CoT-VLNBench: A Benchmark for Visual Chain-of-Thought Reasoning in Vision-Language-Navigation Robots

Xiao Zhao*, Chang Liu*, Ruiteng Ji*, Zheyuan Zhang, Mingxu Zhu, Linna Song, Zhe Ren, Luo Qingliang, YuHang Gao, Zhaolong Du, Chufan Guo, Kuifeng Su

Tencent, AD Lab
{shawzhao,changeliu}@tencent.com

Abstract

Recent advances in vision language models (VLMs) have demonstrated remarkable potential in embodied navigation tasks. However, existing robot-centric datasets primarily focus on traditional 3D tasks such as perception and prediction, lacking adequate support for vision-language tasks. Vision-language-navigation (VLN) is a key capability for achieving human-like and interpretable navigation in complex environments. In this study, we present CoT-VLNBench, the first large-scale benchmark and dataset designed for chain-of-thought (CoT) reasoning in quadruped robot navigation. Our dataset encompasses a diverse range of indoor and outdoor scenes, multi-step navigation trajectories, and rich natural language instructions, all annotated with fine-grained CoT reasoning traces. Specifically, it contains 175K frames, 5.25M 3D bounding boxes, and 875K vision–question–answer (VQA) pairs. This comprehensive resource enables thorough evaluation of embodied agents' perceptual and step-by-step reasoning abilities. Furthermore, we propose a novel CoT-VLN model, a state-of-the-art 7B VLN model that integrates visual, linguistic, and reasoning modules, to facilitate interpretable and effective navigation. Extensive experiments demonstrate that our approach significantly outperforms existing non-VLMs baselines on the new benchmark, underscoring the importance of CoT-VLN in embodied navigation. We hope that CoT-VLNBench will serve as a valuable resource to advance research at the intersection of robotics, vision, language, and reasoning.

Introduction

Embodied navigation requires agents to perceive, reason, and act within complex real-world environments, posing a fundamental challenge in robotics and artificial intelligence. Recent advances in vision language models (VLMs) have demonstrated exceptional capabilities in integrating visual perception with natural language understanding (Liu et al. 2023, 2024a; Malla et al. 2023), thereby opening up new possibilities for embodied agents to follow human instructions and accomplish navigation tasks (Wang et al. 2024; Jiang et al. 2024). Nevertheless, the limitations of existing datasets and benchmarks continue to impede the development of robot robust and interpretable navigation systems.

Most current robot-centric datasets, such as CODa (Zhang et al. 2024) SIT, (Bae et al. 2023) and JRDB (Martin-Martin et al. 2021), primarily focus on 3D scene understanding, object detection, or tracking from the robot's perspective. While these datasets have facilitated significant progress in perception and mapping, they lack the annotations and task structures required to support vision-language navigation (VLN) and advanced reasoning. In particular, they do not provide step-by-step, interpretable reasoning traces, which are essential for understanding how agents make navigation decisions in response to complex or multi-step instructions.

To bridge this gap, we introduce CoT-VLNBench, a large-scale, multimodal benchmark and dataset specifically designed for Chain-of-Thought (CoT) (Wei et al. 2022) reasoning in vision-language navigation for quadruped robots in both indoor and outdoor environments. We collected data by the Unitree Go2 robot (Unitree Robotics 2024). CoT-VLNBench comprises over 175K frames (5 hours), 3.5M 3D bounding boxes, and 875K vision–question–answer (VQA) pairs, as shown in Figure 1. The dataset features rich natural language instructions, multi-step navigation trajectories, and detailed visual observations. Critically, for each navigation scenario, we provide fine-grained CoT reasoning traces that capture the intermediate reasoning steps connecting perception, language, and action. These annotations not only support traditional prediction tasks, but also provide explicit support for CoT-based VLN, enabling a comprehensive evaluation of embodied agents' capabilities.

Inspired by recent advances in CoT reasoning and autonomous driving research (Qian et al. 2024; Tian et al. 2024; Jiang et al. 2025), we adopt the CoT paradigm to bridge the gap between high-level language instructions and low-level navigation actions. To generate high-quality and comprehensive CoT annotations, the agent first produces intermediate subgoal representations (textual descriptions) that reflect its reasoning process prior to executing actions. This visual CoT reasoning enables the model to "think visually" about how to accomplish tasks, thereby enhancing both interpretability and performance. Specifically, we employ a step-by-step annotation pipeline powered by GPT-4.1, which includes detailed environment descriptions, agent motion analysis, 3D scene understanding, and navigation reasoning. The resulting CoT question–answer pairs are further refined through rule-based and manual annotation. By

*These authors contributed equally.

