Move to Understand a 3D Scene: Bridging Visual Grounding and Exploration for Efficient and Versatile Embodied Navigation

Ziyu Zhu[1,2*†], Xilin Wang[2,4], Yixuan Li[2,3], Zhuofan Zhang[1,2], Xiaojian Ma[2], Yixin Chen[2], Baoxiong Jia[2], Wei Liang[3], Qian Yu[4], Zhidong Deng[1†✉], Siyuan Huang[2✉], Qing Li[2✉]
[1]Tsinghua University [3]Beijing Institute of Technology [4]Beihang University
[2]State Key Laboratory of General Artificial Intelligence, BIGAI, China

Project page: mtu3d.github.io

Figure 1. **MTU3D** is a versatile embodied navigation model capable of processing diverse inputs, including *object categories*, *image snapshots*, *natural language descriptions*, *task plan sequences*, and *questions*. It iteratively explores the environment and performs visual grounding to reach the target location. Through large-scale **V**ision-**L**anguage-**E**xploration pre-training, it achieves state-of-the-art performance across various benchmarks including open-vocabulary, multi-modal lifelong, and sequential navigation.

Abstract

*Embodied scene understanding requires not only comprehending visual-spatial information that has been observed but also determining where to explore next in the 3D physical world. Existing 3D Vision-Language (3D-VL) models primarily focus on grounding objects in static observations from 3D reconstruction, such as meshes and point clouds, but lack the ability to actively perceive and explore their environment. To address this limitation, we introduce **M**ove **t**o **U**nderstand (***MTU3D***), a unified framework that integrates active perception with **3D** vision-language learning, enabling embodied agents to effectively explore and understand their environment. This is achieved by three key innovations: 1) Online query-based representation learning, enabling direct spatial memory construction from RGB-D frames, eliminating the need for explicit 3D reconstruction. 2) A unified objective for grounding and exploring, which represents unexplored locations as frontier queries and jointly optimizes object grounding and frontier selection. 3) End-to-end trajectory learning that combines **V**ision-**L**anguage-**E**xploration pre-training over a million diverse trajectories collected from both simulated and real-world RGB-D sequences. Extensive evaluations across various embodied navigation and question-answering benchmarks show that MTU3D outperforms state-of-the-art reinforcement learning and modular navigation approaches by 14%, 23%, 9%, and 2% in success rate on HM3D-OVON, GOAT-Bench, SG3D, and A-EQA, respectively.* MTU3D*'s versatility enables navigation using diverse input modalities, including categories, language descriptions, and reference images. The deployment on a real robot demonstrates* MTU3D*'s effectiveness in handling real-world data. These findings highlight the importance of bridging visual grounding and exploration for embodied intelligence.*

1. Introduction

Embodied scene understanding requires not only recognizing observed objects but also actively exploring and reasoning in the 3D physical world [8, 18, 47, 93]. Imagine stepping into an unfamiliar room with the goal of ***"find something to eat"***. As a human, your instinct would be to explore: perhaps heading to the kitchen first, then scanning the countertops, checking the fridge, or even looking around for a dining area. Your search is driven by a seamless combination of commonsense knowledge, spatial reasoning, and visual grounding [8, 42, 64, 76]. Similarly, an embodied agent navigating a new environment must operate in a continuous closed-loop cycle of exploration, perception, reasoning, and action [31, 64, 73]. A crucial part of this process is understanding both 3D vision and language [1–3, 9, 48, 85, 88] (3D-VL), enabling the agent to think spatially and make informed decisions about where to explore [32, 54, 94].

In recent years, we have seen significant progress in the field of 3D-VL [10, 29, 31, 55, 93, 94]. These models leverage 3D reconstructions to perform visual grounding [9, 26, 69, 87], question answering [3, 48, 92], dense captioning [12], and situated reasoning [31, 48]. Recent approaches, such as 3DVLP [84], PQ3D [94] and LEO [31], aim to handle multiple tasks within a single architecture by pre-training [84, 93] or unified training [5, 29, 31].

However, existing 3D-VL models rely on static 3D representations [9, 63, 67], assuming that a complete reconstruction of the environment is available beforehand [10, 93]. While effective for offline vision language grounding, this assumption is impratical for real-world embodied agents, operating in partially observable and dynamic environments [42, 51, 64]. Moreover, these models typically lack active perception and exploration capabilities [37, 79, 94]. In contrast, reinforcement learning (RL)-based embodied agents can explore environments but often struggle with sample inefficiency [71], poor generalization due to limited training data [20, 57, 62] and the lack of explicit spatial representation. ***Bridging passive 3D-VL grounding and active exploration remains a key challenge in developing intelligent systems capable of efficiently exploring and understanding the 3D world.***

To develop a 3D-VL model with active perception capabilities, three key challenges must be addressed: First, how to effectively learn online representations from raw RGB-D inputs without costly 3D reconstruction, ensuring rich semantics, spatial awareness, and lifelong memory? Second, the joint optimization of object grounding and spatial exploration remains underexplored. Third, training embodied agents requires large-scale trajectory data for robust exploration strategies, yet collecting diverse real-world trajectories presents significant challenges, and methods for effectively leveraging such data remain an open problem.

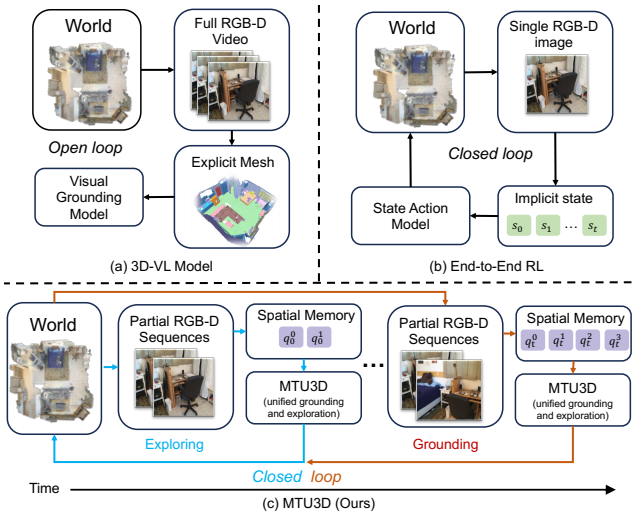


Figure 2. Our approach bridges online exploration with dynamically spatial memory updates for lifelong grounding.

