From End-to-End to Step-by-Step: Learning Composable Navigation Primitives for Vision-Language Navigation

Yang Chen*†‡, Xiaohan Yi‡, Danyang Peng*†, Zhuo-Xia Wang*§, Shi-Yu Tian*§, Jiang Bian‡ and Lan-Zhe Guo*†
*State Key Laboratory of Novel Software Technology, Nanjing University, China
†School of Intelligence Science and Technology, Nanjing University, China
‡Microsoft Research Asia, China
§School of Artificial Intelligence, Nanjing University, China
Email: guolz@nju.edu.cn

***Abstract*—Vision-Language Navigation (VLN) is a fundamental research direction in Embodied AI. Recent VLN studies based on Multimodal Large Language Models (MLLMs) have widely adopted an end-to-end training paradigm on long-horizon instruction datasets. Our analysis shows that, although existing VLN models achieve promising results on long-horizon instructions, they still perform poorly on simple navigation instructions, such as *move to the desk*. Moreover, current end-to-end training and evaluation protocols overlook the acquisition of fundamental navigation skills. We argue that the reliable execution of complex embodied planning and long-horizon navigation tasks depends on the model's ability to compose simple navigation primitives. To address this issue, we propose a primitive-based learning paradigm. We introduce Vision-Language-Move-Base (VLMB), the first controllable dataset centered on the "move-to" primitive, covering 206 scenes and 873 objects. We further construct the corresponding VLMB-Bench to systematically train and evaluate models' mastery of fundamental navigation skills. Based on VLMB, we develop Move-to-Anything (M2A), a model equipped with a Dual-Scale Memory mechanism that balances historical context and current observations. Experiments demonstrate that VLMB enables M2A to learn transferable navigation primitives, yielding strong primitive-level performance, zero-shot generalization to LH-VLN, and better data efficiency than generic VLN supervision. The project page is: https://move2anything.netlify.app/.**

I. Introduction

Vision-Language Navigation (VLN) aims to build autonomous agents that can interpret natural-language instructions and navigate complex environments. Traditional approaches typically adopt modular pipelines that decompose navigation into perception, understanding, and planning modules [3, 19, 9]. However, such decoupled designs often introduce bottlenecks in the flow of information between perception and planning. With the emergence of Multimodal Large Language Models (MLLMs), the field has increasingly shifted toward end-to-end learning [22, 4, 10]. Representative systems, such as StreamVLN [14] and JanusVLN [16], train general-purpose MLLMs on large-scale datasets [1, 6, 5]. By directly aligning visual observations with linguistic queries, these methods provide stronger semantic and spatial understanding for embodied intelligence.

Despite this progress, current end-to-end VLN models are primarily trained and evaluated on legacy benchmarks such as R2R [1], which emphasize long-horizon and complex instruction following. These datasets were originally designed to promote high-level instruction-following ability in navigation models. However, we argue that training exclusively on such complex instruction datasets can cause MLLMs to bypass the acquisition of fundamental navigation skills, such as *Move to the cabinet*. To systematically examine this issue, we evaluate representative SOTA models under simple navigation instructions in Section II. Surprisingly, their performance on simple instructions can be even lower than that on complex instructions. This observation suggests that current training paradigms overlook the need to endow MLLMs with basic navigation skills. Consequently, high success rates on prior benchmarks may partly reflect "blind success", where models exploit dataset biases to align observations with instructions, rather than demonstrating genuine navigation competence.

In complex embodied planning tasks [7, 15], navigation is often abstracted by primitives such as "move to" or "navigate to". This abstraction is also consistent with human action execution, where long-distance navigation is achieved through the sequential execution of a series of basic primitive instructions. Therefore, we focus on mastering the "move-to" primitive and regard it as a foundational skill for building VLN agents.