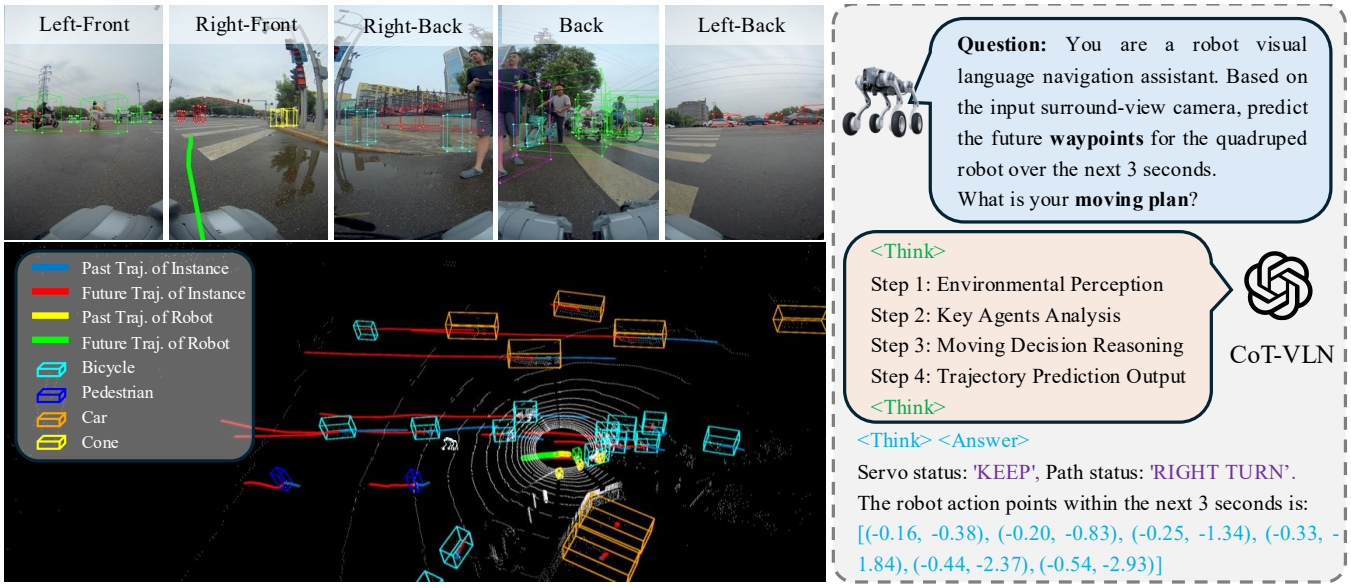


Figure 1: 3D visualization of environmental data collected by a robot moving in an outdoor scene. The first row shows five different camera views covering 360 degrees and the projection of 3D annotations. Below are the point cloud, robot trajectory, instance bounding boxes, and their respective trajectories. On the right is the workflow of our chain-of-thought visual-language navigation (CoT-VLN) method, which ultimately outputs the predicted waypoints for the next 3 seconds.

leveraging both human corrections and the advanced reasoning capabilities of GPT-4.1 to decompose annotations into distinct stages, we ensure the rationality and reliability of the CoT reasoning trajectories.

Furthermore, we propose a novel CoT-VLN model—a state-of-the-art VLN model based on Qwen2.5-VL 7B (Bai et al. 2025)—that integrates visual, linguistic, and reasoning modules. Our model leverages recent advances in unified multimodal foundation models and is trained on our new dataset. Extensive experiments demonstrate that CoT-VLN outperforms existing non-VLM baselines on the new benchmark, highlighting the importance of CoT reasoning for embodied navigation. We hope that CoT-VLNBench will serve as a valuable resource for advancing research at the intersection of robotics, vision, language, and reasoning.

Our main contributions are as follows:

- We propose CoT-VLNBench, the first large-scale dataset and benchmark for visual chain-of-thought reasoning in indoor and outdoor quadruped robot navigation.
- We provide fine-grained CoT annotations that align visual observations, language instructions, and intermediate reasoning steps, enabling interpretable multi-step navigation.
- We introduce a novel CoT-VLN model that integrates visual, language, and reasoning modules, and demonstrate its effectiveness on the new benchmark.

Related Work

Robot Navigation Datasets

Large-scale datasets have played a crucial role in advancing autonomous driving and robotic navigation. Datasets such as nuScenes, Waymo, KITTI, and Argoverse (Caesar et al. 2020; Sun et al. 2020; Geiger et al. 2013; Wilson et al. 2023) provide rich multimodal sensor data and detailed annotations for urban driving scenarios. However, these datasets are primarily designed for vehicles and may not fully capture the unique challenges faced by robots in human-centric environments. For robots, datasets like MIT Stata, NCLT, FusionPortable, and OpenLORIS (Fallon et al. 2013; Carlevaris-Bianco, Ushani, and Eustice 2016; Jiao et al. 2022; Shi et al. 2020) offer long-term simultaneous localization and mapping (SLAM) benchmarks with repeated traversals and multimodal data. SCAND (Karnan et al. 2022) and JRDB (Martin-Martin et al. 2021) further contribute with demonstrations of social navigation and pedestrian annotations, respectively. Notably, SIT (Bae et al. 2023) and CODa (Zhang et al. 2024) provides comprehensive object and terrain annotations in urban environments, supporting research on navigation for robots. Despite these efforts, there remains a lack of datasets that combine fine-grained semantic, terrain, and reasoning annotations specifically tailored for quadruped robot navigation in complex indoor and outdoor environments.

