Towards Learning a Generalist Model for Embodied Navigation

Duo Zheng[1,2*] Shijia Huang[1*] Lin Zhao[3†] Yiwu Zhong[1] Liwei Wang[1‡]
[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory
[3]Centre for Perceptual and Interactive Intelligence
{dzheng23, sjhuang, lwwang}@cse.cuhk.edu.hk

Abstract

Building a generalist agent that can interact with the world is the intriguing target of AI systems, thus spurring the research for embodied navigation, where an agent is required to navigate according to instructions or respond to queries. Despite the major progress attained, previous works primarily focus on task-specific agents and lack generalizability to unseen scenarios. Recently, LLMs have presented remarkable capabilities across various fields, and provided a promising opportunity for embodied navigation. Drawing on this, we propose the first generalist model for embodied navigation, NaviLLM. It adapts LLMs to embodied navigation by introducing schema-based instruction. The schema-based instruction flexibly casts various tasks into generation problems, thereby unifying a wide range of tasks. This approach allows us to integrate diverse data sources from various datasets into the training, equipping NaviLLM with a wide range of capabilities required by embodied navigation. We conduct extensive experiments to evaluate the performance and generalizability of our model. The experimental results demonstrate that our unified model achieves state-of-the-art performance on CVDN, SOON, and ScanQA. Specifically, it surpasses the previous stats-of-the-art method by a significant margin of 29% in goal progress on CVDN. Moreover, our model also demonstrates strong generalizability and presents impressive results on unseen tasks, e.g. *embodied question answering and 3D captioning. Our code is available at* https://github.com/LaVi-Lab/NaviLLM.

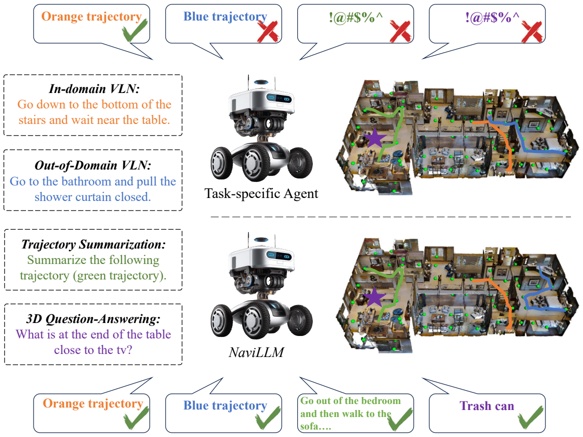

Orange trajectory ✓
Blue trajectory ✗
!@#$%^ ✗
!@#$%^ ✗
In-domain VLN:
Go down to the bottom of the stairs and wait near the table.
Out-of-Domain VLN:
Go to the bathroom and pull the shower curtain closed.
Task-specific Agent
Trajectory Summarization:
Summarize the following trajectory (green trajectory).
3D Question-Answering:
What is at the end of the table close to the tv?
NaviLLM
Orange trajectory ✓
Blue trajectory ✓
Go out of the bedroom and then walk to the sofa.... ✓
Trash can ✓

Figure 1. Comparison between previous methods and ours. Previous methods learn task-specific navigation agents, suffer from a low success rate for out-of-domain VLN, and fall short when facing unseen tasks (*e.g*., QA and summarization). The different colors are used to represent different examples. For instance, orange represents an example from In-domain VLN. Our *NaviLLM* not only excels in diverse tasks required by embodied navigation, but also demonstrates promising generalizability even on unseen tasks.

1. Introduction

The pursuit of artificial intelligence has long been driven by the desire to construct agents that are capable of acquiring knowledge through interacting with the physical world, akin to the complex interaction processes exhibited by humans. This has led to the emergence of embodied navigation [2, 33, 47], where an agent located in 3D environment is required to navigate according to various forms of instructions and provide textual responses based on user queries.

A wide spectrum of tasks have been introduced for embodied navigation, ranging from vision-language navigation that follows step-by-step instructions [3, 34] or coarse-grained directives [17, 43, 61], to the tasks guided by interactions between humans and agents [19, 52], and even to embodied question answering through proactive exploration [18, 55]. To tackle these tasks, a myriad of methodologies have been explored in the past, with some notable approaches leveraging pre-training techniques [24, 25, 32, 40], data augmentation [14, 22, 29], and memory structures [11, 27], etc. These models, while demonstrating considerable proficiency in their specific tasks, unfortunately lack generaliza-

[*] Equal contribution.

[†] Lin Zhao was a research assistant at the Centre for Perceptual and Interactive Intelligence (CPII) under the InnoHK.

[‡] Corresponding author.

tion across diverse scenarios. This naturally raises a question: *Can we train a generalist model that is generalizable to many embodied navigation tasks?*

The advancement of Large Language Models (LLMs) has provided a promising opportunity to construct a generalist model for embodied navigation. In recent years, LLMs [9, 15, 53] have demonstrated human-like capability for text understanding and text generation. Given such impressive performance, numerous works [5, 16, 39, 59] have pioneered adapting LLMs for vision-language tasks through fine-tuning on a variety of data sources. Beyond the image domain, LLMs have exhibited remarkable generalizability in other domains, such as video understanding [36], 3D understanding [28], and robotic manipulation [6, 7, 49]. However, the potential of adapting LLMs to embodied navigation tasks remains largely unexplored.