To enable a step-by-step training paradigm, we introduce **Vision-Language-Move-Base** (VLMB), a dataset specifically designed to enhance the fundamental "move-to" primitive in VLN. Our automated data generation pipeline synthesizes instructions from 206 semantic scenes in MP3D [2] and HM3D-Semantic [11], covering over 873 distinct target instances. The construction process first extracts basic navigation primitives through rule-based procedures, and then expands them with MLLMs to incorporate rich spatial and semantic constraints. This design captures the linguistic diversity and ambiguity inherent in natural language. A rigorous interactive verification process further ensures data quality and refines the model's

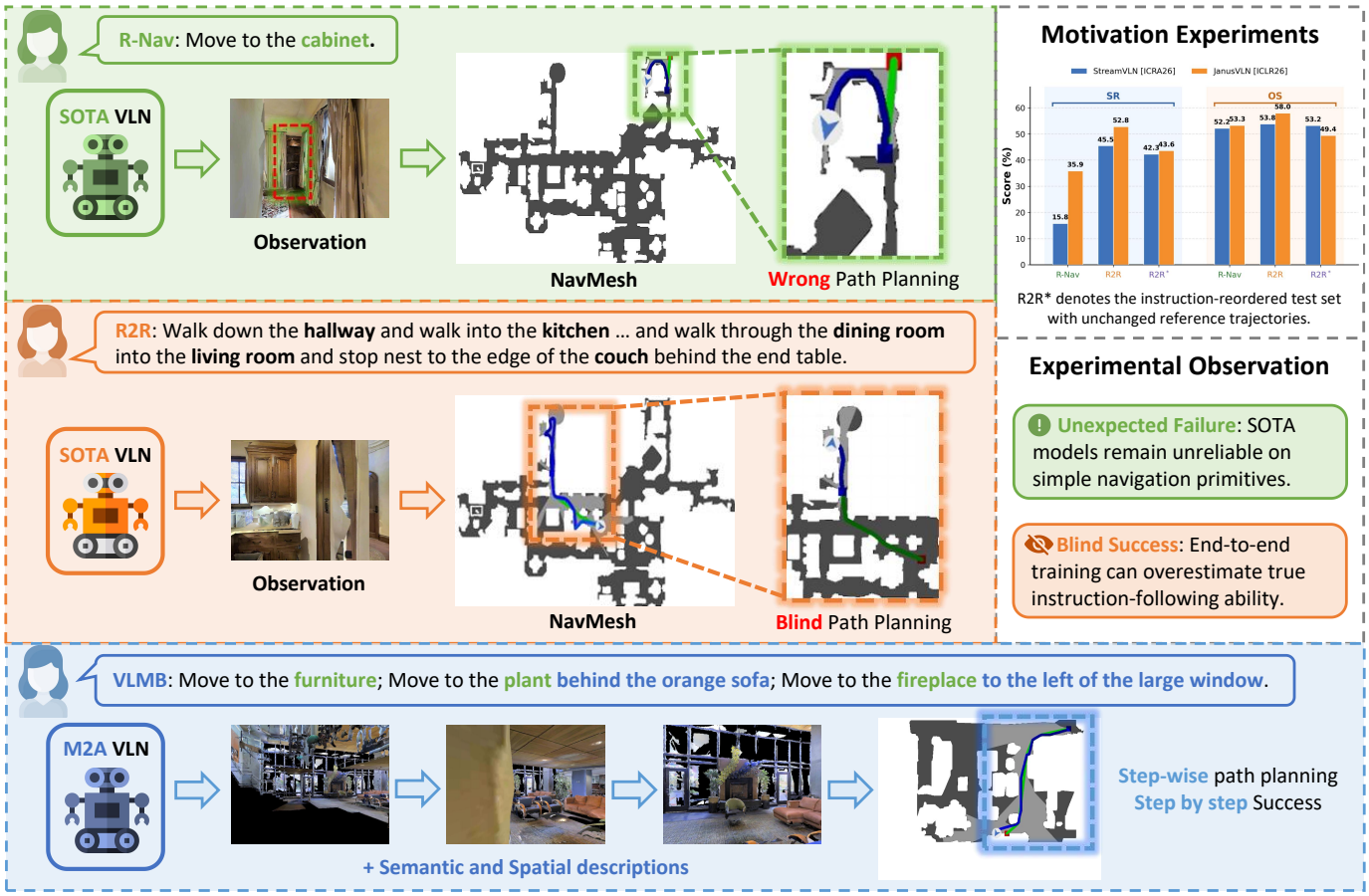

Fig. 1. Motivation. Existing SOTA VLN models can succeed on long-horizon benchmarks while failing on simple navigation primitives. M2A trained on VLMB better follows spatial and semantic target descriptions.