Vision-Language Navigation & Chain-of-Thought

VLMs have significantly advanced the field of robot learning by enabling agents to understand complex visual and linguistic inputs and take appropriate actions. Recent VLMs, such as Flamingo, BLIP, LLaVA, and QwenVL (Alayrac et al. 2022; Li et al. 2022, 2023; Bai et al. 2025), have demonstrated strong capabilities in visual perception, semantic understanding, and reasoning. These models have been integrated into various robotic systems to enhance

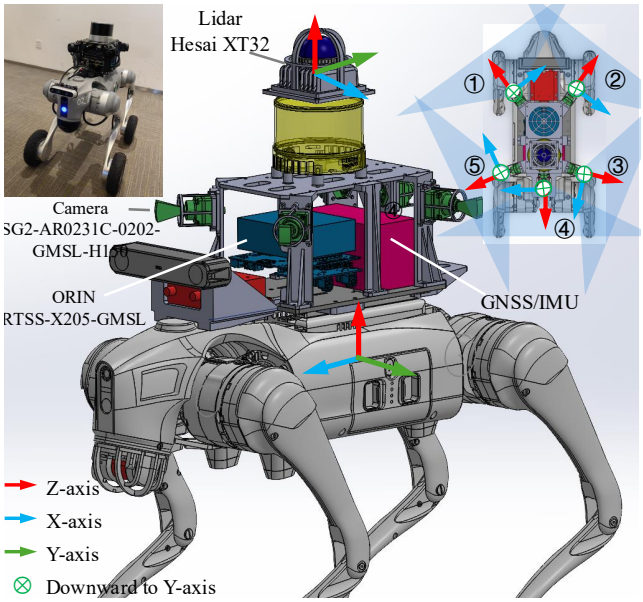


Figure 2: The Go2 platform.

perception, object detection, and instruction following, and have shown good generalization to novel environments and tasks (Kuo et al. 2022; Dorbala, Mullen, and Manocha 2023; Zhao et al. 2025). In the navigation domain, VLMs serve as powerful backbone models for interpreting natural language instructions grounded in visual observations, thereby facilitating more robust and flexible navigation strategies.

CoT reasoning, originally proposed in the natural language domain (Wei et al. 2022), has recently been extended to multimodal and embodied tasks (Mu et al. 2023). CoT reasoning enables models to decompose complex problems into sequential, interpretable steps, thereby improving both performance and transparency. In the context of autonomous driving, recent studies such as DriveLLM and AlphaDrive (Tian et al. 2024; Jiang et al. 2025; Chi et al. 2025) have explored the use of CoT reasoning to generate intermediate representations that guide action prediction and planning. These approaches demonstrate that incorporating step-by-step reasoning can enhance the navigation performance and interpretability of VLM-based models. However, most existing frameworks have not fully leveraged visual CoT reasoning in robot-centric scenarios. The introduction of our CoT-VLAbench addresses this gap.

CoT-VLN Benchmark

Robot Setup

We manually modified and operated a Unitree Go2 robot (Unitree Robotics 2024) to collect wheeled quadruped locomotion data across various environments. Wheeled quadruped robots offer greater flexibility and adaptability to diverse settings while maintaining stability. As shown in Figure 2, the robot is equipped with the following sensors:

- A Hesai XT32 rotating 3D Lidar, operating at 10 Hz.
- Five SG2-AR0231C-0202-GMSL-H150 cameras, capturing high-resolution images at 1920 × 1080 pixels with 150° field of view, covering the 360° view at 10 Hz. Cameras ① to ⑤ correspond to the left-front, right-front, right-back, back, and left-back positions, respectively.
- One IMU and GNSS module, providing fused pose data at 500 Hz. Note that the actual installation differs slightly from the design drawings.

The Go2 is deployed using a Docker image based on Ubuntu 20.04 and ROS2 Humble. Sensor timestamps are synchronized via the PTP protocol, and data from the cameras, Lidar, IMUs, and other sensors are recorded in the MCAP format. The coordinate system definitions for all sensors are shown in Figure 2. We also performed precise intrinsic and extrinsic calibration for all sensors, enabling the Go2 to fully leverage its multimodal sensing capabilities.

Data Collection

Our entire dataset comprises 60 scenes, including densely populated scenes such as intersections, sidewalks, and subway entrances, each sequences containing 300 seconds of continuous data sampled at 10 Hz, as shown in Figure 3(a). The dataset includes Lidar, five surround-view cameras, IMU, and GNSS positioning information, covering both indoor and outdoor environments for mobile robots. In total, it consists of 875K images and 175K Lidar frames. This encompasses approximately 5.25M 3D bounding boxes with object ID annotations and 8.75 million pairs of fine-grained language annotations. Tab. 1 highlights the significant features of the CoT-VLNBench dataset for navigation tasks in comparison to existing robot-centric datasets.

- The CoT-VLNBench dataset provides large-scale 3D scene data—an order of magnitude larger than previous robot-perspective datasets—and offers ultra-long 5-minute sequences to better support motion-related tasks. These data are collected as a wheeled quadruped robot navigates through densely populated indoor and outdoor environments, including building interiors, technology parks, subway entrances, and public sidewalks.
- For the first time, the CoT-VLNBench dataset offers fine-grained visison-question-answer (VQA) and supports chain-of-thought (CoT) reasoning. CoT-VLNBench aligns visual observations, language instructions, and intermediate reasoning steps, enabling interpretable multi-step navigation.