In this work, we aim to learn a generalist model for embodied navigation by adapting LLMs. The main challenge lies in how to unify a wide range of tasks in a single model. To address the challenge, our key idea is to cast all task learning into generative modeling, with the help of pretrained LLMs. Specifically, we propose schema-based instruction and design a series of schemas (*e.g*., descriptions of tasks, visual observation, and navigation history), based on the characteristics of embodied tasks. These schemas are flexible to cast various vision-centric tasks into generation problems. For example, we can effortlessly convert vision-language navigation into the generation of movement direction, and convert object localization into the generation of object IDs. Benefitting from this design, we are able to train a unified model on the data collected for diverse tasks, thereby enabling our model to address a wide spectrum of tasks, ranging from vision-language navigation and object localization, to 3D question answering, trajectory summarization, embodied question answering. Therefore, our approach significantly mitigates the problem of data scarcity and empowers the model to understand instructions of varying formats and granularities, thereby enabling a suite of capabilities to interact with the 3D environment.

We train *NaviLLM* on a combined dataset covering diverse embodied tasks (CVDN, SOON, R2R, REVERIE, ScanQA, LLaVA-23k, and augmented data for R2R and REVERIE), and conduct extensive experiments to evaluate the competencies and generalizability of *NaviLLM*. With only a single model, *NaviLLM* has achieved **new state-of-the-art** results simultaneously on multiple benchmarks, *i.e*. CVDN [52], SOON [61], and ScanQA [4], and demonstrated comparable performance to latest models on R2R [3] and REVERIE [43]. Notably, our model achieves a relative improvement of **29%** over the previous state-of-the-art on the CVDN benchmark. Further, we evaluate the generalizability of our method by excluding CVDN, SOON, and REVERIE from the training data, respectively. Our method outperforms baseline methods on all tasks, significantly improving the Success Rate by **136%** on SOON. Moreover, we observe that our model presents an impressive capability for unseen tasks, *e.g*. embodied question answering and 3D captioning. Collectively, the results from these experiments not only attest to the generalization capability of the model but also highlight the significant potential of our approach in learning a generalist model for embodied navigation.

Our contribution could be summarized as follows:

- We propose the first generalist model for embodied navigation, namely *NaviLLM*, enabling a wide spectrum of capabilities required for embodied navigation.
- We unify various tasks in a single model by adapting LLM and introducing schema-based instruction. By doing this, our model can harness data sources from diverse datasets.
- Our single model achieves SoTA results on CVDN, SOON, and ScanQA, with a significant margin of **29%** compared to the previous SoTA on CVDN. Furthermore, it also exhibits strong generalizability on unseen tasks.

2. Related Work

2.1. Vision-Language Navigation

Vision-language navigation has been extensively explored in the last few years, including a variety of tasks that require different aspects of embodied capabilities. R2R [3] requires the agent to navigate the rooms step by step, following fine-grained instructions, while Thomason *et al.* [52] introduce Cooperative Vision-and-Dialog Navigation (CVDN) which demands the agent to navigate based on a dialog history. Beyond navigation, SOON [61] and REVERIE [43] additionally require the agent to localize the objects queried in instructions. EQA [18, 55] emphasizes the ability to answer questions in a 3D environment by actively exploring the environment. To tackle these tasks, significant efforts have been invested. However, previous approaches [11, 12, 23, 25, 27, 31, 37] primarily focus on designing specialist models for individual tasks, and these models often struggle to generalize and transfer to other tasks. Our method, as an embodied generalist, addresses these tasks simultaneously via a single model and demonstrates strong generalization capabilities.

The work closely related to ours is MT-RCM [54], a multi-task model, designed to alleviate overfitting to specific datasets. On the other hand, our method primarily leverages the LLMs to enhance generalizability, and LLMs have been adapted to a broader range of tasks and datasets. Recently, some work [20, 58] has also explored the generalization and transferability of agents by utilizing off-the-shelf foundation models. In contrast to their approach, our method focuses on building a unified, end-to-end embodied learner, rather than a pipeline chaining up multiple independent models.