To address these challenges, we propose **M**ove **to** **U**nderstand (**MTU3D**), a unified framework that bridges visual grounding and exploration for versatile embodied navigation as shown in Fig. 2. Our approach introduces three key innovations: 1) *Online Query Representation Learning*. Our model processes raw RGB-D frames as input. It generates single-frame local queries and writes them to a global *spatial memory bank*. By leveraging feature extraction and segment priors from 2D foundation models like DINO [6, 53] and SAM [39], our query representations capture rich semantics and precise 3D spatial information [73]. 2) *Unified Exploration-Grounding Objective*: We introduce a joint optimization framework where unexplored regions are represented as frontier queries. This allows for simultaneous learning of object grounding and exploration. By feeding object queries retrieved from spatial memory bank along with frontier queries [75] detected from the occupancy map [30], we enable a cohesive training process that integrates both tasks. 3) *End-to-End Vision-Language-Exploration (VLE) Pre-training*: We train MTU3D using large-scale trajectory data, combining over a million real-world RGB-D trajectories and simulated data from HM3D in total. To enhance training diversity, we develop an automatic trajectory mixing strategy that blends expert and noisy navigation data [57]. After VLE pre-training, MTU3D can be seamlessly transferred to both simulated environments and real-world scenarios for inference.

Extensive experiments on embodied navigation and question-answering benchmarks demonstrate that MTU3D outperforms existing offline modular-based and RL-based approaches, achieving higher exploration efficiency, better generalization to unseen environments, and improved real-

*Work done as an intern at BIGAI. ✉ Corresponding author. † Department of Computer Science, THUAI, BNRist, Tsinghua University, Beijing 100084, China.

Method	Online	Exploration	Grounding	Lifelong
RL	✓	✓	✗	✗
3D-VL	✗	✗	✓	✓
MTU3D (Ours)	✓	✓	✓	✓

Table 1. Comparison of different paradigms. MTU3D uniquely integrates advantages from both sides, supporting online exploration and lifelong visual grounding.

time decision-making. Specifically, MTU3D improves the state-of-the-art results by 13.7%, 23.0%, and 9.1% in SR, and 2.4%, 13.0%, and 6.3% in SPL on HM3D-OVON [79], GOAT-Bench [37], and SG3D [87], respectively. When combined with a large vision-language model, serving as its trajectory generator, our approach improves the embodied question answering for LM-SR by 2.4% and LLM-SPL by 29.5%. Furthermore, we deploy the model on a real robot, demonstrating its effectiveness in handling realistic 3D environments. These results highlight the significance of bridging visual grounding and exploration as a crucial step toward efficient, versatile, and generalizable embodied intelligence.

Our main contributions can be summarized as follows:

- We present **MTU3D**, *bridging visual grounding and exploration* for efficient and versatile embodied navigation.
- We propose *a unified objective that jointly optimizes grounding and exploration*, leveraging their complementary nature to enhance overall performance.
- We propose a novel *vision-language-exploration training* scheme, leveraging large-scale trajectories from simulation and real-world data.
- Extensive experiments validate the effectiveness of our approach, demonstrating *significant improvements in exploration efficiency and grounding accuracy* across open-vocabulary navigation, multi-modal lifelong navigation, task-oriented sequential navigation, and active embodied question-answering benchmarks.

2. Related work

3D Vision-Language Understanding. In recent years, 3D vision-language (3D-VL) learning [18, 31, 32, 54, 93] has attracted significant attention, with focus on grounding language in 3D scenes by understanding spatial relationships [2, 68], object semantics [9, 24, 26, 67], and scene structures [35, 36, 44, 69]. A wide range of tasks has emerged in this domain, including 3D visual grounding [1, 2, 9, 85], question answering [3, 48, 88], and dense captioning [12]. More recently, the field has expanded to cover intention understanding and task-oriented sequential grounding [87], further pushing the limits of 3D-VL models in complex reasoning [4, 58] and interaction [33, 45, 65]. Existing 3D-VL models can be broadly categorized into task-specific models [10, 26, 27, 34, 81] with specialized architectures for individual tasks, pretrained models [84, 93] that leverage large-scale multi-modal data to improve generalization, and unified models [11, 13, 29, 31, 90, 94] that seek to handle multiple tasks within a single framework. Despite progress, a key limitation of existing 3D-VL models is their reliance on static 3D representations (e.g., precomputed meshes [63, 67] or point clouds [9]), making them unsuitable for real-world embodied AI (EAI), where agents must explore and perceive environments in real-time. EmbodiedSAM [73] partly addresses this issue by taking streaming RGB-D video as input for online 3D instance segmentation [25, 72]; however, it lacks active exploration and high-level reasoning capability. In contrast, our MTU3D framework is proposed as a unified model aiming to simultaneously learn scene representations, exploration strategies, and grounding directly from *dynamic spatial memory bank* during online RGB-D exploration.