Annotation Lables

3D Bounding Box Labels. The 3D annotations in the CoT-VLNBench dataset consist of cuboids represented in Lidar coordinates, derived from images captured by 360-degree multi-view cameras and Lidar, along with object IDs. The perception annotation range is up to 30 meters. The GO2 robot primarily operates in pedestrian areas both indoors and outdoors. The annotated categories include Car, Bus, Bicycle, Pedestrian, Half-body (legs only, as GO2 can only capture the lower half of a person at close range), and Cone (encompassing cones, stone piers, bollards, and other static obstacles). Unlike SIT (Bae et al. 2023), we directly annotate

Dataset	Published	3D Ann frames	Sequential	Pose	3D Bbox	VQA Pairs	CoT
OpenLORIS (Shi et al. 2020)	ICRA	×	102s/40hz	SLAM	×	×	×
Rellis-3D (Jiang et al. 2021)	ICRA	13K/0.3hr	60s/10hz	GPS+SLAM	×	×	×
JRDB (Martin-Martin et al. 2021)	TPAMI	60K/2hr	112s/7.5hz	×	1.8M	×	×
SIT (Bae et al. 2023)	NeurIPS	12K/0.33hr	20s/10hz	RTK+SLAM	320K	×	×
CODa (Zhang et al. 2024)	TRO	32K/1hr	145s/10hz	GPS+SLAM	1.1M	×	×
CoT-VLNBench(Ours)	-	**175K/5hr**	**300s**/10hz	GPS+SLAM	**5.25M**	**875K**	✓

Table 1: Comparison between CoT-VLNBench (ours) and similar robotic datasets. The most important entry in each column is highlighted in bold. CoT-VLNBench provides the largest number of 3D bounding box annotations and annotated 3D frames. More importantly, only our dataset offers visual-language annotations.

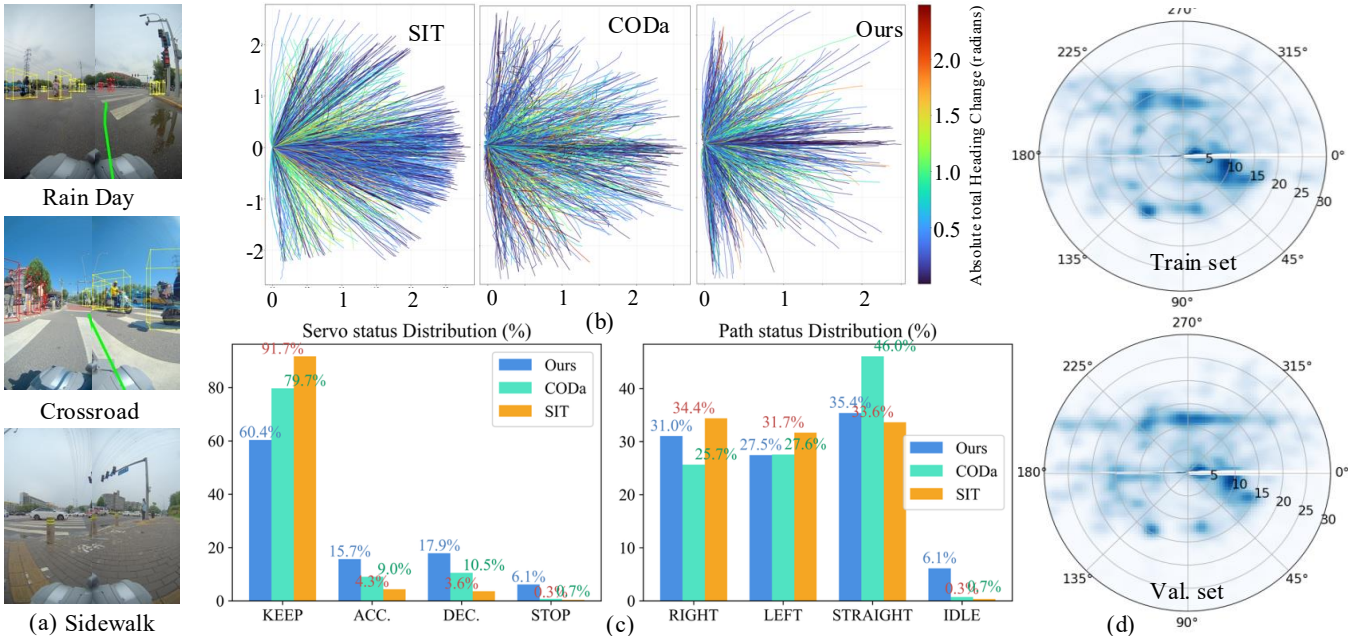


Figure 3: Feature comparison among different datasets. (a) Different scenarios. (b) Comparison of trajectory distributions. (c) Comparison of the proportion of servo states and path states, which, together with (b), demonstrates the diversity and balanced distribution of trajectories in our dataset. (d) Distribution of objects in the training and val. sets.

at a high standard of 10 Hz, rather than interpolating from 5 Hz to 10 Hz. This fine granularity is crucial for meeting the requirements of robotic navigation.

Robot Localization. Trajectory planning tasks require accurate determination of the robot's pose. To accommodate different localization conditions indoors and outdoors, we fuse GNSS and IMU data, integrate SLAM algorithms for relocalization. This enables real-time local and global localization of the robot in both indoor and outdoor environments. The localization system outputs the ego robot's WGS84 coordinates and three orientation angles: roll, pitch, and yaw. By leveraging these localization techniques, seamless transformation between the ego-centric and world coordinate systems is achieved, providing precise absolute target position information, which is essential for navigation-centric robotic systems.