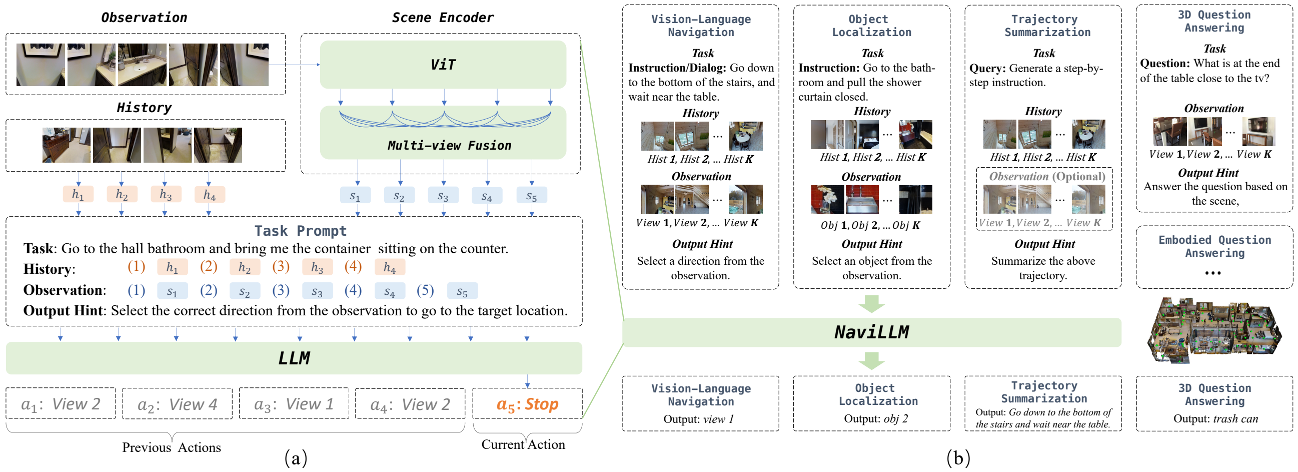

Observation
Scene Encoder
ViT
History
Multi-view Fusion
h_1 h_2 h_3 h_4
s_1 s_2 s_3 s_4 s_5
Task Prompt
Task: Go to the hall bathroom and bring me the container sitting on the counter.
History: (1) h_1 (2) h_2 (3) h_3 (4) h_4
Observation: (1) s_1 (2) s_2 (3) s_3 (4) s_4 (5) s_5
Output Hint: Select the correct direction from the observation to go to the target location.
LLM
a_1: View 2
a_2: View 4
a_3: View 1
a_4: View 2
a_5: Stop
Previous Actions
Current Action
(a)
Vision-Language Navigation
Task
Instruction/Dialog: Go down to the bottom of the stairs, and wait near the table.
History
Hist 1, Hist 2, ... Hist K
Observation
View 1, View 2, ... View K
Output Hint
Select a direction from the observation.
Object Localization
Task
Instruction: Go to the bathroom and pull the shower curtain closed.
History
Hist 1, Hist 2, ... Hist K
Observation
Obj 1, Obj 2, ... Obj K
Output Hint
Select an object from the observation.
Trajectory Summarization
Task
Query: Generate a step-by-step instruction.
History
Hist 1, Hist 2, ... Hist K
Observation (Optional)
View 1, View 2, ... View K
Output Hint
Summarize the above trajectory.
3D Question Answering
Task
Question: What is at the end of the table close to the tv?
Observation
View 1, View 2, ... View K
Output Hint
Answer the question based on the scene,
Embodied Question Answering
...
NaviLLM
Vision-Language Navigation
Output: view 1
Object Localization
Output: obj 2
Trajectory Summarization
Output: Go down to the bottom of the stairs and wait near the table.
3D Question Answering
Output: trash can
(b)

Figure 2. The overview of *NaviLLM*. The left figure presents the architecture and workflow of our model, while the right figure illustrates the schema-based instruction and multi-task learning process in our method.

2.2. Multimodal Instruction Tuning

Large Language Models (LLMs) [9, 15, 53] have revolutionized text understanding and text generation. Recent advancements [39, 56] further expanded their capabilities to digest visual inputs. For example, multimodal instruction tuning methods [5, 16, 36, 39, 59] have been widely proposed for 2D images [16] or videos [36]. One of the distinctions between our work and other multi-modal LLMs is that, our work is designed for embodied AI, including navigation and 3D understanding, which previous works don't consider. More recently, Hong et al. [28] introduced 3D-LLM by adapting LLMs to 3D data. However, it does not address the problem of calibrating LLMs for embodied navigation that requires the ability of sequential decision-making.

2.3. Large Language Models as Embodied Agents

Two lines of work have emerged in the exploration of integrating LLMs into embodied tasks. The first line focuses on translating visual information into textual format, which is then processed by the LLMs to generate plans [8, 30, 41, 58], landmarks [48], or code [38]. These methods leverage the inherent knowledge within the pre-trained, frozen LLMs. The second line [6, 7, 49], in contrast, directly fine-tunes LLMs on datasets that comprise action sequences of robotic manipulations. Similar to the second line of methods, our work also fine-tunes LLMs, yet focuses on addressing various tasks in embodied navigation, rather than robotic manipulations.

3. Method

3.1. Problem Formulation

In embodied navigation, an embodied agent situated in the 3D environment is required to complete tasks described in natural language. The agent leverages past trajectories and current observations to predict actions that enable task completion. The actions encompass navigation moves, bounding boxes for objects, and textual responses.

3.2. NaviLLM

NaviLLM is an embodied model grounded in LLM, comprising two modules, *i.e.*, a scene encoder and an LLM. As depicted in Figure 2 (a), the scene encoder takes the current visual observation as input, and transforms it into a series of scene representations (§3.2.1). Utilizing these scene representations, we construct various schemas for different tasks and these schemas serve as input for LLM, to produce the next action (§3.2.2).

3.2.1 Scene Encoding

The scene encoder extracts scene presentations given an observation composed of a set of images $\{I_i\}_{i=1}^n$, with each image representing a unique viewpoint. The visual encoder initially extracts visual features for each individual image via a Vision Transformer (ViT) [21]. These features from different viewpoints are then integrated via a multi-view fusion process, yielding scene representations $\{s_i\}_{i=1}^n$.
Visual Feature Extraction. A pre-trained ViT is utilized to extract visual features from images. Specifically, given an image I_i, it is first divided into a sequence of patches, with a special [CLS] token appended at the beginning. The sequence is then fed into the transformer network of the ViT. Finally, the last hidden states of the [CLS] token serve as visual features, which could be denoted as:

$$f_i = \mathrm{ViT}(I_i) \quad \text{for } i = 1, 2, ..., n, \tag{1}$$

where f_i is the visual features of the i-th viewpoint.

Multi-View Fusion. Upon acquiring the visual features, multi-view fusion is performed to model complex interdependencies among different viewpoints. The image features $\{f_i\}_{i=1}^n$ for all viewpoints are fed into a transformer encoder, to learn the spatial relationships between different viewpoints, formulated as:

$$\{s_i\}_{i=1}^n = \text{Transformer-Encoder}(\{f_i\}_{i=1}^n), \qquad (2)$$

where s_i is the scene representation for the i-th viewpoint. To enhance the scene representations, we also incorporate the angle and GPS information of each view into the scene encoding. We omit these details for brevity.

