Can MLLMs Find Their Way in a City? Exploring Emergent Navigation from Web-Scale Knowledge

Anonymous ACL submission

Abstract

Leveraging multimodal large language models (MLLMs) to construct embodied agents represents a promising approach for addressing real-world tasks. However, current benchmarks predominantly focus on language-centric tasks or heavily rely on simulated environments, thereby restricting their capacity to evaluate performance in realistic settings. To bridge this gap, we introduce CityNav, a comprehensive benchmark encompassing four diverse global cities, explicitly designed to evaluate the decision-making capabilities of raw MLLM-driven agents in real-world navigation tasks. Specifically, agents must rely solely on visual observations and internal multimodal reasoning to sequentially execute a significant number of decisions (50+) without additional environmental annotations or specialized architectural enhancements. Extensive evaluation reveals that popular linguistic reasoning techniques (e.g., Chain-of-Thought, Self-Consistency, Reflection) fail to deliver substantial improvements in performance. To address this, we propose *Verbalization of Path (VoP)*, a novel method that explicitly grounds the agent's internal multimodal reasoning through verbalized navigational paths, substantially enhancing navigation success. Nonetheless, overall performance in particularly challenging cities remains insufficient, underscoring the critical necessity for advanced reasoning frameworks and robust methods capable of handling complex, long-range sequential decision-making tasks.

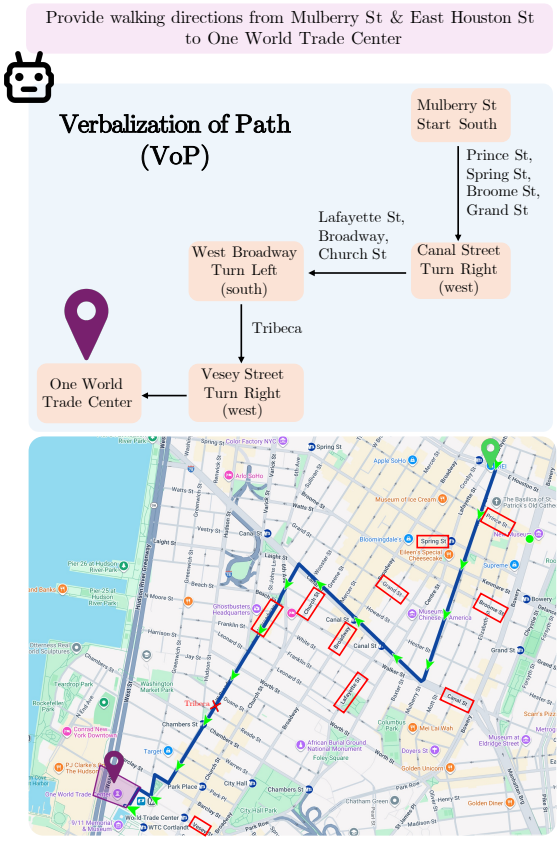

Figure 1: The figure illustrates our proposed Verbalization of Path (VoP) method. The red bounding boxes on the New York map highlight the street and location names referenced by the MLLM during its verbalized reasoning. Note, in New York City, street names typically refer to the entire street segment.

1 Introduction

Pretraining multimodal large language models on web-scale, interleaved image–text corpora induces broad, transferable world knowledge—supporting zero- and few-shot generalization across diverse vision–language tasks (Alayrac et al., 2022; Huang et al., 2023a; Radford et al., 2021). This emergent behavior has enabled foundation models (Brown et al., 2020; Achiam et al., 2023b; Touvron et al., 2023; Yang et al., 2024) to exhibit robust general reasoning and instruction-following abilities. Furthermore, their multimodal successors (OpenAI, 2024; Reid et al., 2024; Liu et al., 2024a; Wang et al., 2024; Chen et al., 2023), enhanced by recent scaling methods (Chen et al., 2025), integrate high-resolution visual processing, extended context handling, and precise OCR and grounding capabilities. Such advancements facilitate sophisticated perception-to-action pipelines, enabling re-

searchers to develop embodied agents capable of complex perception, planning, and action tasks (Huang et al., 2022, 2023b; Ahn et al., 2022; Song et al., 2023; Zhou et al., 2024; Lin et al., 2025; Yue et al., 2024).