Embodied Navigation and Reasoning. The recent advances in embodied AI, particularly embodied navigation [28, 38, 61, 66] and reasoning [17, 42, 50, 59, 64, 64], primarily rely on three crucial capabilities: perception, reasoning, and exploration. Several benchmarks are proposed to assess navigation capabilities [37, 79] across different goal specifications (images, objects, or language instructions), examine the sequential awareness [87], or question-answering in an embodied setting [51]. Efforts to tackle these challenges generally follow two main approaches [37, 46, 83, 95]: end-to-end reinforcement learning and modular architectures. End-to-end approaches, like PIRLNav [57] and VER [71], utilize RNNs or transformers [19, 82] trained directly for navigation tasks, integrating perception and reasoning to take actions. However, its end-to-end nature without explicit 3D representations limits its performance under complex instructions in intricate environments [28, 80]. In contrast, modular approaches [7, 43, 59, 77, 78, 80] decompose navigation into specialized components, maintaining distinct models for mapping and navigation policies. More recently, CLIP on Wheels [23, 40, 56] leverages pre-trained vision-language models to interpret navigation goals without fine-tuning [49]. Unlike these methods as in Tab. 1, our model employs a vision-language-exploration training paradigm, actively exploring to construct a scene representation and reasoning within an end-to-end system.

3. MTU3D

In this section, we present the architecture of our model in Fig. 3 and the training pipeline. We begin by detailing our approach to query representation learning, which extracts object and frontier queries from partial RGB-D sequences and dynamically stores them in a spatial memory bank. Next, we introduce our unified grounding and exploration objective, where queries are processed through a spatial reasoning layer for selection. Finally, we describe our trajectory collection strategy and training procedure.

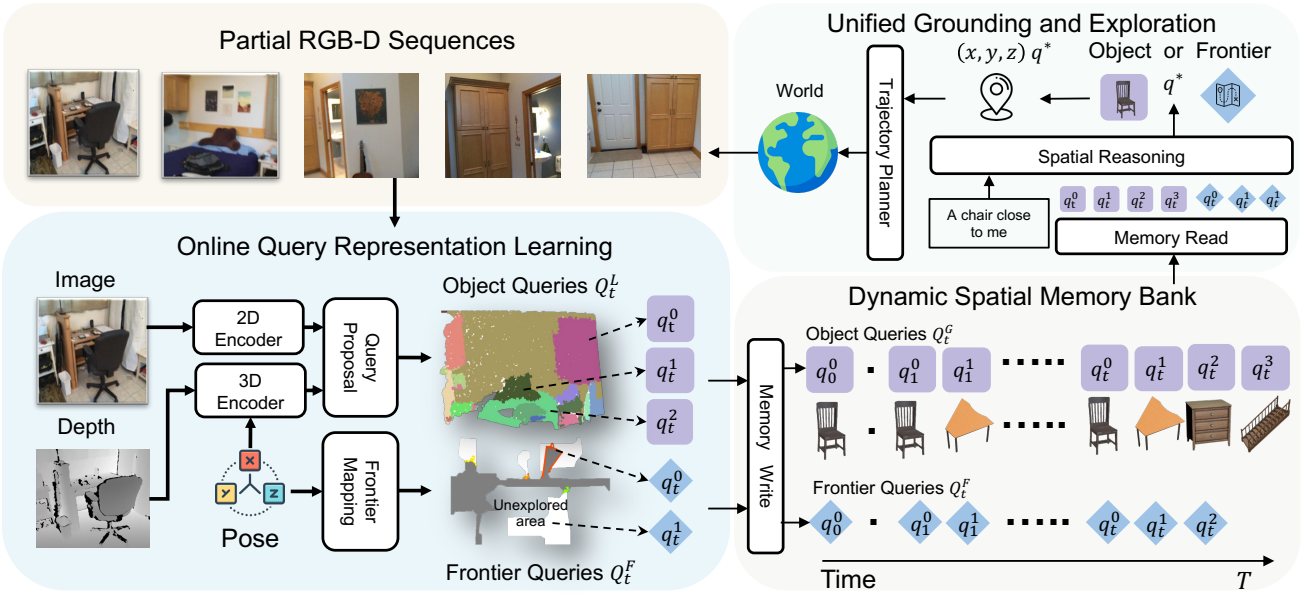


Figure 3. Our proposed model processes RGB-D sequences to generate object queries, which are stored in a memory bank. The spatial reasoning layer then selects either object queries or frontier points for exploration. These selected query locations are passed to a trajectory planner that guides navigation and facilitates the acquisition of new RGB-D sequences, creating a continuous perception-action loop.

3.1. Online Query Representation Learning

Our model takes as input a partial RGBD sequence spanning an arbitrary time range, denoted as $O = [o_{t_1}, o_{t_2}, \ldots, o_t]$. At each timestep t, we extract local queries Q_t^L, which are subsequently aggregated across the interval t_1 to t_2 to generate global queries $Q_{t_2}^G$ that are stored in a memory bank. This process operates in an online manner, enabling flexible processing at any timestep. Our online representation learning framework comprises three essential components, as illustrated below.

2D and 3D Encoding. For each input observation $o_t = [I_t, D_t, P_t]$, we process three components: RGB image $I_t \in \mathbb{R}^{H \times W \times 3}$, depth image $D_t \in \mathbb{R}^{H \times W}$, and camera pose $P_t \in SE(3)$. Following ESAM [73], we apply FastSAM [39, 89] to segment the image into distinct regions, generating a segment index map $S_t \in \mathbb{N}^{H \times W}$ where each value represents a region ID. For 2D feature extraction, we process I_t through a DINO backbone [6, 53] to obtain pixel-level features $F_t^{2D} \in \mathbb{R}^{H \times W \times C}$. These features are pooled according to S_t, producing segment-level representations [94] $\hat{F}_t^{2D} \in \mathbb{R}^{M \times C}$, with M denoting the number of segments identified by FastSAM. Each segment corresponds to a coarse image region associated with a group of pixels and their respective 3D points.

For 3D feature extraction, we project the depth image D_t into a point cloud and uniformly downsample it to N points $\mathcal{P}_t \in \mathbb{R}^{N \times 3}$, which is processed through a sparse convolutional U-Net [14, 15, 63] to generate 3D features $F_t^{3D} \in \mathbb{R}^{N \times C}$, with C being the feature dimension. We then perform segment-wise pooling [22] using S_t, resulting in segment-level 3D representations $\hat{F}_t^{3D} \in \mathbb{R}^{M \times C}$.