3.2.2 Schema-Based Instruction

Schema-based instruction was proposed in the language models for multi-turn dialog [10, 35, 46], serving as an effective way of generalizing to novel tasks. In the context of LLMs, we extend schema-based instruction to multimodal modeling so that it can digest multimodal information. Our schema is designed to be a unified format that can adapt to different data sources and enable flexibility for a wide range of tasks.

Task. It consists of a word sequence that the agent is expected to execute, which could manifest in various forms, such as a navigation instruction, an invisible object required to find, or a question raised by a user.

Observation. This refers to the visual observation at the current location of the agent. The observation schema consists of the scene representations $\{s_i\}_{i=1}^n$. To distinguish representations between different views, we prepend each representation with an ID, denoted by

$$[Emb(1), s_1, ..., Emb(n), s_n], \qquad (3)$$

where $Emb(i)$ is the embedding of the ID for the i-th view.

History. It records the sequence of past visual observations upon the t-th step. This schema provides a temporal context that helps the agent understand its past trajectory within the environment and the visual feedback associated with each decision. Given history representations $\{h^i\}_{i=1}^t$, we prepend each representation with an ID to indicate the order of the past observations. The History schema is constructed as

$$[Emb(1), h^1, ..., Emb(t), h^t], \qquad (4)$$

$Emb(i)$ is the embedding of the ID for the i-th step.

Output Hint. The schema hints at the output information that the agent is expected to produce, *e.g.*, an identifier for a desired viewpoint to move towards, an answer response to a question, or a summarization for a previous trajectory. This schema helps the model understand how to generate actions that align with the task requirements.

3.3. Multi-task Learning

As illustrated in Figure 2 (b), we summarize the key tasks for embodied navigation and transform these tasks into generation problems using schema-based instruction. Then we can optimize our model on different tasks with a unified cross-entropy objective. We detail each task as follows.

Vision-Language Navigation (VLN) requires the agent to navigate in 3D environment to accomplish a given task. We present the schemas for VLN as follows:

- *Task*: A navigation instruction with a brief task description.
- *Observation*: Scene representations of all reachable viewpoints at the current location.
- *Output Hint*: *e.g.*, *select a direction from the observation*.

The LLM takes the above schemas as input, to predict the ID of a viewpoint to move towards, where the ID is a number. As the agent moves, the history representations are updated with the scene representation corresponding to the agent's most recently selected viewpoint.

Object Localization. It requires identifying the correct object from a set of visible objects after the agent successfully reaches the destination. In addition to *History* schema, it also contains the following schemas:

- *Task*: An object localization command.
- *Observation*: Object representations of all visible objects at the current position. The object representations are extracted from a pre-trained ViT and subsequently converted into the same dimension as word embeddings.
- *Output Hint*: *e.g.*, *select an object from the observation*.

With these schemas, the agent is required to generate the ID of the selected object.

Trajectory Summarization. We follow [22] to include the task of synthesizing instructions from given trajectories. For this task, it shares the *History* and *Observation* schemas as VLN, where the *History* schema is optional depending on the dataset. Besides the two schemas, we also include:

- *Task*: A concise description of the style for summarizing, *e.g.*, fine-grained and coarse-grained.
- *Output Hint*: *e.g.*, *Summarize the above trajectory*.

3D Question Answering (3D-QA) asks the agent to answer a question in a 3D scene. Different from the previous tasks, in this task, the *History* schema is not required. The following schemas are provided for 3D-QA.

- *Task*: A question about an indoor scene.
- *Observation*: Scene representations of images from different positions. We also utilize the previous scene encoding to process the scene representations.
- *Output Hint*: *e.g.*, *Answer the question based on the scene*.

The model demands to generate a textual answer based on the above schemas.

Embodied Question Answering (EQA). The agent is asked to first navigate to the location referred by a question, and then respond to the question accordingly. We utilize the schemas of VLN and 3D-QA in two stages, respectively.

4. Experimental Setup

In this section, we first introduce the implementation details (§4.1). Then we describe the evaluation datasets, metrics, and baseline methods for VLN (§4.2), 3D-QA (§4.3) and EQA (§4.4), respectively.

4.1. Implementation Details

Model Details. We fine-tune the multi-view fusion module and LLM, where the former consists of a 2-layer transformer encoder with a hidden size of 1024 and the LLM is built upon Vicuna-7B-v0 [42]. The ViT in the scene encoder is EVA-CLIP-02-Large (428M) [51] and is kept frozen during the training phase. In addition, we leverage the object features extracted from ViT-B16 by Chen *et al.* [14].

Training Details. We follow the previous works [12, 13, 23] to employ a two-stage training strategy. Throughout both stages, we utilize the Adam optimizer with a learning rate of 3e-5. The model is trained for 10,000 steps in the pre-training stage and 5,000 steps in the multi-task fine-tuning stage with a batch size of 64. It takes approximately 80 hours with 8 Nvidia A100 GPUs. During the testing phase, we employ a sampling strategy with a temperature of 0.01 for action generation in the SOON and REVERIE tasks, to encourage more exploration. For other tasks, we opt for a greedy strategy in generating actions.

Training Data. In the pre-training stage, we perform teacher-forcing training on the combined dataset from CVDN, SOON, R2R, REVERIE, ScanQA, and augmented data from R2R and REVERIE. In the multi-task fine-tuning stage, we alternate between teacher forcing and student forcing on the combined dataset from CVDN, SOON, R2R, REVERIE, ScanQA, and LLaVA-23k [39].