For the predicted trajectory points over the next 3 seconds, we define two meta states, Servo status S_s and Path status P_s,to further represent the ego robot's state. Servo status indicates the status of the servo motors, analogous to the throttle state in autonomous vehicles. Path status represents the direction of the future trajectory.

$$S_s = [\text{Keep, Accelerate, Decelerate, Stop}] \quad (1)$$

$$P_s = [\text{Right Turn, Left Turn, Straight, IDLE}] \quad (2)$$

Figure 3 illustrates the overall distribution of meta-states in our dataset and compares it with the SIT (Bae et al. 2023) and CODa (Zhang et al. 2024) datasets. It can be observed that the SIT dataset still contains a large number of redundant trajectories after clustering. Combined with its short duration of only 0.33 hours, this indicates limited trajectory diversity compared to ours. The CODa dataset, on the other hand, mainly consists of straight trajectories, and the clustered trajectories are highly concentrated, suggesting a lack

of dispersion and similarly limited diversity. Overall, the distribution of various states in our dataset is more balanced. Figure 3(d) further shows that, from a robot-centric perspective, the distribution of targets in the training and validation sets is more reasonable.

VQA Pair Generation

Directly generating CoT data often faces challenges such as incoherent reasoning processes, inconsistent expressions, and difficulties in quantitative evaluation, which are particularly pronounced in high-degree-of-freedom task scenarios. To maximize the value of the dataset for training and evaluating planning models, we draw inspiration from frameworks such as DriveVLM (Tian et al. 2024) and EMMA (Hwang et al. 2024), and unify the multi-task annotations generated for each segment into planning-oriented VQA pairs. This format directly maps visual inputs, textual outputs, and action trajectory predictions into the sequence space of large models, facilitating end-to-end learning of the perception-reasoning-decision process.

Specifically, as illustrated in Figure 4, we systematically decompose the reasoning chain into four hierarchical steps:

Step 1: Environmental Perception. Focuses on understanding and describing the overall scene context.

Step 2: Key Agents Analysis. Provides detailed analysis of static road features and movable objects in the scene, explicitly introducing depth information to enable comprehensive 3D spatial understanding.

Step 3: Walking Decision Reasoning. In this critical step, we innovatively introduce a multiple-choice format, abstracting future trajectories and action decisions into a finite set of options to enhance the accuracy and consistency of the reasoning process.

Step 4: Trajectory Prediction Output. Outputs the final structured action trajectory prediction results.

In particular, for Step 3, considering the high degree of freedom in the movement of quadruped robots in open indoor and outdoor spaces, allowing the model to freely describe future trajectories can easily lead to inconsistent expressions and difficulties in quantitative evaluation. Therefore, we abstract future trajectories and actions into higher-level semantic options, namely Servo status and Path status. This multiple-choice format not only reduces the complexity of reasoning and annotation but also significantly improves the accuracy and comparability of CoT annotations.

At each hierarchical step, we employ the GPT-4.1 model to independently perform VQA reasoning, obtaining high-quality step-by-step reasoning ground truth, which is then manually reviewed and revised. Through this layered reasoning process, we are able to extract rich structured information beyond simple captions and reasonably assign the reasoning results to one of four unstructured scene categories. Note that, following Omnidrive (Wang et al. 2024), image inputs are concatenated: the left-front and right-front views are stitched together, as are the left-rear, rear, and right-rear views.

Subsequently, leveraging the ground-truth labels from the 3D scene data, human experts integrate multi-faceted feedback and domain knowledge to logically consolidate the outputs of each step. This process ultimately enables the automated generation of complete single-turn VQA pairs and step-by-step reasoning instructions with corresponding standard answers. By synthesizing insights obtained at each stage of the reasoning process, the constructed CoT is able to produce rational and consistent decision outcomes. The structured CoT outputs provide detailed contextual information and decision rationale for subsequent task annotation and model training.

COT-VLN Framework

As shown in Figure 4, we freeze the parameters of the image encoder and train only the projection layer and the language model. The COT-VLN model is optimized using supervised fine-tuning (SFT) (Longpre et al. 2023).

Benchmark Construction

To systematically evaluate the spatial reasoning and planning capabilities of VLMs in real-world scenarios, we constructed a large-scale visual question answering benchmark for wheeled quadruped robots. This benchmark comprises 140K training instances (80%) and 35K validation instances (20%), all sourced from our GO2 platform. The navigation task aims to predict waypoints for the next 3s (6 sets of trajectory points at 0.5s intervals). We also performed data distribution analysis to ensure that the training and the validation set is representative of all defined data distribution characteristics, as shown in Figure 3(d). We designed navigation-centric chain-of-thought reasoning tasks that require the model to analyze complex spatial relationships among multiple targets. The reasoning process includes general descriptions and fundamental spatial understanding, such as yaw angle classification, pixel-level localization, and depth range estimation, each assessing the model's performance across different target-centric reasoning dimensions.

Experiments

Experimental Settings

Training. We evaluated 5 current mainstream visual language models on CoT-VLNBench. These primarily include InternVL3-8B (Qwen2.5-7B) (Zhu et al. 2025), Qwen2.5-VL (Qwen2.5-7B) (Bai et al. 2025), Janus-Pro (DeepSeek-7B) (Chen et al. 2025), and LLaVA-1.6 (Vicuna-7B) (Liu et al. 2024b). We trained three variants of each model: pure language models, multimodal non-reasoning, and reasoning. Non-reasoning models generate answers directly based on the input without explicit intermediate reasoning. Reasoning models, on the other hand, adopt a step-by-step CoT process—triggered by the keyword "think" during training—to perform complex reasoning and produce more human-like answers. All models are trained on 8 NVIDIA L40 GPUs.