Local Query Proposal. After obtaining 2D and 3D features, we generate object queries through a multi-layer perceptron: $Q_t = \text{MLP}(\hat{F}_t^{2D}, \hat{F}_t^{3D}) \in \mathbb{R}^{M \times C}$. These initial queries are refined via PQ3D-inspired [94, 96] decoder layers, producing local output queries Q_t^L. Each refined local query $q_t^L \in Q_t^L$ comprises multiple components: $q_t^L = [b_t, m_t, f_t, v_t, s_t]$, where $b_t \in \mathbb{R}^6$ encodes a 3D bounding box in global coordinates, $m_t \in \mathbb{R}^M$ represents segment-level mask predictions for instance segmentation, $v_t \in \mathbb{R}^C$ captures open-vocabulary feature embeddings for semantic alignment, $f_t \in \mathbb{R}^C$ contains output features from the query decoder, and $s_t \in \mathbb{R}$ denotes a confidence score indicating detection reliability. Each mask in shape M can be converted to a 2D mask ($H \times W$) or a single frame 3D point cloud mask (N points). As the input point cloud is transformed into world coordinates using pose information, the predicted b_t exists in a shared coordinate system, enabling subsequent query merging operations.

Dynamic Spatial Memory Bank. We merge local queries with historical queries from our spatial memory bank [52, 73, 91] by calculating bounding box IoU between current local queries Q_t^L and previous global queries Q_{t-1}^G, generating updated global queries Q_t^G, which contains f_t and v_t for further grounding and exploration. For matched query pairs, we employ a fusion strategy where bounding box parameters b_t, feature embeddings f_t, semantic vectors v_t, and confidence scores s_t are combined using exponential mov-

ing averages, while instance segmentation masks m_t are fused through a union operation following [73]. To identify unexplored regions, we maintain an occupancy map $\mathcal{M} \in \mathbb{R}^{X \times Y \times 3}$ that categorizes spatial segments as occupied, unoccupied, or unknown. Central to our exploration approach is the concept of frontiers, denoted as Q_t^F, which represent boundaries between explored (known) and unexplored (unknown) areas. These frontiers are identified by traversing the boundaries of the explored space and detecting adjacent unknown areas. Each element of Q_t^F corresponds to a coordinate in 3D space $\mathbb{R}^3$, serving as a potential exploration target [75]. These frontier waypoints are periodically recalculated upon reaching each target position, directing autonomous agents to explore unmapped regions.

3.2. Unified Grounding and Exploration

Our unified approach combines object grounding and exploration in a single decision framework. Given a natural language goal L (e.g., "find a red chair"), our model decides between grounding an object from current global queries Q_t^G or selecting a frontier from Q_t^F for exploration.A spatial reasoning transformer, described in detail in the appendix, integrates language instructions, object representations, and frontier information to produce a unified score $S_t^U = f(Q_t^G, Q_t^F, L)$ for each candidate decision. Frontier points are encoded through a simple two-layer MLP before being fed into the transformer. The input for Q_t^G consists of f_t and v_t. We incorporate type embeddings to distinguish between object and frontier queries. For language goals, we use a CLIP text encoder [56], while image-based goals employ the CLIP image encoder [41, 56]. These embeddings are projected into the transformer's feature space for cross-attention with query representations. The final decision selects the query with the highest score: $q^* = \arg\max_{q_i \in Q_t^G \cup Q_t^F} S_t^U(q_i)$. If $q^* \in Q_t^G$, the system grounds the corresponding object; if $q^* \in Q_t^F$, it navigates to that frontier location. We employ the Habitat-Sim shortest path planner to generate local trajectories. This mechanism dynamically balances exploration and grounding based on current scene knowledge and goal [8].

3.3. Trajectory Data Collection

Our unified model training requires diverse trajectory data, which is difficult to collect manually [57, 70]. We implement a systematic collection process spanning simulated and real environments, combining *visual grounding data* and *exploration data* as shown in Tab. 2.

Visual Grounding Trajectory. Our visual grounding data follows the structure (Object Q_t^G, Language L) $\longrightarrow$ (Decision S_t^U). Leveraging offline RGB-D videos in ScanNet [16, 60] as trajectories, we can directly use these paired examples for pre-training. Each sample associates object queries with language descriptions to generate decisions.

Source	Scan	Sim	Real	VG	Exp	Traj	Dec	Goal
ScanRefer [9]	ScanNet	✗	✓	✓	✗	1202	37k	37k
ScanQA [3]	ScanNet	✗	✓	✓	✗	1202	26k	30k
Multi3DRefer [85]	ScanNet	✗	✓	✓	✗	1202	44k	44k
SG3D-VG-HM3D [87]	HM3D	✓	✗	✓	✗	145	30k	30k
SG3D-VG-ScanNet [87]	ScanNet	✗	✓	✓	✗	1202	13k	13k
Nr3D [2]	ScanNet	✗	✓	✓	✗	1202	30k	30k
SceneVerse-HM3D [36]	HM3D	✓	✗	✓	✗	145	48k	48k
HM3D-OVON [79]	HM3D	✓	✗	✗	✓	290k	1940k	290k
GOAT-Bench [37]	HM3D	✓	✗	✗	✓	680k	14119k	5098k
SG3D-Nav [87]	HM3D	✓	✗	✗	✓	2k	34k	11k

Table 2. Data source statistic for Vision-Language-Exploration Pre-training. **Sim** denotes simulation, **VG** denotes visual grounding, **Exp** denotes exploration, **Traj** denotes trajectory, **Dec** denotes decision and **Goal** denotes number of goals.