For object localization, we utilize the corresponding annotations from REVERIE and SOON. For trajectory summarization, we convert the instruction-trajectory pairs of VLN datasets into trajectory-instruction pairs, where the trajectory serves as input and the instruction as output. As for 3D-QA, in addition to ScanQA, we also construct question-answer pairs from the fine-grained annotations on R2R [26]. In concrete, the constructed questions ask the model to predict the corresponding sub-instruction for a selected viewpoint. Additionally, we held out the task of EQA to verify the generalization capability on out-of-domain tasks.

4.2. Setup for VLN

Datasets. We adopt four datasets, each addressing distinct challenges posed by VLN. These datasets are split into train, val-seen, val-unseen, and test sets according to environments.

- **CVDN** [52] requires the agent to navigate towards the target based on a dialog history, thereby requiring the ability to comprehend the dialog and interpret it as actions.
- **SOON** [61] asks the agent to locate a thoroughly described object, which necessitates intricate alignment between rich semantic descriptions and the corresponding visual cues.
- **R2R** [3] demands the agent to navigate following a step-by-step instruction. To make effective decisions, it requires the agent to dynamically track the progress, demanding fine-grained alignment between history and instructions.
- **REVERIE** [43] requires the agent to localize a distant target object according to a concise high-level instruction.

Metrics. We follow [3] to evaluate our method on the following metrics: 1) Sucess Rate (**SR**), whether the agent successfully reaches the target location within a predefined distance threshold. 2) Success Rate Weighted by Path Length (**SPL**), calculated as the SR weighted by the ratio of the ground truth length and actual path length. 3) Oracle Success Rate (**OSR**), SR given the oracle stop strategy. 4) Trajectory Length (**TL**), the overall distance covered by the agent during navigation. 5) Goal Process (**GP**), the progress in meters towards the goal. GP is adopted for the CVDN dataset, while SPL is employed as the primary evaluation metric for other datasets.

Baseline Methods. We compare our method with the latest SoTA methods on the CVDN, SOON, R2R, and REVERIE datasets. We do not consider methods with pre-exploration (e.g., AuxRN [60], RREx-BoT [50]), and models augmented by new environments (e.g., HM3D-AutoVLN [13]).

4.3. Setup for 3D-QA

Dataset. ScanQA dataset [4] is a widely used dataset for 3D-QA, which is divided into train, val, and test sets. Here, we use val and 'test w/ objects' sets for comparison. For the results on 'test w/o objects' set, please refer to the appendix.

Metrics. We follow [28] to evaluate our method with Exact Match (**EM**), METEOR, ROUGE-L, CIDER, and BLEU-4.

Baseline Methods. We include some representative methods for comparison, including VoteNet+MCAN [57], ScanRefer+MCAN [57], and 3D-LLM [28]. 3D-LLM is the current SoTA method on the ScanQA benchmark.

4.4. Setup for EQA

Dataset. We test the zero-shot inference capability on the val split of the MP3D-EQA [55] dataset. Since MP3D-EQA is generated from functional programs, there are some data with inaccurate endpoints. Therefore, we manually check the dataset and filter out those invalid data in our experiments.

Metrics. We report SR and SPL for the navigation phase, and Accuracy (**ACC**) for the question-answering phase.

Baseline Methods. We compare our *NaviLLM* with the fully-supervised VQA model [18] and zero-shot DUET models [12] separately trained on R2R, REVERIE, and SOON.

	CVDN		**SOON**		**R2R**		**REVERIE**		**ScanQA**	
	Val-U	Test	Val-U	Test	Val-U	Test	Val-U	Test	Val	Test
Separate Model For Each Task										
PREVALENT [25]	3.15	2.44	-	-	53	51	-	-	-	-
HOP [44]	4.41	3.24	-		57	59	26.11	24.34	-	-
HAMT [11]	5.13	5.58	-	-	61	60	30.20	26.67	-	-
VLN-BERT [27]	-	-	-	-	57	57	24.90	23.99	-	-
GBE [61]	-	-	13.34	9.23	-	-	-	-	-	-
DUET [12]	-	-	22.58	21.42	60	58	33.73	36.06	-	-
Meta-Explore [31]	-	-	-	25.80	62	**61**	34.03	-	-	-
AZHP [23]	-	-	-	-	61	60	**36.63**	35.85	-	-
VLN-SIG [32]	5.52	5.83	-	-	62	60	-	-	-	-
VLN-PETL [45]	5.69	6.13	-	-	60	58	27.67	26.73	-	-
BEV-BERT [1]					**64**	60	36.37	**36.41**	-	-
3D-LLM [28]	-	-	-	-	-	-	-	-	20.5	19.1
Unified Model For All Tasks										
MT-RCM+Env [54]	4.65	3.91	-	-	49	40	-	-	-	-
NaviLLM	**6.16**	**7.90**	**29.24**	**26.26**	59	60	35.68	32.33	**23.0**	**26.3**

Table 1. Overall comparison with state-of-the-art methods on all tasks. 'Val-U' denotes val-unseen split. We report SPL for CVDN, SOON, R2R, and REVERIE, and report Accuracy for ScanQA.

	ScanQA-Val					ScanQA-Test				
	EM	ROUGE-L	METEOR	CIDER	BLEU-4	EM	ROUGE-L	METEOR	CIDER	BLEU-4
VoteNet+MCAN [57]	17.3	29.8	11.4	54.7	6.2	19.7	30.9	12.0	58.2	6.0
ScanRefer+MCAN [57]	18.6	30	11.5	55.4	7.9	20.6	30.7	11.9	57.4	7.5
ScanQA [4]	21.0	33.3	13.1	64.9	10.1	23.5	34.3	13.5	67.3	12.0
3D-LLM (flamingo) [28]	20.4	32.3	12.2	59.2	7.2	23.2	34.8	13.5	65.6	8.4
3D-LLM (BLIP2-flant5) [28]	20.5	35.7	14.5	69.4	12.0	19.1	35.3	14.9	69.6	11.6
NaviLLM (Ours)	**23.0**	**38.4**	**15.4**	**75.9**	**12.5**	**26.3**	**40.2**	**16.6**	**80.8**	**13.9**

Table 2. Detail comparison with state-of-the-art methods on ScanQA.