Evaluation. We conduct open-loop evaluation on the collected data to specifically assess the trajectory prediction accuracy of the quadruped robot in open scenarios when benefiting from CoT-VLN. Performance is mainly measured by the L2 distance (in meters) between the predicted and

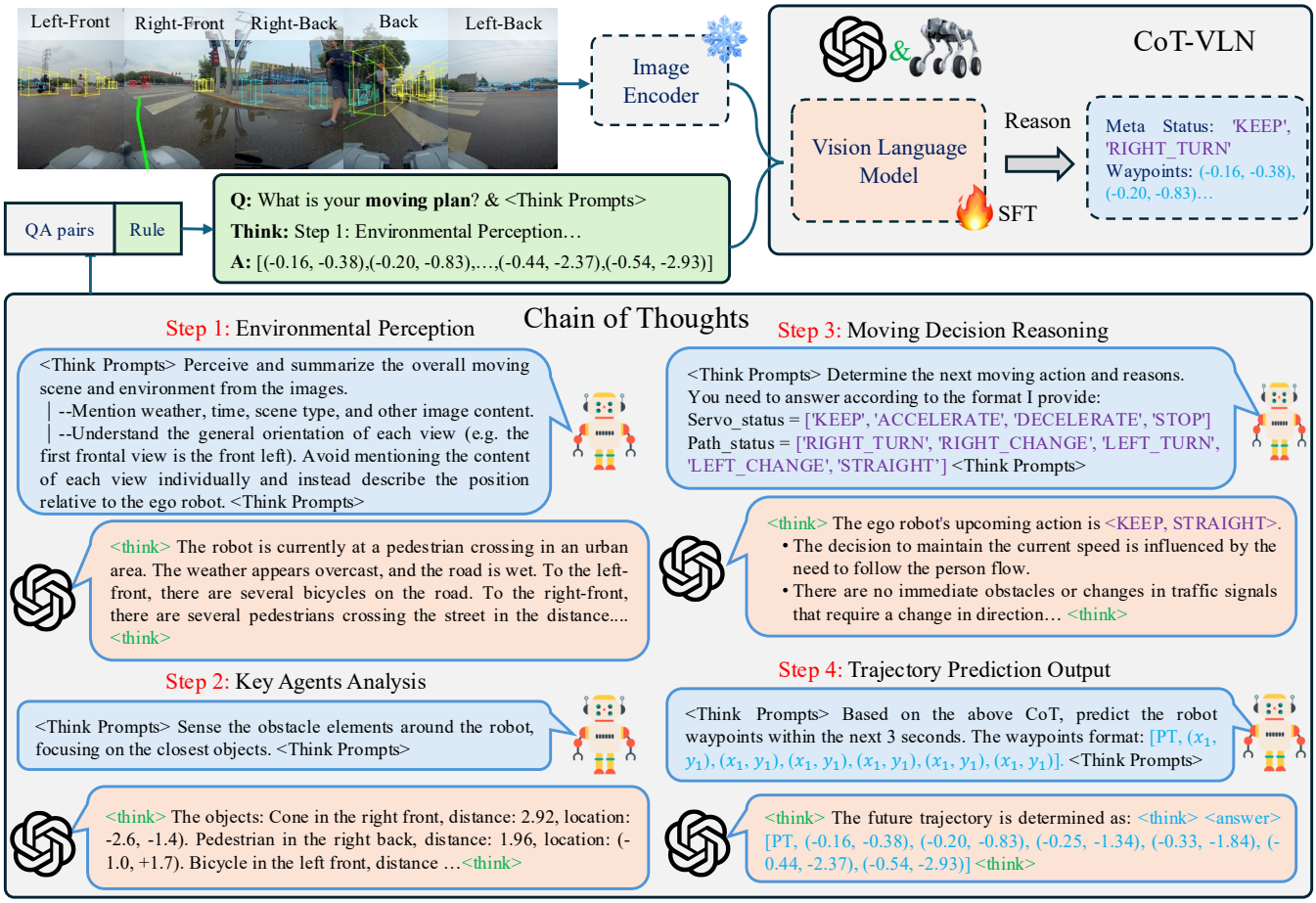


Figure 4: The chain-of-thought reasoning in CoT-VLNBench is structured into consecutive stages, including environmental perception, key agent analysis, moving decision-making, and final trajectory suggestion, to incrementally guide the VLM.

ground-truth trajectories within future time horizons of 1, 2, and 3 seconds, as well as the average L2 error.

$$\mathrm{L2}_t = \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2 \tag{3}$$

$$\mathrm{L2}_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2 \tag{4}$$

where $\hat{\mathbf{p}}_t$ is the predicted value at time t, and $\mathbf{p}_t$ is the ground truth at time t.

To evaluate the quality of the step-by-step CoT reasoning process, we adopt an industry-standard LLM-based evaluation protocol. We use the powerful GPT-4.1 as an automatic judge to compare the outputs generated by the fine-tuned VLMs with the ground truth VQA. The LLM scores each step on a scale of 1 to 100 based on factual accuracy, logical consistency, and semantic completeness.