While existing works have demonstrated reasoning capabilities of embodied agents in indoor navigation (Zhou et al., 2024; Lin et al., 2025; Yue et al., 2024), code generation and planning (Song et al., 2023; Liang et al., 2022), and robotic arm manipulation (Huang et al., 2023b; Ahn et al., 2022; Singh et al., 2022), these evaluations predominantly occur within simulated environments and are not knowledge intensive tasks. Therefore do not rigorously assess the agents' ability to leverage their extensive internal knowledge repositories to execute sequential decisions in dynamic, real-world scenarios.

Whereas, outdoor navigation is inherently a knowledge-intensive task demanding extensive cognitive capabilities, such as comprehensive environmental knowledge, sequential decision-making, spatial reasoning, and robust visual grounding using recognizable landmarks (Mirowski et al., 2018a). Prior studies on outdoor navigation have typically supplied agents with explicit landmark information embedded within images (Chen et al., 2019; Zeng et al., 2024; Schumann et al., 2024), significantly alleviating cognitive load by eliminating the necessity for agents to internally retrieve knowledge for self-localization and planning. In contrast, our work introduces a novel task termed *Sparsely Grounded Long-Range Navigation*, where MLLMs must navigate without any landmark annotations or explicit city navigation instructions, relying exclusively on images observed at each intersection. This task necessitates that MLLMs leverage their intrinsic world knowledge to facilitate spatial understanding, accurate self-positioning, and effective sequential decision-making.

Existing datasets (Mirowski et al., 2018a; Chen et al., 2019) exhibit several limitations: 1) they contain relatively short path lengths, 2) they are restricted to one-two cities, 3) many of them are closed source (not available even on request) 4) given their widespread usage over the years, there is a significant likelihood that MLLMs have been exposed to these datasets during training. To address these limitations, we introduce a novel, diverse dataset comprising paths of length greater than 2 Km, and include over 50+ decision points, spanning four distinct cities, thereby testing varied capabilities of MLLMs and significantly enhancing the task complexity. Our dataset is constructed using Google Street View panograph (Anguelov et al., 2010). Alongside the dataset, we provide a robust evaluation platform capable of deploying MLLMs directly onto the Google Street View navigation graph. The platform is explicitly designed to handle practical navigation challenges such as dead ends, missing street connections, and abrupt transitions inherent to Google Street View.

We further introduce *Verbalization of Path (VoP)*, a mechanism designed to explicitly extract and leverage the latent world knowledge internalized by MLLMs. By prompting agents to verbalize navigation paths—as illustrated in Fig. **??**—VoP substantially enhances the performance of MLLM-based agents on long-range navigation tasks. Since navigating unstructured environments is widely regarded as a fundamental hallmark of intelligence (Mirowski et al., 2018a), our results highlight the effectiveness of VoP in bridging the gap between static reasoning capabilities and dynamic, real-world sequential decision-making.

Our main contributions are 1) We introduce a new task and dataset designed to MLLMs on long-range sequential decision-making that requires leveraging their internal world knowledge. 2) We propose a zero-shot framework *Verbalization of Path (VoP)* to elicit and utilize the internal world knowledge of MLLMs for effective outdoor navigation. 3) We show that MLLMs can successfully navigate complex urban environments such as New York City, indicating that these models possess extensive structured world knowledge capable of supporting real-world spatial reasoning. 4) We demonstrate that structured reasoning techniques (such as CoT, reflection)- effective in static reasoning contexts—fail when MLLMs must coax and deploy their internal knowledge for dynamic, sequential decision-making.