5. Experimental Results

We conduct a series of experiments to answer three critical questions about *NaviLLM*: (1) Can *NaviLLM*, when trained with diverse tasks, demonstrate superior performance compared to existing SoTA methods (§5.1)? (2) How well does *NaviLLM* generalize to unseen tasks, compared to previous task-specific models (§5.2)? (3) What is the impact of each component in our method (§5.3)? Lastly, we also provide visualization for *NaviLLM* on unseen scenes and tasks (§5.4).

5.1. Comparision with SoTA Methods

Delivering SoTA Results with a Single Model. We present the comparison across all tasks in Table 1. Our single model achieves SoTA performance on the test sets of the CVDN, SOON, and ScanQA datasets, and demonstrates comparable results to the latest SoTA methods on R2R and REVERIE.

Significant Improvement on CVDN Can Be Credited to Our Innovative Design. Compared to the 6.13 GP of VLN-PETL [45], our method shows a significant increase in the GP at 7.90, winning the first place on the leaderboard of CVDN. Compared to other datasets, the improvement on CVDN is the most pronounced. We attribute this significant improvement primarily to two factors: 1) The dialog structure in CVDN is relatively complex, and the knowledge inherited from LLM in our model can better help comprehend the dialog. 2) Given that the size of the CVDN dataset is smaller compared to other datasets, the unification of these datasets effectively mitigates the issue of data scarcity.

Our Method Also Excels in 3D Tasks. As shown in Table 2, our method achieves SoTA results on the val and test sets in all metrics. It obtains a 26.3% EM, with a 7.2% improvement over 3D-LLM, which is specially designed for 3D tasks.

Better Performance on Tasks with Complex Instructions. We count the average length of instructions across different datasets, with CVDN averaging 81.6 words per instruction, SOON averaging 38.6 words, R2R averaging 29 words, and REVERIE averaging 18 words. This reflects the complexity of the instructions to some extent. We notice that our method exhibits superior performance on datasets with com-

	CVDN		SOON				REVERIE			
	TL	**GP↑**	TL	**OSR↑**	**SR↑**	**SPL↑**	TL	**OSR↑**	**SR↑**	**SPL↑**
DUET (R2R)	21.12	3.38	26.83	7.64	4.66	2.84	7.88	29.11	24.91	20.00
DUET (REVERIE)	76.13	3.30	33.72	20.86	10.24	6.06	-	-	-	-
DUET (SOON)	48.61	2.40	-	-	-	-	38.10	43.45	10.91	3.64
NaviLLM	26.37	**4.46**	28.66	**33.11**	**19.81**	**14.29**	18.96	**51.47**	**28.10**	**21.04**

Table 3. Held-out results on val-unseen splits of CVDN, SOON and REVERIE. We only perform the multi-task fine-tuning for held-out experiments. Trajectory Length (TL) serves as a statistical indicator rather than an evaluation metric.

#		GT Path	Navigation			QA
			TL	**SR↑**	**SPL↑**	**ACC↑**
1	DUET (R2R) [12]		16.47	47.00	30.51	-
2	DUET (REVERIE) [12]		12.59	39.22	11.47	-
3	DUET (SOON) [12]		47.01	17.43	3.91	-
4	***NaviLLM*** (Ours)		14.15	47.78	35.60	44.5
5	EQA (habitat-lab)† [18]	✓	-	-	-	46.0
6	***NaviLLM*** (Ours)	✓	-	-	-	47.4

Table 4. Zero-shot inference results on MP3D-EQA. 'GT Path' means using the ground truth trajectory for question answering. † indicates the method is finetuned on the training set of MP3D-EQA. Trajectory Length (TL) serves as a statistical indicator rather than an evaluation metric.

plex instructions, such as CVDN, SOON, and R2R, achieving performance better than or comparable to SoTA methods. However, there is still a slight gap with DUET on datasets with relatively simple instructions, such as REVERIE. This may indicate that our method possesses excellent instruction comprehension capabilities, which helps improve the performance on tasks with complex instructions.

***NaviLLM* Demonstrates An Excellent Object Localization Capability.** In the object localization task of REVERIE, our method achieves 19.83% RGS and 16.04% RGSPL, and outperforms the 14.88% and 13.08% achieved by HAMT [11], demonstrating an excellent object localization capability. Given that existing methods typically integrate object features with image features, we believe that our method could be further enhanced by combining these features.

5.2. Generalization Ability on Unseen Tasks

We evaluate the zero-shot inference performance of our method on unseen tasks, and compare it with zero-shot DUET models (DUET (R2R), DUET (REVERIE), and DUET (SOON)), with each model being separately trained on its corresponding dataset.