understanding of both object-level and scene-level cues. To provide an objective and fair evaluation of whether models can correctly execute navigation actions, we further construct navigation instruction test sets, R-NAV and VLMB-Bench.

Building on general MLLM training strategies, we design a **Dual-Scale Hierarchical Memory** mechanism that combines spatiotemporal grounding with historical context, allowing the model to balance current observations and past information. Models trained on VLMB achieve leading performance on VLMB-Bench and LH-VLN [12]. Our main contributions: (1) We identify a critical bottleneck in end-to-end VLN systems: current models struggle with fundamental navigation primitives, often producing actions that mismatch instructions. (2) We construct VLMB and VLMB-Bench, a primitive-centric training dataset and evaluation benchmark, and propose an automated step-by-step data generation pipeline for navigation skill acquisition. (3) We introduce Move-to-Anything (M2A), which outperforms existing SOTA models with substantially lower training cost, demonstrating the effectiveness of step-wise primitive-centric training.

II. Motivation

To ground our analysis, we conduct two motivation experiments before presenting the full evaluation. These experiments are designed to validate our motivation and design principles, and to reveal why our proposed method is necessary for VLN. In this section, we aim to answer the following questions.

a) Can SOTA models handle simple navigation?: We construct **R-Nav**, a raw-instruction navigation benchmark using native object-label instructions from MP3D and HM3D. R-Nav tests whether agents can execute simple commands such as *move to the table*. Surprisingly, SOTA end-to-end VLN models perform worse on R-Nav than on complex VLN-CE tasks, revealing that long-horizon training does not guarantee mastery of primitive navigation.

b) Does end-to-end VLN measure instruction following?: We further perturb R2R instructions by reordering instruction segments while keeping the original trajectories unchanged. If a model truly follows instructions, such temporal inconsistency should strongly reduce success. Instead, models retain relatively high success rates, indicating that they may exploit visual priors or dataset biases. Together with the gap between Oracle Success Rate (OS) and Success Rate (SR), this suggests a form of "blind success" and motivates primitive-level supervision and evaluation.