2 Related Works

Embodied Environment. Existing evaluation suites for embodied agents differ along domain, action granularity, and multimodality. Household high-level instruction-following benchmarks (Shridhar et al., 2020b,a; Li et al., 2023; Szot et al., 2023; Choi et al., 2024; Li et al., 2024b) focus primarily on symbolic task decomposition and sequencing in indoor scenes, typically within one or two simulators. In low-level control, VLM-bench targets vision–language manipulation with-

out capability scoring, while VLABench adds capability-oriented, fine-grained evaluation with native LLM/VLM support (Zheng et al., 2022; Zhang et al., 2024). Finally, multi-domain evaluations at high level split into text-only, LLM-centric AgentBench and multimodal VisualAgentBench, which standardize agent scaffolds but remain at abstract action levels (Liu et al., 2023, 2024b). (Yang et al., 2025) unifies these strands with both high- and low-level action spaces and introducing a capability-oriented evaluation protocol. In contrast to simulator-bound or single-skill suites, our CityNav benchmark places *raw MLLMs* in *real, long-horizon city navigation* directly on the Google Street View graph—*without* landmarks, maps, or auxiliary annotations. Our evaluation platform also handles practical graph irregularities (dead ends, missing links).

MLLMs and Sequential Decision Making. Recent advancements in prompting (Wei et al., 2022; Yao et al., 2023a; Wang et al., 2022; Yao et al., 2024; Manas et al., 2024) have shown LLMs can exhibit sophisticated reasoning, and significantly improve performance on tasks requiring intermediate reasoning in static environment. (Yao et al., 2023b; Shinn et al., 2023; Gao et al., 2023) extend this further by enabling iterative planning through interactive feedback and reflection. Complementing these approaches, prompting-based planning methods (Wang et al., 2023; Xu et al., 2023) illustrate how strategic prompt refinement can enhance task-specific planning performance. In this work, we show that while these methods perform effectively in static contexts, they degrade significantly in sequential decision-making tasks that require methods to *coax* the internal world knowledge.

Instruction-based outdoor navigation. Vision-and-Language Navigation (VLN) (Gu et al., 2022) addresses the challenge of jointly grounding linguistic instructions and visual perception in realistic environments. Prior works (Chen et al., 2019; Schumann and Riezler, 2021) introduced landmark-rich navigation dataset. Subsequent work (Schumann and Riezler, 2022) demonstrated that agents trained on outdoor VLN tasks often fail to generalize to unseen urban regions, highlighting domain shift and data sparsity as major challenges. (Hermann et al., 2020) aligned textual instructions with visual observations in partially observable Street View environments, while (Mirowski et al., 2018b) employed reinforcement learning to improve navigation robustness. More recent methods (Schumann et al., 2024; Zeng et al., 2024; Xu et al., 2025; Tian et al., 2024) fine-tune MLLMs with city landmark–based instruction following, whereas (Li et al., 2024a) leverages driving videos to provide dense visual supervision for route following. All prior approaches rely on explicit textual or landmark-based instructions and operate over short trajectories (typically under 350 m). In contrast, our dataset CityNav focuses on long-range navigation (average path length $\approx$ 2 km) and provides no auxiliary environmental information—only images at each intersection—requiring the model to infer spatial relations, coax it's internal world knowledge and plan trajectories.

3 CityNav

We model autonomous city navigation as a Partially Observable Markov Decision Process (POMDP) defined on an undirected graph $G = (V, E)$, where V denotes intersections and E represents undirected street segments, each associated with a positive length $\ell(e) > 0$ for $e \in E$.

At any given discrete time step t, the state s_t of the agent corresponds directly to its current intersection: $s_t = v_t \in V$. From an intersection v_t, the set of available actions is defined as the set of street segments incident to v_t: $\mathcal{A}(v_t) = \{e \in E \mid v_t \in e\}$. When an agent at intersection v_t selects an action $a_t = e \in \mathcal{A}(v_t)$, it deterministically transitions to the adjacent intersection v' connected by the chosen street segment $e = \{v_t, v'\}$.

The system exhibits *sparse grounding*, as visual observations are only available at intersections and entire state in form of map is not available. Specifically, upon reaching an intersection v_t, the agent receives a set of images: $o_t = \{I_e \mid e \in \mathcal{A}(v_t)\}$, where each image I_e corresponds to visual input associated with the street segment represented by edge e. Between intersections, while navigating along a street segment, the agent receives no visual observations.