Generalize to Out-of-Domain VLN Tasks. We conduct held-out experiments to verify the generalization ability to out-of-domain VLN tasks. Specifically, we individually exclude CVDN, SOON, and REVERIE from the training set, train three separate models, and then test their zero-shot performance on the respective excluded datasets. DUET specially designs different hyper-parameters and pre-training schemes for each VLN task, giving the model strong in-domain navigation capabilities. However, such learned task-specific agents lack out-of-domain generation abilities. As illustrated in Table 3, our method significantly outperforms DUET on CVDN and SOON, improving 32% of GP on CVDN and 136% SPL on SOON, respectively. Since instructions in REVERIE are relatively simpler and similar to R2R, DUET (R2R) delivers an SR of 24.91% on REVERIE, but we still achieve a better SR of 28.10%. This demonstrates that our schema-based instruction and multi-task learning empower the out-of-domain generation abilities.

Skill Combination for EQA. We perform a zero-shot evaluation on MP3D-EQA to show that *NaviLLM* can combine the learned navigation and question-answering ability to solve more complex tasks. We ask our agent to first execute the navigation process and then answer the question after reaching the goal. As illustrated in Table 4, our model achieves 47.78% SR and 35.60% SPL, surpassing DUET (R2R) by 5.1% in SPL (row 1 vs. 4). At the same time, it can also answer questions at a decent accuracy of 44.5%, while the DUET models are incapable of performing question answering. Moreover, when the ground truth trajectories are provided, our zero-shot model presents superior performance over the fully-supervised EQA model (rows 5 vs. 6).

5.3. Ablation

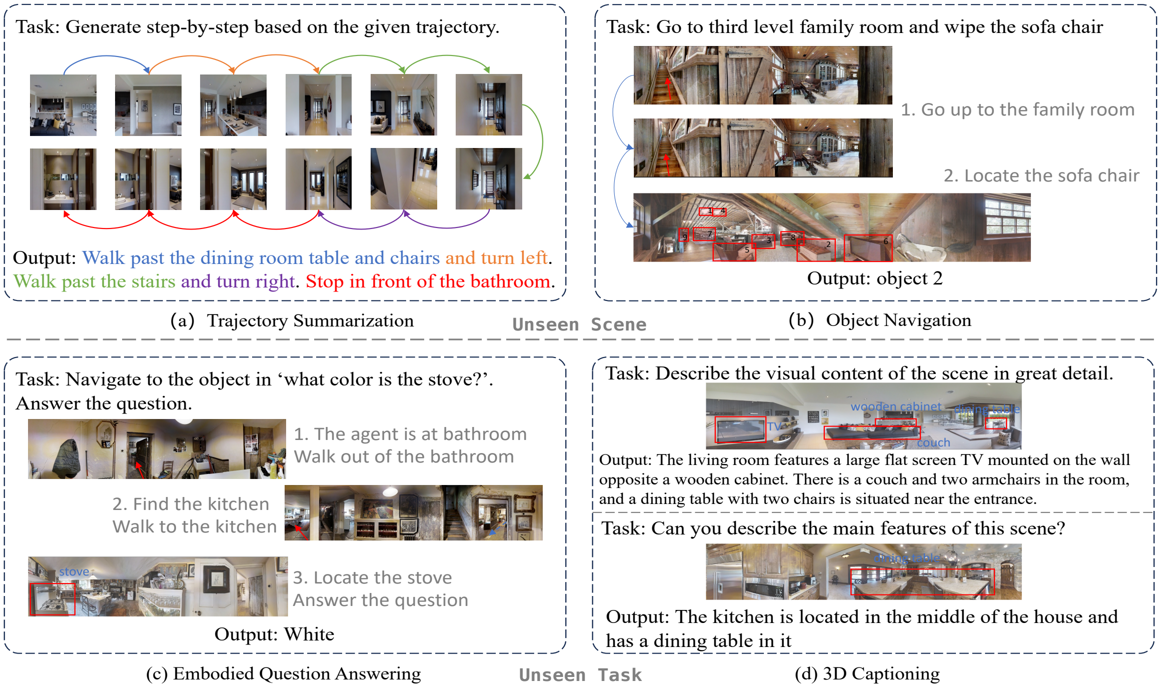
Multi-task Learning Enhances the Performance On All Tasks. Table 5 illustrates that multi-task learning improves performance on all tasks (row 1 vs. 3). This demonstrates that expanding the volume and diversity of training data is crucial for learning a generalist model for embodied navigation.

LLM plays a Key Role in Our Method. Comparing rows 2 and 3, we can observe a significant performance drop when the LLM is randomly initialized, underscoring the substantial role that the LLM plays.

Limited Benefits of Pre-Training on Augmented Data. Previous works [12, 23] have consistently shown notable improvements after pre-training on augmented data from R2R and REVERIE. However, comparing rows 3 and 4, we find only a slight enhancement on R2R, CVDN, and SOON after pre-training. We speculate that the quality of the data may play a more crucial role than its quantity for our method.

#	LLM	Multi-Task	Pretrain	CVDN	SOON		R2R		REVERIE		ScanQA	
				GP	SR	SPL	SR	SPL	SR	SPL	EM	ROUGE-L
1	✓			5.54	28.37	21.37	64	57	30.95	24.10	21.8	37.0
2		✓		3.64	20.73	17.39	49	40	31.49	26.87	12.4	22.8
3	✓	✓		5.91	35.44	28.09	**67**	58	**44.56**	**36.63**	**23.3**	**38.2**
4	✓	✓	✓	**6.16**	**38.33**	**29.24**	**67**	**59**	42.15	35.68	22.1	37.6

Table 5. Ablation study of *NaviLLM* across all tasks. 'LLM', 'Multi-Task', and 'Pretrain' denote the utilization of pretrained LLM weights for initialization, the execution of multi-task learning, and the performance of pre-training, respectively. The results reported are from the val-unseen splits for VLN tasks and the val split for ScanQA.