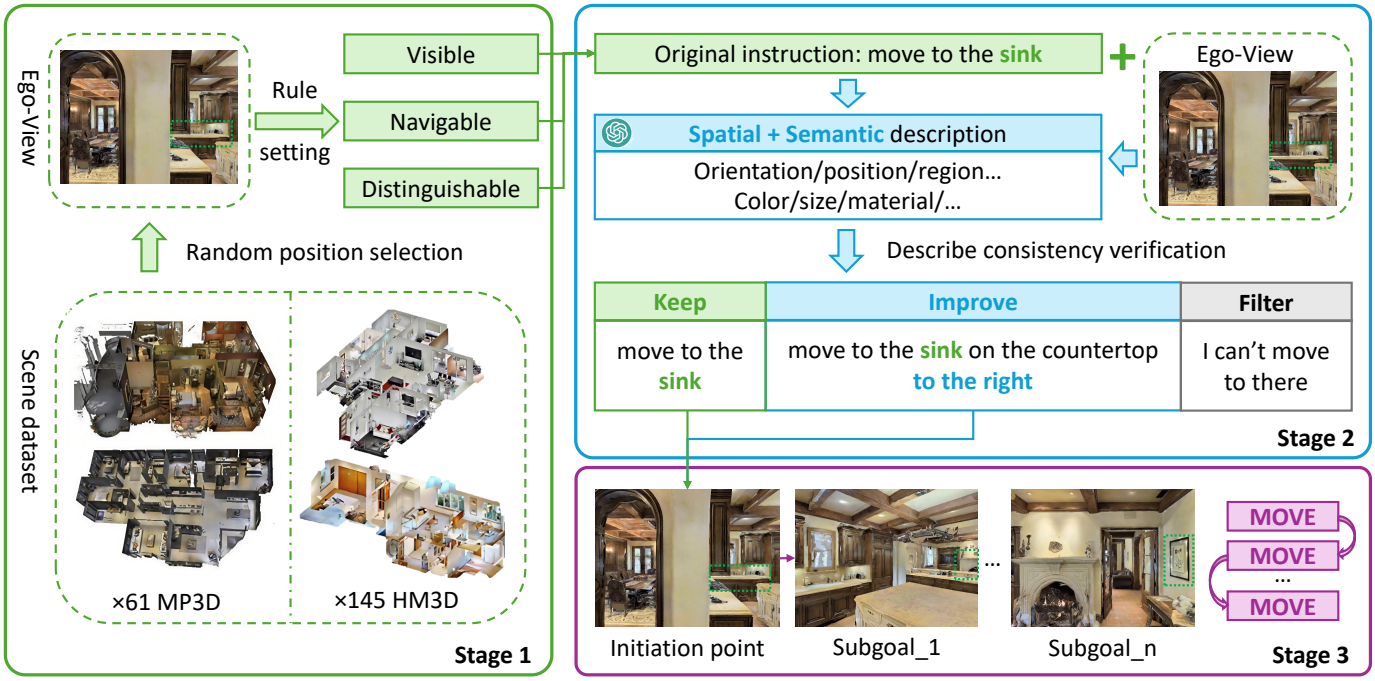

Fig. 2. VLMB data generation pipeline. We sample reachable visible objects, enrich primitive instructions with spatial and semantic descriptions, and compose verified primitives into multi-step instructions.

III. Dataset: Vision-Language-Move-Base

A. Problem Setup

The agent receives egocentric RGB observations and a language instruction, then predicts discrete actions: turning, moving forward, or stopping. An episode succeeds if the agent stops within a predefined radius of the target object. VLMB is built from 61 MP3D scenes and 145 HM3D scenes, while VLMB-Bench uses disjoint unseen scenes for evaluation.

B. Data Construction

VLMB is generated in three stages. First, we automatically sample reachable and visually clear target objects from Habitat-Sim scenes. Targets are required to be visible, semantically meaningful, and reachable through the NavMesh. Second, GPT-4o verifies the target visibility and enriches the raw instruction with attributes and spatial relations, reducing ambiguity and increasing linguistic diversity. Third, verified primitives are composed into structured multi-step instructions, enabling controlled long-horizon evaluation.

We initially collect 30K samples and filter them into a curated 15K training set. Each sample contains both a raw R-Nav instruction and an enriched VLMB instruction. We balance scenes, room types, goals, and instruction patterns, and manually verify a subset to confirm annotation quality. VLMB-Bench is built from unseen scenes with additional manual filtering to support fair evaluation.

IV. Method: Move-to-Anything

M2A is trained on R-Nav and VLMB. To handle long-horizon navigation under a bounded visual token budget, it uses a **Dual-Scale Memory** architecture and **Explicit Spatiotemporal Grounding**.

A. Dual-Scale Memory

Dual-Scale Memory separates visual history into a short-term High-Fidelity Perceptual Buffer (PB) and a long-term Object-Aware Episodic History (OH). PB stores recent frames as uncompressed visual tokens, preserving fine-grained perception for local movement. OH stores sparse keyframes selected by semantic novelty, retaining landmarks and scene transitions while avoiding redundant history.

Given the current observation feature f_t and the latest keyframe feature f_k, M2A measures object salience with cosine distance:

$$\mathcal{S}_t = 1 - \frac{f_t^\top f_k}{\|f_t\|_2 \|f_k\|_2}. \tag{1}$$

A frame is written into OH only when $\mathcal{S}_t > \tau$. This keeps long-term memory compact while preserving distinctive visual evidence.