Action selection at each intersection is governed by a policy π, defined as a mapping from the current state v_t and observation set o_t to an action a_t: $\pi : V \times O \to E, \quad a_t = \pi(v_t, o_t)$, where O represents the space of possible observation sets. The policy thus dictates the agent's decision-making process, leveraging available visual information to select the next street segment to traverse.

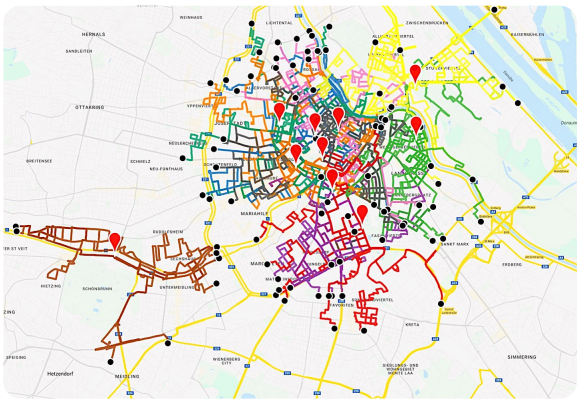

Figure 2: Dataset paths visualization of Vienna. Here the black dot's mark the starting point, and the red blobs mark the destination point.

3.1 Dataset

We curated a diverse dataset explicitly designed to evaluate multiple dimensions of navigation planning in LLM-based embodied agents. Our city selection methodology targeted locations exhibiting considerable diversity across primary language usage, architectural style, signage characteristics, and street layout topology, including variations in grid density and road complexity. This strategic diversity ensures exposure of the model to a comprehensive array of real-world navigational scenarios.

Specifically, we selected four globally distributed cities, each presenting distinct navigational challenges designed to rigorously assess the adaptability of the model. For example, Tokyo, Japan, predominantly employs Japanese-language signage and place nomenclature, thus posing significant linguistic barriers to LLMs primarily trained on English-dominated corpora. Within each city, we systematically identified and annotated 100 distinct origin-destination pairs, forming standardized evaluation tasks.

Manual Annotation. The destinations are not always just a building, sometimes they are as big as park. So in those case, the destination doesn't necessarily need to be just a node, it can be collection of nodes. Since, it's not possible to figure this out algorithmically we manually annotate the destinations for each of the chose places. So we draw a polygon around the destination that acts as a boundary for the destination and when agent reaches this boundary we call it reached.

Random Sampling of Starting Location. Starting from a seed node $v_s \in V$, our crawler aims to reach a target radial distance d_{target} from v_s. The traversal operates in two distinct phases: first, a deterministic corridor-following phase continues along nodes having an effective out-degree of 1 (excluding the backward link), until encountering the first decision junction (nodes with out-degree ≥ 2). The second phase involves a depth-first search (DFS) using an explicit junction stack with backtracking. At each junction node v_j, the crawler chooses among candidate edges $e_j^i \in E(v_j)$ according to a probability distribution computed via softmax over their angular deviation θ_i relative to the desired heading (typically directed away from the seed node). Specifically, the selection probability is: $P(e_j^i) = \frac{\exp(\cos(\theta_i)/T)}{\sum_k \exp(\cos(\theta_k)/T)}$ where the temperature parameter T anneals with increasing straight-line distance $d(v_j, v_s)$ from an initial random exploration ($T \to \infty$) toward a progressively more directional selection (lower T). To mitigate loops and encourage diverse coverage, we impose an exponential revisit penalty factor γ^{n_v} to the selection probability, where n_v is the visit count for node v, and $0 < \gamma < 1$. Once the crawler reaches the radial target distance d_{target}, it optionally continues for a small number of steps to terminate at a node with degree greater than or equal to a threshold $d_{\text{min_final}}$.