Exploration Trajectory. Exploration data follows the format (Object Q_t^G, Frontier Q_t^F, Goal G) $\longrightarrow$ (Decision S_t^U), with frontiers changing during exploration [75]. Training with only optimal frontiers (closest to target) leads to overfitting. We address this by implementing random frontier selection and a hybrid strategy combining random and optimal approaches [57]. We collect trajectories from simulation scans (HM3D [74] via Habitat-Sim [61]) to apply these varied exploration strategies. Exploration succeeds when the target becomes visible and reachable. To prevent unnecessary exploration, we maintain a visited frontier list and only explore when better potential frontiers exist (much closer to target); otherwise, we raise an exception and terminate the current collection. The complete collection process appears in the appendix.

3.4. Vision-Language-Exploration Training

Stage 1: Low-level Perception Training. We utilize RGB-D trajectories from ScanNet and HM3D to train query representation with instance segmentation loss. Our loss function combines multiple components: $\mathcal{L} = \lambda_b \mathcal{L}\text{box} + \lambda_m \mathcal{L}\text{mask} + \lambda_v \mathcal{L}\text{vocab} + \lambda_s \mathcal{L}\text{score}$. $\mathcal{L}$box is the 3D box IoU loss; $\mathcal{L}$mask and $\mathcal{L}$score are binary cross-entropy losses; $\mathcal{L}$vocab is cosine similarity loss. This approach ensures that output local queries Q_t^L effectively capture spatial, semantic, and confidence information.

Stage 2: Vision-Language-Exploration Pre-training. In VLE pre-training, we utilize stage 1 output queries to jointly train exploration and grounding. Using the dataset in Tab. 2, we train our decision model on over one million trajectories. The unified decision scores S_t^U are optimized with binary cross-entropy loss, teaching the model to assign higher scores to appropriate query locations based on the current state and goal.

Stage 3: Task-Specific Navigation Fine-tuning. During this stage, we employ the same objective as stage 2 and fine-tune our MTU3D to specific navigation trajectories, optimizing performance for targeted deployment scenarios.

4. Experiment

4.1. Experimental setting

Dataset and Benchmarks. We evaluate on diverse benchmarks: GOAT-Bench [37] (multi-modal lifelong navigation), HM3D-OVON [79] (open-vocabulary navigation), SG3D [87] (sequential task navigation), and A-EQA [51] (embodied question answering). Common metrics include Success Rate ($SR = \frac{N_{\text{success}}}{N_{\text{total}}}$) and Success weighted by Path Length ($SPL = \frac{1}{N_{\text{total}}} \sum_{i=1}^{N_{\text{total}}} S_i \cdot \frac{l_i}{\max(p_i, l_i)}$), where S_i indicates success, l_i is shortest path length, and p_i is agent path length. SG3D uses task SR (*t-SR*) to measure step coherence. A-EQA employs LLM match score (*LLM-SR*) and score averaged by exploration length (*LLM-SPL*). We compare against a diverse range of baseline methods, including modular approaches such as GOAT [37] and VLFM [78], end-to-end reinforcement learning approaches [57, 79], and video-based approaches [83].

Implementation Details. In Stage 1, we train for 50 epochs using AdamW (learning rate 1e-4, β_1 = 0.9, β_2 = 0.98) with loss weights $\lambda_b = 1.0, \lambda_m = 1.0, \lambda_v = 1.0, \lambda_s = 0.5$. Stages 2 and 3 use identical optimizer settings for 10 epochs each. Both Stages 1 and 2 use 4 transformer layers. Query proposal is trained in stage 1 then frozen, and spatial reasoning is trained in later stages. All training runs on four NVIDIA A100 GPUs around 164 GPU hours. For simulation evaluation, we follow [37, 79, 87] using Stretch embodiment (1.41m tall, 17cm base radius), processing 360×640 RGB images I_t, depth maps D_t, and pose P_t with actions: `MOVE_FORWARD`(0.25m), `TURN_LEFT`, `TURN_RIGHT`, `LOOK_UP`, and `LOOK_DOWN`. Spatial reasoning is activated upon arrival at each target position. We subsample 18 frames along the trajectory between consecutive target positions.For A-EQA [51], our model is used solely to generate the exploration trajectory and collect the corresponding video for each question.

4.2. Quantitative Results

Open Vocabulary Navigation. The results in Tab. 3 demonstrate that our proposed MTU3D significantly outperforms all baselines in terms of SR across both Val Seen and Val Unseen settings. Notably, MTU3D achieves the highest SR in Val Unseen (40.8%), showcasing its strong generalization ability to unseen episodes. This highlights the effectiveness of our approach in handling unseen scenarios compared to prior methods such as RL and behavior cloning (BC), which are not pre-trained on large-scale language data. However, we observe that MTU3D's SPL is lower than Uni-Navid, especially in the Val Synonyms and Val Unseen settings. Given that the HM3D-OVON dataset consists of relatively short trajectories, video-based models inherently gain an advantage because they can directly navigate to the goal upon recognizing a target.

	Val Seen		Val Seen Synonyms		Val Unseen	
Method	SR (↑)	SPL (↑)	SR (↑)	SPL (↑)	SR (↑)	SPL (↑)
BC	11.1	4.5	9.9	3.8	5.4	1.9
DAgger	18.1	9.4	15.0	7.4	10.2	4.7
RL	39.2	18.7	27.8	11.7	18.6	7.5
BCRL	20.2	8.2	15.2	5.3	8.0	2.8
DAgRL	41.3	21.2	29.4	14.4	18.3	7.9
VLFM	35.2	18.6	32.4	17.3	35.2	19.6
Uni-NaVid	41.3	21.1	43.9	**21.8**	39.5	19.8
TANGO	–	–	–	–	35.5	19.5
DAgRL+OD	38.5	21.1	39.0	21.4	37.1	**19.9**
MTU3D (Ours)	**55.0**	**23.6**	**45.0**	14.7	**40.8**	12.1

Table 3. Open-vocab navigation results on HM3D-OVON [79].