B. Explicit Spatiotemporal Grounding

Long-context visual inputs mix current frames and historical keyframes. To make this structure explicit, M2A adds temporal and segment embeddings to each visual token:

$$\tilde{x} = x + E_{\text{time}}[t] + E_{\text{seg}}[s]. \tag{2}$$

The temporal embedding encodes order, while the segment embedding marks whether a token comes from the recent buffer or long-term history. This helps the model distinguish immediate perceptual evidence from retrieved context.

TABLE I
COMPARISON WITH PREVIOUS METHODS ON R-NAV AND VLMB-BENCH. NAVID WAS NOT INCLUDED IN THE HM3D COMPARISON BECAUSE IT WAS NOT TRAINED ON THE HM3D SCENES.

Method	R-Nav								VLMB-Bench							
	MP3D				HM3D				MP3D				HM3D			
	SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	OS↑	NE↓
MLLM-driven Agents																
Video-LLaVA [8]	14.1	12.5	34.7	4.91	15.2	14.6	43.1	5.02	13.4	12.3	34.3	4.96	15.3	14.1	39.2	5.21
LLaVA-NEXT [20]	12.1	10.5	31.5	5.46	14.2	13.5	39.4	5.45	11.2	10.5	33.1	5.34	12.3	11.5	35.7	5.56
GPT-4o	11.6	11.4	13.6	5.23	22.6	22.5	24.8	4.53	15.3	15.2	16.9	4.98	20.29	20.29	20.45	4.44
GPT-5-mini	14.1	14.1	14.2	5.02	25.4	25.3	27.1	4.29	16.1	16.1	17.1	4.92	25.77	25.77	26.5	4.36
End-to-end VLN Models																
NaVid [18]	46.7	42.6	59.8	4.74	–	–	–	–	43.8	39.0	60.1	4.89	–	–	–	–
StreamVLN [14]	15.8	10.8	52.2	8.48	32.4	28.2	49.3	5.43	22.0	16.0	60.9	7.24	44.2	39.0	59.3	4.43
JanusVLN [16]	35.9	31.5	53.3	6.28	37.2	31.5	53.3	6.16	42.8	37.8	64.6	4.77	45.3	39.9	67.0	4.62
M2A (ours)	**62.5**	**61.4**	**69.1**	**3.09**	**70.6**	**69.6**	**75.8**	**2.70**	**60.6**	**59.6**	**67.6**	**3.21**	**71.4**	**70.3**	**76.3**	**2.71**

TABLE II
COMPARISON WITH PREVIOUS METHODS ON MULTI-STEP NAVIGATION TASK.

Method	R-Nav				VLMB-Bench				LH-VLN				
	SR↑	SPL↑	OS↑	NE↓	SR↑	SPL↑	OS↑	NE↓	SR↑	NE↓	ISR↑	CSR↑	CGT↑
StreamVLN [14]	11.5	2.36	50.4	10.3	20.4	5.25	61.9	10.26	10.9	7.87	17.1	13.7	12.7
JanusVLN [16]	22.4	14.73	34.8	7.60	21.7	5.64	35.3	9.76	17.6	7.16	26.7	21.8	20.7
NaVid [18]	31.9	28.5	44.2	7.24	21.3	6.41	34.6	9.64	-	-	-	-	-
M2A (ours)	**36.3**	**29.1**	**54.9**	**6.24**	**35.4**	**28.7**	**53.1**	**6.38**	**35.6**	**6.14**	**39.5**	**34.8**	**29.2**

V. EXPERIMENTS

A. Experiment Setup

We evaluate M2A on R-Nav and VLMB-Bench for primitive execution, a compositional multi-step split for long-horizon composition, and LH-VLN [12] for external zero-shot evaluation. R-Nav uses raw object-name instructions to test whether agents can execute basic move-to commands, while VLMB-Bench further evaluates spatially and semantically enriched instructions in unseen scenes.

