2024.
- [8] S. Yuan, H. Huang, Y. Hao, C. Wen, A. Tzes, and Y. Fang, “Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance,” in *Advances in Neural Information Processing Systems*, ser. NeurIPS 2024, 2024.
- [9] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, “Instructnav: Zero-shot system for generic instruction navigation in unexplored environment,” in *Conference on Robot Learning*, ser. CoRL 2024, 2024.
- [10] F. Ziliotto, T. Campari, L. Serafini, and L. Ballan, “TANGO: Training-free embodied AI agents for open-world tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [Online]. Available: <https://arxiv.org/abs/2412.10402>
- [11] B. Yu, Y. Liu, L. Han, H. Kasaei, T. Li, and M. Cao, “Vln-game: Vision-language equilibrium search for zero-shot semantic navigation,” *arXiv preprint arXiv:2411.11609*, 2024.
- [12] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, “Spatialvlm: Endowing vision-language models with spatial reasoning capabilities,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 455–14 465.
- [13] Y. Liu, W. Chen, Y. Bai, X. Liang, G. Li, W. Gao, and L. Lin, “Aligning cyber space with physical world: A comprehensive survey on embodied ai,” *arXiv preprint arXiv:2407.06886*, 2024.
- [14] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, “Cobra: Extending mamba to multi-modal large language model for efficient inference,” in *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. AAAI Press, 2025.
- [15] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, “Target-driven visual navigation in indoor scenes using deep reinforcement learning,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
- [16] S. K. Ramakrishnan, A. Gokaslan, E. Wijnmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. 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*, 2021. [Online]. Available: <https://arxiv.org/abs/2109.08238>
- [17] X. Zhao, W. Cai, L. Tang, and T. Wang, “Imaginenav: Prompting vision-language models as embodied navigator through scene imagination,” in *International Conference on Learning Representations*, ser. ICLR 2025, 2025.
- [18] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, “Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks,” in *Robotics: Science and Systems*, 2025.
- [19] P. Anderson, A. X. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir, “On Evaluation of Embodied Navigation Agents,” *arXiv preprint arXiv:1807.06757*, 2018.
- [20] K. Yadav, J. Krantz, R. Ramrakhya, S. K. Ramakrishnan, J. Yang, A. Wang, J. Turner, A. Gokaslan, V.-P. Berges, R. Mootaghi, O. Maksymets, A. X. Chang, M. Savva, A. Clegg, D. S. Chaplot, and D. Batra, “Habitat challenge 2023,” <https://aihabitat.org/challenge/2023/>, 2023.
- [21] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.
- [22] D. Soboleva, F. Al-Khateeb, R. Myers, J. R. Steeves, J. Hestness, and N. Dey, “SlimPajama: A 627B token cleaned and deduplicated version of RedPajama,” <https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama>, 2023. [Online]. Available: <https://huggingface.co/datasets/cerebras/SlimPajama-627B>
- [23] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.
- [24] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijnmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, “Habitat: A Platform for Embodied AI Research,” in *ICCV*, 2019.
- [25] B. Yamauchi, “A Frontier-Based Approach for Autonomous Exploration,” in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA’97: Towards New Computational Principles for Robotics and Automation’*. IEEE, 1997, pp. 146–151.
- [26] R. Ramrakhya, D. Batra, E. Wijnmans, and A. Das, “PIRLNav: Pretraining with Imitation and RL Finetuning for ObjectNav,” in *CVPR*, 2023.
- [27] E. Wijnmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, “Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames,” *arXiv preprint arXiv:1911.00357*, 2019.
- [28] N. H. Matthias Minderer, Alexey Gritsenko, “Scaling open-vocabulary object detection,” *NeurIPS*, 2023.
- [29] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, “Faster Segment Anything: Towards Lightweight SAM for Mobile Applications,” *arXiv preprint arXiv:2306.14289*, 2023.
- [30] K. Yadav, A. Majumdar, R. Ramrakhya, N. Yokoyama, A. Baevski, Z. Kira, O. Maksymets, and D. Batra, “Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav,” *arXiv preprint arXiv:2303.07798*, 2023.
- [31] T. Silwal, A. Guo, K. Narayanan, and S. Karaman, “Learn to navigate in dynamic environments with normalized lidar scans,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- [32] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, “The realestate10k dataset for video prediction and beyond,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.