Google Street View Pre-processing. We enhance the underlying Google Street View graph for reliable navigation tasks, we systematically identify and resolve common structural issues. One of the primary challenges in constructing our navigation graph from Google Street View data arises from structural inconsistencies such as dead ends, incomplete coverage, and asymmetric links. We define a *dead end* as a node with an outgoing edge to a neighboring node that does not lead to any further intersections (i.e., a terminal street segment with no valid successors). To ensure graph connectivity and eliminate such artifacts, we algorithmically identify and prune dead ends during preprocessing. Additionally, the Street View panograph often exhibits asymmetric connectivity, where a link from node a to node b exists, but the reverse link from b to a is absent. This breaks the undirected graph assumption required for consistent navigation. To resolve this, we crawl the underlying graph and explicitly add the missing reverse edges whenever such inconsistencies are detected, thereby restoring bidirectional connectivity and ensuring that the resulting graph is well-formed for navigation tasks.

Cities	Region	Diversity	Distance (Km)	Decision Points
New York	USA	Grid-Based, Well Spaced, Rich Street Signs	1.8	44
São Paulo	Brazil	Non-Block Structure, Portuguese Language	2.0	55
Tokyo	Japan	Short Sightlines in Narrow Alleys, Japanese Language	1.9	80
Vienna	Austria	Road blocks becasue of rails, German Language	2.1	60

Table 1: Dataset statistics across four cities. *Diversity*: qualitative descriptors of urban form and visual/linguistic variety that affect navigation (e.g., grid regularity, sightlines, signage language). *Distance*: average path length (km) for routes in our test split. *Decision Points*: mean number of discrete navigation decisions per route (intersections).

3.2 Evaluation Metrics

We employ standard reasoning and navigation metrics:

• **Success:** If the agent reaches the destination node successfully, it receives a score of 1; otherwise, the score is 0.

• **SPL (Success weighted by Path Length):** SPL evaluates the agent's navigation efficiency by comparing the optimal (shortest possible) path distance d_{opt} to the actual distance traveled by the agent d_{agent}, scaled by the binary success indicator $S \in \{0, 1\}$. $\text{SPL} = S \times \frac{d_{\text{opt}}}{\max(d_{\text{agent}}, d_{\text{opt}})}$ where $S = 1$ if the agent successfully reaches the destination, and $S = 0$ otherwise.

• **Decision Accuracy (D.A.):** The percentage of correct navigation decisions (e.g., correct turns at junctions) made by the agent. A decision is classified as correct if the remaining walking distance to the destination (calculated using the Google Street View API) decreases after executing that decision.

4 AgentNav

4.1 Grounding with Verbalized Paths

Successful outdoor navigation fundamentally requires accurate self-localization and comprehensive world knowledge. Here, we demonstrate a targeted approach to probe such knowledge explicitly from MLLMs. We augment the agent's prompt-as-policy framework with three distinct phrases that consistently elicit robust navigation performance by explicitly grounding the agent's internal state and reasoning in the external world. Specifically, we incorporate the following structured prompts: 1) *Write the exact location of the destination*: This explicitly defines the navigation goal, anchoring the agent's decision-making process to a clear terminal state. 2) *Write the current estimated exact location*: This compels the agent to continuously estimate and update its current position, serving as a precise initial condition for subsequent decisions. 3) *Write the walking directions from the current position to the destination*: Crucially, this leverages the agent's generalist knowledge, prompting it to generate actionable instructions grounded in real-world spatial relationships and pathfinding logic.

4.2 Memory of AgentNav

For extensive runs averaging over 50+ decisions per trajectory, maintaining an efficient memory system is critically important. Traditional episodic memory architectures—which store comprehensive information (images, decisions, analyses) for each step across multiple past episodes—quickly become computationally intractable, scaling exponentially with episode length. To mitigate this issue, we strategically decompose memory management into three core components that significantly enhance efficiency (approximately a 100-fold reduction in memory overhead) within a Partially Observable Markov Decision Process (POMDP) framework: *Markovian Memory*, *Decision History*, and *Previous Visit Tracking*.

Markovian Memory. We implement *Markovian Memory* by explicitly prompting the agent to produce a memory state at each decision step. Formally, at time step t, the agent's input includes the previous memory state m_{t-1}, and along with next action a_t it outputs updated memory state m_t. This process can be expressed as: $(a_t, m_t) = \pi(v_t, o_t, m_{t-1})$ where m_t represents a sufficient statistic summarizing past observations, effectively transforming the partially observable process into a Markovian one within an augmented state space $\tilde{s}_t = (v_t, m_t)$. This mechanism eliminates the need for full episodic memory, thereby significantly reducing computational and storage costs. Furthermore, as the model becomes increasingly capable, it learns to selectively preserve only the most relevant information for effective decision-making, resulting in a more compact and adaptive memory representation.