Main Results

Tab. 2 presents the accuracy of each model under three evaluation settings, including both reasoning and non-reasoning structures in the VLN mode, as well as the non-reasoning structure in the LN mode. This comprehensive comparison aims to investigate how the acquisition of visual and linguistic information affects final decision-making in autonomous driving. Specifically, the LN mode refers to training on QA pairs pre-processed by GPT-4.1 without incorporating visual information, whereas the VLN mode includes visual inputs.

Overall, all models achieved their highest accuracy under the CoT reasoning setting, with Qwen2.5-VL (Bai et al. 2025) attaining the best performance, reaching an average error of 0.29m. For the VLMs listed in Tab. 2, direct inference without CoT reasoning led to suboptimal results; even a powerful model like Qwen2.5-VL experienced a performance drop of 13.7% (0.04m) in this setting. The lowest accuracy (0.39m) was observed when using language-only models without CoT reasoning. Our experimental results indicate that, when tackling complex spatial understanding tasks such as navigation, state-of-the-art (SOTA) VLMs require the joint guidance of both visual and linguistic information to make optimal decisions. On average, models without reasoning capabilities exhibited a 8.8% increase in average L2 error, while the absence of visual input led to a 13.5% decrease in accuracy. These findings were consistently observed across multiple VLMs. Figure 5 shows the CoT reasoning results of the best Qwen2.5-VL.

Models	L2 Error (m) ↓											
	Lang. Navi.				Vision Lang. Navi.							
	Non-Reasoning				Non-Reasoning				Reasoning			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
InternVL3-8B (Zhu et al. 2025)	0.19	0.33	0.61	**0.38**	0.18	0.31	0.54	0.34	0.15	0.27	0.49	0.31
Qwen2.5-VL (Bai et al. 2025)	0.20	0.34	0.62	0.39	0.16	0.30	0.52	**0.33**	0.14	0.28	0.47	**0.29**
Janus-Pro (Chen et al. 2025)	0.24	0.41	0.73	0.46	0.21	0.37	0.68	0.42	0.18	0.35	0.64	0.39
LLaVA-1.6 (Liu et al. 2024b)	0.22	0.39	0.71	0.44	0.20	0.36	0.63	0.40	0.16	0.33	0.57	0.35
Avg.	0.21	0.37	0.67	0.42	0.19	0.34	0.59	0.37	0.16	0.31	0.54	0.34

Table 2: Trajectory prediction L2 errors (m) on the CoT-VLNBench dataset with VLM-based method.

Model	Scores (%) ↑
InternVL3-8B (Zhu et al. 2025)	74.33
Qwen2.5-VL (Bai et al. 2025)	**75.67**
Janus-Pro (Chen et al. 2025)	69.56
LLaVA-1.6 (Liu et al. 2024b)	66.48

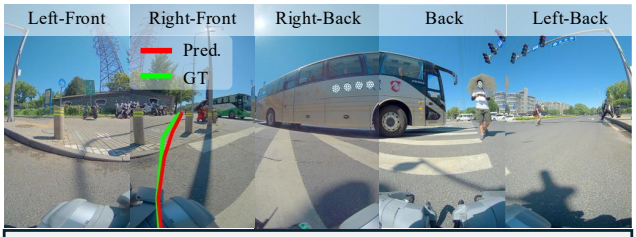
Table 3: Comprehensive evaluation with GPT4.1.

Model	L2 Error (m) ↓			
	1s	2s	3s	Avg.
SparseDrive (Sun et al. 2024)	0.20	0.41	0.65	0.42
DiffusionPlanner (Zheng et al. 2025)	0.17	0.32	0.61	0.37
CoT-VLN(Qwen2.5-VL)	0.14	0.28	0.47	**0.29**

Table 4: Trajectory prediction L2 errors (m) on the CoT-VLNBench dataset with Non-VLM method.

Furthermore, Tab. 3 reports the subjective evaluation of different VLMs' reasoning results by GPT-4.1, with Qwen2.5-VL achieving the highest score of 75.67. The subjective assessment results for all models are generally consistent with the objective metrics in Tab. 3, though some discrepancies exist. For example, LLaVA-1.6 (Liu et al. 2024b) received the lowest subjective score with reasoning, but outperformed Janus-Pro (Chen et al. 2025) in terms of L2 error.

In addition, Tab. 4 compares the VLMs with SOTA non-VLM trajectory prediction models in the autonomous driving domain, including the image-based end-to-end method SparseDrive (Sun et al. 2024) and the structured data-based DiffusionPlanner (Zheng et al. 2025), to analyze the advantages of VLMs over traditional approaches. To adapt these models to the COT-VLNBench dataset, we made appropriate modifications, such as removing the encoding of lane information in DiffusionPlanner while keeping other settings unchanged. The results demonstrate that our proposed COT-VLN model outperforms both non-VLM trajectory prediction models. Results indicate that, in open-world scenarios, path selection is more flexible compared to the relatively fixed routes in autonomous driving. VLMs are able to generate more accurate trajectories through effective reasoning.



Q: What is your **moving plan**? & <Think Prompts>
<Think> Step 1: The robot is situated at an intersection under clear, sunny weather conditions. To the left-front, there is a pedestrian crossing with several parked scooters. To the right-front, there are several cones obstacle ahead, while in the distance a large bus and several pedestrians crossing the street can be seen… Step 2: The Cone in the right front, Distance: 2.8, Location: [-1.3, -2.7]… Step 3: <Keep, Right Turn>. The robot is currently at an intersection, combined with the presence of a pedestrian crossing and cones, necessitates a controlled speed to safely navigate the turn while avoiding any obstacles. Step 4: Based on the above perception, object analysis, and action prediction, the future trajectory is determined as <Think><Answer> [PT, (-0.25, -0.78), (-0.55, -1.50), (-1.25, -2.00), (-1.65, -2.80), (-2.35, -3.43), (-4.10, -4.25)] <Think>

Figure 5: CoT-VLN(Qwen2.5-VL) reasoning result.