Task-oriented Sequential Navigation. Tab. 4 presents results on task-oriented sequential navigation, a challenging task requiring an embodied agent to understand relationships between task steps. The results show that MTU3D achieves the highest s-SR (23.8%), t-SR (8.0%), and SPL (16.5%) on the SG3D benchmark, demonstrating its effectiveness in sequential task execution and task understanding. Unlike other benchmarks, SG3D emphasizes task consistency across multiple steps, making it more complex. While MTU3D significantly outperforms Embodied Video Agent [21] and SenseAct-NN Monolithic [37, 87], overall success rates remain lower than in GOAT-Bench and HM3D-OVON, highlighting SG3D's inherent difficulty in requiring both navigation and sustained task accuracy.

	s-SR(↑)	t-SR(↑)	SPL(↑)
Embodied Video Agent	14.7	3.8	10.2
SenseAct-NN Monolithic	12.1	7.7	10.1
MTU3D (Ours)	**23.8**	**8.0**	**16.5**

Table 4. Sequential task navigation results on SG3D-Nav [87].

Multi-modal Lifelong Navigation. The results in Tab. 5 highlight the significant performance improvement of our MTU3D over baseline methods in lifelong setting. Notably, MTU3D achieves the highest SR across all settings, with 52.2% in Val Seen, 48.4% in Val Seen Synonyms, and 47.2% in Val Unseen, demonstrating its superior navigation capability. Compared to open-vocabulary navigation, multi-modal lifelong navigation is a more challenging task due to the need for continuous spatial memory and long-term reasoning. Our model's lifelong spatial memory allows it to retain and utilize past experiences more effectively, leading to a much larger performance gain over baselines. Additionally, MTU3D achieves the highest SPL across all settings, with 30.5% in Val Seen and 27.7% in Val Unseen, indicating that it not only reaches the goal more accurately but also follows more efficient trajectories. This suggests that our approach effectively balances success rate and efficiency,

Method	Val Seen		Val Seen Synonyms		Val Unseen	
	SR (↑)	SPL (↑)	SR (↑)	SPL (↑)	SR (↑)	SPL (↑)
Modular GOAT	26.3	17.5	33.8	24.4	24.9	17.2
Modular CLIP on Wheels	14.8	8.7	18.5	11.5	16.1	10.4
SenseAct-NN Skill Chain	29.2	12.8	38.2	15.2	29.5	11.3
SenseAct-NN Monolithic	16.8	9.4	18.5	10.1	12.3	6.8
TANGO	-	-	-	-	32.1	16.5
MTU3D (ours)	**52.2**	**30.5**	**48.4**	**30.3**	**47.2**	**27.7**

Table 5. Multi-modal lifelong navigation results on GOAT-Bench [37].

a crucial factor in lifelong navigation. Overall, these results emphasize that lifelong spatial memory is key to multi-modal navigation, and MTU3D's substantial improvements validate its effectiveness in handling this complex task.

Active Embodied Question Answering. Tab. 6 demonstrate that our MTU3D-enhanced GPT-4V significantly outperforms the baseline GPT-4V model, achieving 44.2% LLM-SR vs. 41.8% and a much higher LLM-SPL of 37.0% vs. 7.5%. This indicates that our approach enables a more efficient trajectory, avoiding the exhaustive search across all locations that baseline models rely on. Furthermore, GPT-4o with MTU3D achieves even better performance, reaching 51.1% LLM-SR and 42.6% LLM-SPL.

Method	LLM-SR(↑)	LLM-SPL(↑)
Blind LLMs		
GPT-4	35.5	N/A
LLaMA-2	29.0	N/A
LLM with captions		
GPT-4 w/ LLaVA-1.5	38.1	7.0
LLaMA-2 w/ LLaVA-1.5	30.9	5.9
VLMs		
GPT-4V	41.8	7.5
VLMs with MTU3D Trajectory		
GPT-4V (Ours)	44.2	37.0
GPT-4o (Ours)	**51.1**	**42.6**

Table 6. Embodied question answering results on A-EQA [51].

4.3. Discussions

Does Vision-Langauge-Exploration Pe-training benefit navigation? The results in the Fig. 4a show that Vision-Language Exploration (VLE) Pre-training significantly improves navigation performance, as indicated by the SR across all datasets. Specifically, SR increases from 27.8% to 33.3% in OVON, 22.2% to 36.1% in GOAT, and 22.9% to 27.9% in SG3D, demonstrating a consistent benefit of VLE across different task settings and distribution.

Does grounded training lead to efficient exploration? The results in Fig. 4c demonstrate that MTU surpasses frontier exploration in both SR and SPL as exploration steps increase. Unlike frontier exploration, which blindly selects the nearest frontier, MTU utilizes semantic guidance for more efficient, goal-directed exploration. Notably, at step 6, MTU achieves a higher SR (50.0% vs. 33.3%) and improved SPL (35.3% vs. 30.3%).

Can spatial memory bank enhance lifelong navigation? We reset spatial memory in w/o mem for each sub-episode in GOAT-Bench, and the experimental results in Fig. 4b show that memory significantly improves SR across all goal types. Notably, SR increases from 10.5% to 52.6% for Object goals, 28.6% to 71.4% for Description goals, and 26.7% to 60.0% for Image goals, demonstrating that memory helps retain useful spatial information.

