DecoVLN: Decoupling Observation, Reasoning, and Correction for Vision-and-Language Navigation
-Supplementary Material-

Zihao Xin[1]*, Wentong Li[1*,†], Yixuan Jiang[1], Bin Wang[2], Runmin Cong[2], Jie Qin[1‡], Shengjun Huang[1‡]
[1]Nanjing University of Aeronautics and Astronautics [2]Shandong University

A. Additional Implementation Details

A.1. Dataset and Training Details

To train a navigation model capable of handling the complexities of the physical world, we construct a simulation-based training dataset incorporating diverse challenges. The dataset generation is organized into two stages: Supervised Fine-Tuning (SFT) and Error-Correction Fine-Tuning (ECF).

As in mainstream VLMs [3, 4, 6, 7, 10, 11, 13–16], we leverage the SFT stage to establish strong cross-modal alignment between navigation instructions and expert trajectories. We collect training data in the Habitat simulator from three datasets: R2R-CE [1], R2R-EnvDrop [9], and RxR-CE [2]. To bridge the gap between ideal simulator motion and the inherent uncertainty of real-world actuators, we employ a stochastic step strategy during data collection. In real-world scenarios, factors such as wheel slip and limited motor precision introduce small deviations at every movement step, causing gradual trajectory drift. To mimic this phenomenon, we introduce slight random offsets into motion actions during data collection, improving policy's robustness to real-world noise and drift. To further enhance visual diversity, we apply several image augmentation strategies such as Gaussian blur and random masking, with parameters listed in Tab. 1. Additionally, we employ a reverse augmentation strategy to expand the navigation paths. Specifically, we utilize the Qwen3 LLM [12] to rewrite each original instruction I into a reverse navigation instruction I_{rev} representing the same path in the opposite direction.

During the ECF stage, the SFT-trained model π_{SFT} performs autonomous rollouts in the simulator environment. When the model deviates from the expert trajectory, we introduce a *ShortestPathFollower* (SPF) as an expert policy to provide corrective actions that guide the agent back toward the optimal path. By collecting these high-quality state-action pairs, the model learns to recover effectively from errors. Notably, during this phase, we retain only visual-domain randomization and remove the action-level stochastic offsets to ensure the accuracy and stability of the expert correction signals.

Augmentation Strategy	Parameters
Random Brightness	factor = 0.2, prob = 0.5
Random Saturation	factor = 0.2, prob = 0.5
Posterization	bits = 4, prob = 0.5
Random Sharpness	factor = 0.2, prob = 0.5
AutoContrast	prob = 0.3
Gaussian Blur	prob = 0.2
Random Masking	prob = 0.2

Table 1. Image augmentation strategies and their corresponding parameters used in the SFT stage.

A.2. Navigation Instructions

To help the VLM to effectively understand its role within the POMDP [8] and differentiate between information modalities, we adopted a structured prompt format, explicitly categorizing the input information into three types: System Message, Task Instruction, and Visual Context. First, the System Message is used to define the agent's role, with the relevant prompt being,"You are a robot designed for navigation tasks," which activates reasoning capabilities related to embodied tasks. Next, the Task Instruction includes the dataset's natural language instruction I along with predefined action space, following standard VLN conventions: {*Move Forward 25cm, Turn Left 15 degrees, Turn Right 15 degrees, Stop*}. Finally, the Visual Context is divided into two parts: historical memory and current observation. The historical memory consists of the refined sequence of historical frames obtained by the adaptive memory refinement strategy, while the current observation is the current egocentric view. This explicit separation is designed to guide the model to prioritize anchoring its decisions on the cur-

*Equal contribution
†Project lead
‡Corresponding author

rent visual state, while treating the "historical memory" as a queryable memory bank to assist with long-horizon reasoning and resolve perceptual aliasing issues.

B. Real-World Deployment

In our real-world experiments, we deploy DecoVLN on a Unitree GO2 quadruped robot. The system adopts a device-server collaboration architecture. The robot is equipped with a Jetson Orin, which serves as the onboard computational core. It is responsible for running a lightweight Automatic Speech Recognition (ASR) model and the low-level motion control APIs. A remote server with a single NVIDIA RTX 4090 GPU runs the computationally intensive DecoVLN model. The system's asynchronous data loop operates as follows: the onboard ASR model converts the user's voice commands into text instructions I in real-time. These instructions and the current ego-centric video stream o_t, are transmitted (*uplinked*) to the server via a wireless video transmission protocol. Upon receiving this data, the server performs VLM inference and returns (*downlinked*) the resulting symbolic action chunk, containing four discrete navigation actions, which is then executed by the onboard controller.

C. Limitations and Future Work

Our current VLN framework is is built upon LLaVA-Video-7B [5], whose computational cost prevents fully onboard inference on devices such as the Jetson Orin. Current real-time operation relies on a device–server collaborative architecture, where heavy VLM inference is offloaded to a remote GPU server via wireless communication. This setup introduces additional network latency and limits deployment in environments with poor connectivity. Future work will explore model distillation and parameter-efficient tuning to transfer the high-level navigation capabilities of the 7B model to more lightweight models ranging from 0.5B to 2B parameters, enabling fully onboard real-time inference.

In challenging real-world scenes, the agent may still lose orientation when visual landmarks disappear or perceptual aliasing occurs. The current system lacks a deep introspective mechanism to address such scenarios of total deviation. To solve this problem, we plan to incorporate a Chain-of-Thought (CoT)–based error-recovery module that leverages historical memory to backtrack, identify a reliable waypoint, and replan a corrected trajectory, improving robustness in long-horizon navigation.

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*, 2018. 1

[2] 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. 1

[3] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025. 1

[4] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025. 1

[5] 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 the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 2

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[7] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1

[8] Matthijs TJ Spaan. Partially observable markov decision processes. In *Reinforcement learning: State-of-the-art*, pages 387–414. Springer, 2012. 1

[9] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019. 1

[10] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[11] Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2025. 1

[12] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1

[13] Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14147–14157, 2025. 1

[14] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024.

[15] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18970–18980, 2025.

[16] Yuqian Yuan, Wenqiao Zhang, Xin Li, Shihao Wang, Kehan Li, Wentong Li, Jun Xiao, Lei Zhang, and Beng Chin Ooi. Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity. *arXiv preprint arXiv:2510.23603*, 2025. 1