M2A is built on LLaVA-Video-7B [21], with Qwen2-7B [13] as the language backbone and a SigLIP encoder [17] for visual perception. At each step, the agent receives front-view RGB observations and predicts one of four discrete actions: *turn left*, *turn right*, *move forward*, and *stop*. We report standard navigation metrics including SR, SPL, OS, and NE.

B. Can M2A Master Primitive-Level Navigation Skills?

Table I compares M2A with general-purpose MLLM agents and representative end-to-end VLN models on R-Nav and VLMB-Bench. The results show that strong multimodal understanding does not directly translate into executable navigation. Training-free MLLM agents, including GPT-4o and GPT-5-mini, achieve low success rates across both benchmarks, indicating that current foundation models still lack reliable primitive-level action grounding.

Although end-to-end models are trained on a large amount of VLN data, they also exhibit limited performance on these tasks. This suggests that high-level VLN training alone may not provide sufficient supervision for learning fine-grained, executable navigation primitives. M2A achieves the best results across all R-Nav and VLMB-Bench settings. These gains demonstrate that VLMB effectively improves fundamental navigation skills, especially for object grounding, spatial disambiguation, and visually guided local decision-making.

C. Can VLMB-Trained Primitives Generalize to Long-Horizon Navigation?

We next evaluate whether primitive-level skills learned from VLMB transfer to longer-horizon navigation. As shown in Table II, M2A is tested on compositional multi-step VLMB tasks and LH-VLN [12]. M2A consistently outperforms prior VLN models on the compositional multi-step tasks, indicating that VLMB-trained primitives can be composed into longer instruction-following behaviors. On LH-VLN, M2A further surpasses StreamVLN and JanusVLN under the same zero-shot setting, improving SR over JanusVLN by 18.0 points. These results show that M2A learns transferable navigation primitives rather than merely fitting VLMB-Bench.

VI. CONCLUSION

This work identifies a critical weakness in end-to-end VLN systems: despite success in long-horizon navigation, they struggle with basic primitives. We introduce VLMB, the first dataset designed to enhance fundamental navigation capabilities. Centered on the "move-to" primitive, VLMB constructs a complete training set and benchmark via spatial-semantic augmentation, enabling controllable composition. Experiments demonstrate that our dataset and DSM design offer a principled approach to equipping general-purpose MLLMs with reliable navigation capabilities. Overall, we present a scalable, step-by-step paradigm that helps models acquire foundational skills for VLN applications.

REFERENCES

[1] 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.

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

[3] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. $\hat{a}2$ nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *arXiv preprint arXiv:2308.07997*, 2023.

[4] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. In *Proceedings of Robotics: Science and Systems*, 2025.

[5] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Proceedings of European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[6] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

[7] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li E Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37: 100428–100534, 2024.

[8] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of Empirical Methods in Natural Language Processing*, 2024.

[9] Tianyi Ma, Yue Zhang, Zehao Wang, and Parisa Kordjamshidi. Breaking down and building up: Mixture of skill-based vision-and-language navigation agents. *arXiv preprint arXiv:2508.07642*, 2025.

[10] Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*, 2025.

[11] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[12] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12078–12088, 2025.

[13] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[14] Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

[15] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *International Conference on Machine Learning*, pages 70576–70631. PMLR, 2025.

[16] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025.

[17] X Zhai, B Mustafa, A Kolesnikov, and L Beyer. Siglip: Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

[18] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based VLM plans the next step for vision-and-language navigation. In *Proceedings of Robotics: Science and Systems*, 2024.

[19] Lingfeng Zhang, Xiaoshuai Hao, Yingbo Tang, Haoxiang Fu, Xinyu Zheng, Pengwei Wang, Zhongyuan Wang, Wenbo Ding, and Shanghang Zhang. $nav\hat{a}3$: Understanding any instruction, navigating anywhere, finding anything. *arXiv preprint arXiv:2508.04598*, 2025.

[20] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024.

[21] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[22] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634, 2024.