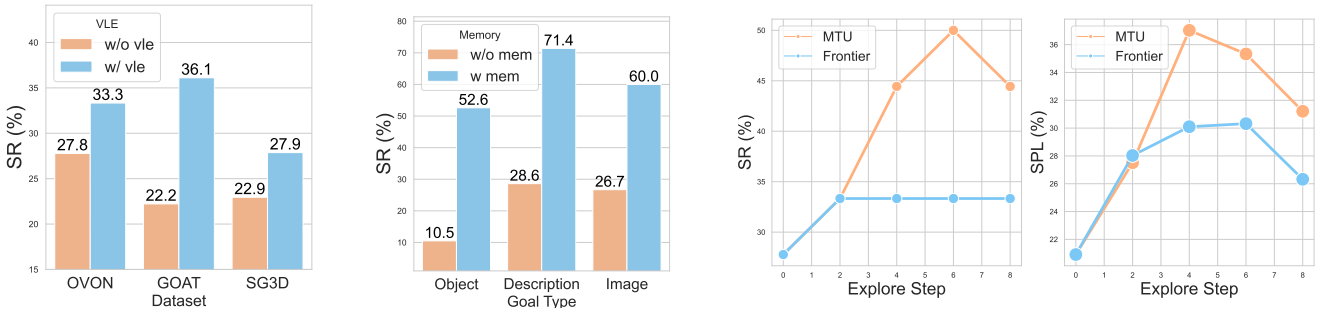
Can MTU3D run in real time? Tab. 7 shows that our model achieves efficient query proposal (192 ms) and fast reasoning (31 ms) while maintaining a competitive FPS (3.4) with 266M parameters. These results highlight the model's optimal balance between speed and performance, making it well-suited for real-time applications.

Query Proposal	Spatial Reasoning	FPS	Params
192 ms	31 ms	3.4	266M

Table 7. Model speed and parameter metrics, results are average from 5 runs across multiple frames and episodes on 3090 Ti.

4.4. Qualitative results

These qualitative results in Fig. 5 showcase the agent's ability to navigate and complete diverse goal types, including language, image, description, and task planning goals. The trajectories illustrate how the agent efficiently locates objects based on visual and semantic cues, demonstrating its capability to understand both image-based and textual instructions. Notably, for task planning goals, the agent follows a structured sequence of actions.



(a) Effect of VLE. (b) Usefulness of spatial memory. (c) Exploration efficiency comparison.

Figure 4. Ablation studies showing (a) the impact of vision-language-exploration pretraining, (b) exploration efficiency on seen environments, and (c) the contribution of spatial memory to navigation performance.

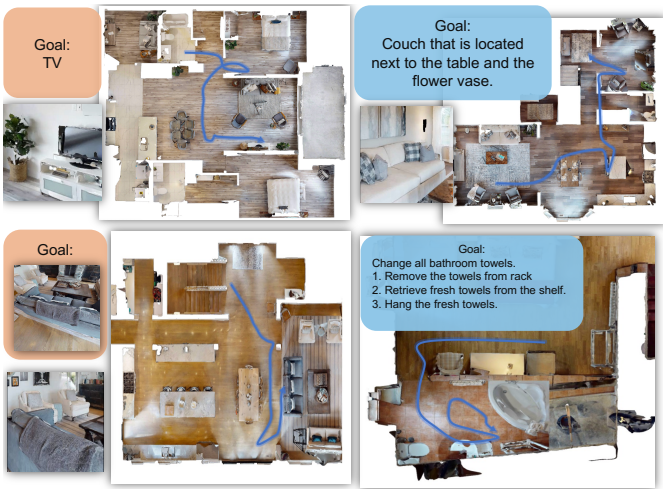


Figure 5. Visualization of results in Habitat-Sim.

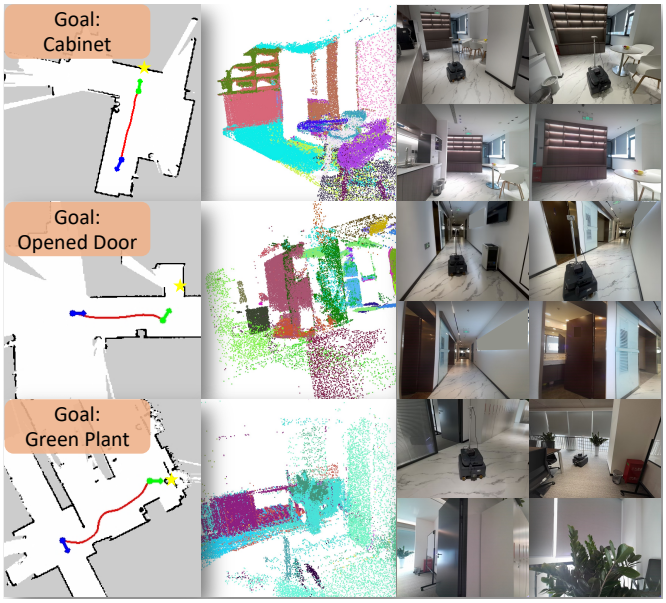


Figure 6. Real World Testing.

4.5. Real World Testing

We evaluate MTU3D in realistic 3D environments by deploying it on an NVIDIA Jetson Orin with a Kinect [86] for real-time RGB-D data and a mobile robot equipped with Lidar for exploration. Without any real-world fine-tuning, we test the model in three diverse scenes: home, corridor, and meeting room. As shown in Fig. 6, MTU3D effectively navigates to the target position. Since MTU3D is trained on both simulated and real data, it overcomes the Sim-to-Real transfer challenges commonly faced in RL-based methods. This ability not only enhances its real-world applicability but also makes it highly scalable and impactful for advancing embodied intelligence in the future.

5. Conclusions

In this paper, we introduce Move to Understand (**MTU3D**), a unified framework that bridges visual grounding and exploration to advance embodied scene understanding. By jointly optimizing grounding and exploration, MTU3D enables efficient navigation across diverse input modalities, fostering a deeper understanding of spatial environments. Our **V**ision-**L**anguage-**E**xploration (VLE) training leverages large-scale trajectories, achieving state-of-the-art performance on multiple Embodied AI benchmarks. Experimental results highlight the crucial role of *spatial memory* in enabling lifelong multi-modal navigation and spatial intelligence, allowing agents to reason about and adapt to complex environments. Furthermore, real-world deployment demonstrates MTU3D 's generalization ability, validating the effectiveness of our mixed training with both simulation and real-world data. By bridging visual grounding and exploration, MTU3D paves the way for more capable, scalable, and generalizable embodied agents, bringing us closer to the goal of truly intelligent embodied AI.