Task: Generate step-by-step based on the given trajectory.
Output: Walk past the dining room table and chairs and turn left. Walk past the stairs and turn right. Stop in front of the bathroom.
Task: Go to third level family room and wipe the sofa chair
1. Go up to the family room
2. Locate the sofa chair
Output: object 2
Unseen Scene
Task: Navigate to the object in 'what color is the stove?'. Answer the question.
1. The agent is at bathroom
Walk out of the bathroom
2. Find the kitchen
Walk to the kitchen
stove
3. Locate the stove
Answer the question
Output: White
Task: Describe the visual content of the scene in great detail.
wooden cabinet
dining table
TV
couch
Output: The living room features a large flat screen TV mounted on the wall opposite a wooden cabinet. There is a couch and two armchairs in the room, and a dining table with two chairs is situated near the entrance.
Task: Can you describe the main features of this scene?
dining table
Output: The kitchen is located in the middle of the house and has a dining table in it
Unseen Task

(a) Trajectory Summarization

(b) Object Navigation

(c) Embodied Question Answering

(d) 3D Captioning

Figure 3. The visualization for our method on unseen scenes and unseen tasks. In Figure (a), lines and text of the same color represent sub-trajectories and their corresponding sub-instructions. In Figures (b) and (c), the text in gray is the description of the actions of the agent during navigation, while the red arrow indicates the direction that the agent moves towards.

5.4. Visualization

Figure 3 (a) and (b) are examples of trajectory summarization and object navigation on unseen scenes, respectively. The first example illustrates that our model can generate accurate step-by-step instructions given trajectories, which could be further used for data augmentation. Figure 3 (c) and (d) respectively present examples of EQA and 3D captioning, which are not encountered in training data, demonstrating the generalizability of *NaviLLM*. Specifically, as shown in Figure 3 (d), our model is capable of producing captions of varying granularity according to the instructions.

6. Conclusion

In this paper, we present the first generalist model for embodied navigation, *NaviLLM*, which adapts LLMs to a variety of tasks by introducing schema-based instruction. Benefiting from this design, we unify diverse tasks into a generation problem, allowing our model to utilize data sources from various datasets. Our experiments show that our single model can achieve SoTA results on CVDN, SOON, and ScanQA, and comparable performance to the latest models on R2R and REVERIE. Moreover, it also demonstrates strong generalizability and presents promising results on unseen tasks.

7. Acknowledgements

This work was supported by National Key R&D Program of China (Project No. 2022ZD0161200, 2022ZD0161201). It was also partially funded by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. Liwei Wang is a Principal Investigator of CPII under the InnoHK. This work has also been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 24200223) as well as partially supported by Hong Kong Innovation and Technology Commission Project No. ITS/228/22FP.

References

[1] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6

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

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1, 2, 5

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 2, 5, 6

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2, 3

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 2, 3

[8] 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*, pages 287–318. PMLR, 2023. 3

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 2, 3

[10] Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7521–7528, 2020. 4

[11] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 5834–5847. Curran Associates, Inc., 2021. 1, 2, 6, 7

[12] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16537–16547, 2022. 2, 5, 6, 7

[13] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *ECCV*, 2022. 5

[14] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *European Conference on Computer Vision*, pages 638–655. Springer, 2022. 1, 5

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 2, 3

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 3

[17] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018. 1

[18] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 7

[19] Harm De Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018. 1

[20] Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*, 2022. 2

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[22] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 4

[23] Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. Adaptive zone-aware hierarchical planner for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14911–14920, 2023. 2, 5, 6, 7

[24] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643, 2021. 1

[25] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146, 2020. 1, 2, 6

[26] Yicong Hong, Cristian Rodriguez, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3360–3376, 2020. 5

[27] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, 2021. 1, 2, 6

[28] 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, 2023. 2, 3, 5, 6

[29] Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhaes, Jason Baldridge, and Eugene Ie. Transferable representation learning in vision-and-language navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7404–7413, 2019. 1

[30] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR, 2023. 3

[31] Minyoung Hwang, Jaeyeon Jeong, Minsoo Kim, Yoonseon Oh, and Songhwai Oh. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6683–6693, 2023. 2, 6

[32] Mohit Bansal Jialu Li. Improving vision-and-language navigation by generating future-view image semantics. In *CVPR*, 2023. 1, 6

[33] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017. 1

[34] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 1

[35] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. *arXiv preprint arXiv:2109.07506*, 2021. 4

[36] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 3

[37] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2583–2592, 2023. 2

[38] 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023. 3

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 5

[40] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 259–274. Springer, 2020. 1

[41] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and

Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023. 3

[42] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 5

[43] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 1, 2, 5

[44] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: history-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15418–15427, 2022. 6

[45] Yanyuan Qiao, Zheng Yu, and Qi Wu. Vln-petl: Parameter-efficient transfer learning for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15443–15452, 2023. 6

[46] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8689–8696, 2020. 4

[47] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 1

[48] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR, 2023. 3

[49] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 2, 3

[50] Gunnar A Sigurdsson, Jesse Thomason, Gaurav S Sukhatme, and Robinson Piramuthu. Rrex-bot: Remote referring expressions with a bag of tricks. *arXiv preprint arXiv:2301.12614*, 2023. 5

[51] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 5

[52] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 1, 2, 5

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 3

[54] Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. Environment-agnostic multitask learning for natural language grounded navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 413–430. Springer, 2020. 2, 6

[55] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019. 1, 2, 5

[56] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 3

[57] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 5, 6

[58] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023. 2, 3

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 2, 3

[60] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020. 5

[61] Fengda Zhu, Xiwen Liang, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration, 2021. 1, 2, 5, 6