Conclusion

We introduce CoT-VLNBench, the first large-scale benchmark for Chain-of-Thought (CoT) reasoning in quadruped robot navigation, featuring diverse real-world scenes, multi-step trajectories, and rich natural language instructions with fine-grained reasoning annotations. Extensive experiments demonstrate that our proposed CoT-VLN model, which integrates visual, linguistic, and reasoning modules, achieves SOTA performance and outperforms both non-VLM and language-only baselines. The combination of visual and language information with explicit reasoning is crucial for robust and accurate navigation, while the absence of either modality or reasoning capability leads to notable performance drops. These findings underscore the importance of multimodal, interpretable reasoning in embodied navigation.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bae, J. W.; Kim, J.; Yun, J.; Kang, C.; Choi, J.; Kim, C.; Lee, J.; Choi, J.; and Choi, J. W. 2023. Sit dataset: socially interactive pedestrian trajectory dataset for social navigation robots. *Advances in neural information processing systems*, 36: 24552–24563.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Carlevaris-Bianco, N.; Ushani, A. K.; and Eustice, R. M. 2016. University of Michigan North Campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9): 1023–1035.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Chi, H.; Gao, H.-a.; Liu, Z.; Liu, J.; Liu, C.; Li, J.; Yang, K.; Yu, Y.; Wang, Z.; Li, W.; et al. 2025. Impromptu VLA: Open Weights and Open Data for Driving Vision-Language-Action Models. *arXiv preprint arXiv:2505.23757*.

Dorbala, V. S.; Mullen, J. F.; and Manocha, D. 2023. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5): 4083–4090.

Fallon, M.; Johannsson, H.; Kaess, M.; and Leonard, J. J. 2013. The mit stata center dataset. *The International Journal of Robotics Research*, 32(14): 1695–1699.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11): 1231–1237.

Hwang, J.-J.; Xu, R.; Lin, H.; Hung, W.-C.; Ji, J.; Choi, K.; Huang, D.; He, T.; Covington, P.; Sapp, B.; et al. 2024. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*.

Jiang, B.; Chen, S.; Liao, B.; Zhang, X.; Yin, W.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*.

Jiang, B.; Chen, S.; Zhang, Q.; Liu, W.; and Wang, X. 2025. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*.

Jiang, P.; Osteen, P.; Wigness, M.; and Saripalli, S. 2021. Rellis-3d dataset: Data, benchmarks and analysis. In *2021 IEEE international conference on robotics and automation (ICRA)*, 1110–1116. IEEE.

Jiao, J.; Wei, H.; Hu, T.; Hu, X.; Zhu, Y.; He, Z.; Wu, J.; Yu, J.; Xie, X.; Huang, H.; et al. 2022. Fusionportable: A multi-sensor campus-scene dataset for evaluation of localization and mapping accuracy on diverse platforms. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3851–3856. IEEE.

Karnan, H.; Nair, A.; Xiao, X.; Warnell, G.; Pirk, S.; Toshev, A.; Hart, J.; Biswas, J.; and Stone, P. 2022. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4): 11807–11814.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2022. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 22631–22648. PMLR.

Malla, S.; Choi, C.; Dwivedi, I.; Choi, J. H.; and Li, J. 2023. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1043–1052.

Martin-Martin, R.; Patel, M.; Rezatofighi, H.; Shenoi, A.; Gwak, J.; Frankel, E.; Sadeghian, A.; and Savarese, S. 2021. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 6748–6765.

Mu, Y.; Zhang, Q.; Hu, M.; Wang, W.; Ding, M.; Jin, J.; Wang, B.; Dai, J.; Qiao, Y.; and Luo, P. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36: 25081–25094.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. Nuscenes-qa: A multi-modal visual question answer-

ing benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4542–4550.

Shi, X.; Li, D.; Zhao, P.; Tian, Q.; Tian, Y.; Long, Q.; Zhu, C.; Song, J.; Qiao, F.; Song, L.; et al. 2020. Are we ready for service robots? the openloris-scene datasets for lifelong slam. In *2020 IEEE international conference on robotics and automation (ICRA)*, 3139–3145. IEEE.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; and Zheng, S. 2024. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Unitree Robotics. 2024. Unitree Go2. Accessed: 2024-06-10.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2024. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *CoRR*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; et al. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*.

Zhang, A.; Eranki, C.; Zhang, C.; Park, J.-H.; Hong, R.; Kalyani, P.; Kalyanaraman, L.; Gamare, A.; Bagad, A.; Esteva, M.; et al. 2024. Toward robust robot 3-d perception in urban environments: The ut campus object dataset. *IEEE Transactions on Robotics*, 40: 3322–3340.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1702–1713.

Zheng, Y.; Liang, R.; Zheng, K.; Zheng, J.; Mao, L.; Li, J.; Gu, W.; Ai, R.; Li, S. E.; Zhan, X.; et al. 2025. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.