Progress-Think: Semantic Progress Reasoning for Vision-Language Navigation

Supplementary Material

1. More Implementation Details

1.1. Observation Processing

At each navigation step, the agent receives a monocular RGB observation as input. To incorporate temporal context efficiently, we construct a compact history by uniformly sampling 8 frames from all previous observations. Thus, the visual input to the model consists of the current frame together with 8 sampled historical frames. All images are single-view RGB with a resolution of 448×448 and a 90° field of view (FoV). No depth, panoramic view, odometry, or any privileged information is used during inference. This processing strategy maintains long-term context while keeping memory and computation overhead low, and we found it to be a stable and effective choice for VLN-CE.

1.2. PRM and PG-VLA Architecture Details

Both the Progress Reasoning Module (PRM) and the Progress-Guided VLA Module (PG-VLA) follow a standard vision–language model architecture. The visual input is encoded using SigLIP [2], where the final patch embeddings are projected into the language embedding space through a lightweight MLP projector. The projected visual tokens are then concatenated with the instruction tokens and processed by a Qwen-2 [3] language model to produce either progress reasoning outputs (in PRM) or action predictions (in PG-VLA). During the training of either module, all components, including SigLIP, the projector, and the Qwen-2 language model, are updated. The two modules do not share weights, preventing interference between progress estimation and policy learning while preserving a unified processing pipeline.

1.3. DAgger Data Collection

We adopt a DAgger [1] strategy to further enrich the training set with non-oracle trajectories. We first train a base navigation model using only the oracle demonstrations provided in the training dataset. The resulting model is then deployed in the R2R-CE training environments to perform inference without access to expert supervision. For trajectories that successfully reach the target, we record the full observation–action sequences and include them as additional training samples. This process enables the model to learn from its own successful behaviors and provides more diverse state–action pairs than the oracle demonstrations alone.

1.4. Prompts for Different Tasks

For the progress reasoning, the prompt is set to be:

> Assume you are a robot designed for navigation. Based on the visual observations <image>, ..., <image>, describe the instruction you have finished.

We use the following prompt to drive the PG-VLA to predict navigation actions:

> Imagine you are a robot programmed for navigation tasks. You have been given a video of historical observations: <image>,...,<image>and and current observation: <image>. Your assigned task is: [Instruction]. You have finished the task: [Progress]. Analyze information to decide your next move, which could involve turning left or right by a specific degree, moving forward a certain distance, or stop if the task is completed.

Among them, [Instruction] is the language instruction given for the current task and [Progress] is the progress predicted by PRM.

2. Real-world Deployment Details

To evaluate Progress-Think beyond simulation, we deploy it in real indoor environments using a Unitree G1 humanoid robot (Fig. 1). The model runs in real time on an NVIDIA RTX 4090 server and continuously receives the robot's egocentric RGB observations. At every step, the model first infers the progress state through progress reasoning and then predicts the next navigation action conditioned on the inferred progress. The resulting action is executed on the robot and the loop repeats autonomously. Across multiple scenes with different layouts, the robot consistently follows multi-step instructions and adapts its behavior according to the evolving progress, demonstrating that Progress-Think transfers effectively from simulation to the real world.

References

[1] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 1

[2] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil

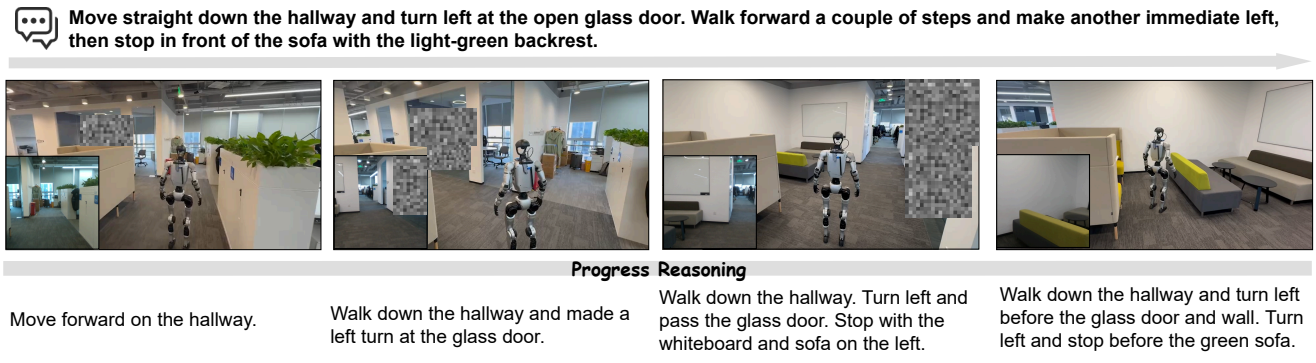

Figure 1. Real-world deployment of Progress-Think on a humanoid robot. We evaluate Progress-Think in real indoor environments, where the humanoid robot receives language instructions and must navigate based on egocentric observations. For each scene, we visualize the executed trajectory alongside the intermediate progress reasoning generated during navigation. Progress-Think produces step-level reasoning that remains aligned with the instruction and current scene context, enabling the robot to adjust its behavior throughout execution. The examples (a)–(c) show that the robot can reliably follow long-horizon instructions across different environments, successfully grounding progress in real-world visual observations.

Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1

[3] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1