Acknowledgements. This work was supported in part by the National Science Foundation of China (NSFC) under Grant No. 62176134 and by a grant from the Assisted Medical Consultation Project Based on DeepSeek.

References

[1] Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, and Panos Achlioptas. Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, pages 3524–3534, 2024. 2, 3

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 5

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5

[4] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022. 3

[5] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4

[7] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[8] 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 (NeurIPS)*, 33: 4247–4258, 2020. 2, 5

[9] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5

[10] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[11] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[12] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[13] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *International Conference on Computer Vision (ICCV)*, pages 18109–18119, 2023. 3

[14] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[15] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[16] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 5

[17] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[18] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. 2, 3

[19] Ainaz Eftekhar, Luca Weihs, Rose Hendrix, Ege Caglar, Jordi Salvador, Alvaro Herrasti, Winson Han, Eli VanderBil, Aniruddha Kembhavi, Ali Farhadi, et al. The one ring: a robotic indoor navigation generalist. *arXiv preprint arXiv:2412.14401*, 2024. 3

[20] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16238–16250, 2024. 2

[21] Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. *arXiv preprint arXiv:2501.00358*, 2024. 6

[22] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59:167–181, 2004. 4

[23] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[24] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision (ECCV)*, pages 540–557. Springer, 2022. 3

[25] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 3

[26] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *International Conference on Computer Vision (ICCV)*, pages 15372–15383, 2023. 2, 3

[27] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3

[28] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[29] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 20482–20494, 2023. 2, 3

[30] Adam Hoover and Bent David Olsen. A real-time occupancy map from multiple video streams. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, pages 2261–2266. IEEE, 1999. 2

[31] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *International Conference on Machine Learning (ICML)*, 2024. 2, 3

[32] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*, pages 9118–9147. PMLR, 2022. 2, 3

[33] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*, 2023. 3

[34] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[35] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Subramanian Iyer, Soroush Saryazdi, Nikhil Varma Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. In *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023. 3

[36] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *European Conference on Computer Vision (ECCV)*, 2024. 3, 5

[37] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024. 2, 3, 5, 6, 7

[38] Taewoong Kim, Cheolhong Min, Byeonghwi Kim, Jinyeon Kim, Wonje Jeung, and Jonghyun Choi. Realfred: An embodied instruction following benchmark in photo-realistic environments. In *European Conference on Computer Vision*, pages 346–364. Springer, 2024. 3

[39] 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 *International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 4

[40] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *arXiv preprint arXiv:2209.13430*, 2022. 3

[41] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[42] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 2, 3

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[44] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 3

[45] 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. In *International Conference on Robotics and Automation (ICRA)*, 2023. 3

[46] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and

Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *arXiv preprint arXiv:2403.07376*, 2024. 3

[47] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024. 2

[48] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[49] 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. 3

[50] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[51] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498, 2024. 2, 3, 6, 7

[52] David S Olton. Spatial memory. *Scientific American*, 236 (6):82–99, 1977. 4

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4

[54] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2, 3

[55] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–824, 2023. 2

[56] 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 (ICML)*, pages 8748–8763. PMLR, 2021. 3, 5

[57] Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17896–17906, 2023. 2, 3, 5, 6

[58] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *7th Annual Conference on Robot Learning*, 2023. 3

[59] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024. 3

[60] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision (ECCV)*, pages 125–141. Springer, 2022. 5

[61] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9339–9347, 2019. 3, 5

[62] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[63] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 2, 3, 4

[64] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 2, 3

[65] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[66] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[67] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[68] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[69] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai

Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 2, 3

[70] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 5

[71] Erik Wijmans, Irfan Essa, and Dhruv Batra. Ver: Scaling on-policy rl leads to the emergence of navigation in embodied rearrangement. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:7727–7740, 2022. 2, 3

[72] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[73] Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam: Online segment any 3d thing in real time. *arXiv preprint arXiv:2408.11811*, 2024. 2, 3, 4, 5

[74] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023. 5

[75] Brian Yamauchi. Frontier-based exploration using multiple robots. In *Proceedings of the second international conference on Autonomous agents*, pages 47–53, 1998. 2, 5

[76] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. 2

[77] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17294–17303, 2025. 3

[78] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *International Conference on Robotics and Automation (ICRA)*, 2024. 3, 6

[79] Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5543–5550. IEEE, 2024. 2, 3, 5, 6

[80] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. Frontier semantic exploration for visual target navigation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 3

[81] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[82] Kuo-Hao Zeng, Zichen Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, and Luca Weihs. Poliformer: Scaling on-policy rl with transformers results in masterful navigators. *arXiv preprint arXiv:2406.20083*, 2024. 3

[83] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024. 3, 6

[84] Taolin Zhang, Sunan He, Dai Tao, Bin Chen, Zhi Wang, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding. *arXiv preprint arXiv:2305.10714*, 2023. 2, 3

[85] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *International Conference on Computer Vision (ICCV)*, pages 15225–15236, 2023. 2, 3, 5

[86] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 8

[87] Zhuofan Zhang, Ziyu Zhu, Junhao Li, Pengxiang Li, Tianxu Wang, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding and navigation in 3d scenes. *arXiv preprint arXiv:2408.04034*, 2024. 2, 3, 5, 6

[88] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Yinjie Zhao, Lipeng Wang, and Xibo Fan. Towards explainable 3d grounded visual question answering: A new benchmark and strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2, 3

[89] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 4

[90] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 3

[91] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3769–3777, 2023. 4

[92] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision (ECCV)*, pages 151–168. Springer, 2024. 2

[93] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *International Conference on Computer Vision (ICCV)*, pages 2911–2921, 2023. 2, 3

[94] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024. 2, 3, 4

[95] Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. Tango: Training-free embodied ai agents for open-world tasks. *arXiv preprint arXiv:2412.10402*, 2024. 3

[96] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15116–15127, 2023. 4