Decision History. *Decision History* maintains a structured record of the sequence of actions chosen by the agent at each intersection along its trajectory. Formally, this can be represented

Figure 3: Illustration of state transition from S (marked by the yellow dot) to $S+1$ using the agent's internal reasoning through Verbalization of Path (VoP). At each state, the agent perceives visual cues and references its memory to update decisions and navigation strategy. The purple marker denotes the destination (One World Trade Center), while the green marker indicates the starting point.

		New York			Tokyo			Vienna			Sao Paulo		
MLLM	**Agent Config**	**Success**	**SPL**	**D.A.**	**Success**	**SPL**	**D.A.**	**Success**	**SPL**	**D.A.**	**Success**	**SPL**	**D.A.**
GPT 4o	AgentNav	88	0.539	72.91	14	0.099	40.90	26	0.170	46.32	20	0.06	43.45
	Base	13	0.064	39.04	4	0.046	36.79	4	0.031	35.67	3	0.040	34.69
GPT 5	AgentNav	94	0.711	82.98	30	0.163	54.97	56	0.226	54.82	29	0.126	48.96
	Base	54	0.375	55.97	10	0.088	41.24	11	0.092	40.72	7	0.051	36.98
GPT 4.1	AgentNav	92	0.557	75.27	17	0.101	43.67	32	0.182	49.95	22	0.080	44.05
	Base	15	0.097	42.27	5	0.044	38.83	2	0.037	34.66	5	0.049	35.46
Gemini 2.5 Flash	AgentNav	73	0.471	74.75	17	0.066	46.87	17	0.137	46.35	12	0.085	43.65
	Base	12	0.060	41.57	8	0.049	39.97	1	0.010	29.31	5	0.049	35.80
Qwen 2.5 VL 32b	AgentNav	32	0.153	56.39	12	0.094	40.03	12	0.119	44.94	9	0.059	37.80
	Base	7	0.089	35.11	2	0.023	30.01	0	0.0	26.1	2	0.011	29.87

Table 2: AgentNav and base MLLM performance on CityNav

as: $\mathcal{H}_t = \{a_1, a_2, \ldots, a_t\}$. Maintaining this ordered sequence enables the agent to leverage its own behavioral trajectory for reasoning about prior choices, route corrections, and void repeated loops. By focusing on compact action traces instead of complete episodic histories, this mechanism provides a balance between computational efficiency, and long-horizon reasoning.

Previous Visit. *Previous Visit* memory provides the agent with awareness of its past interactions at specific intersections. Each time the agent revisits an intersection v_t, it retrieves the record of its previous decisions taken at that node, enabling it to reason about prior outcomes. Repeatedly encountering the same intersection typically indicates that the agent is caught in a local loop or has failed to make progress toward the destination. To mitigate this, the system encodes a visit count n_v for each node v, which influences the policy's exploration behavior. As n_v increases, the agent is progressively discouraged from repeating the same action—promoting exploration and preventing cyclic behavior. For example, if the agent has chosen to go west multiple times from an intersection without improvement, the memory mechanism biases future actions toward unexplored directions such as east. This structured representation of visit history thus endows the agent with self-awareness of its traversal patterns, improving navigational robustness in complex city graphs.

5 Experimentation

In this section, we conduct comprehensive experiments to evaluate the performance of various MLLMs and methods on our proposed dataset CityNav, followed by ablation studies (Sec. 5.4) and

Method	New York			Tokyo			Vienna			Sao Paulo		
	Success(%)	SPL	D.A.(%)	Success(%)	SPL	D.A.(%)	Success(%)	SPL	D.A.(%)	Success(%)	SPL	D.A.(%)
GPT-4.1	15	0.097	42.27	5	0.044	38.83	2	0.037	34.66	5	0.049	35.46
CoT	21	0.173	44.59	9	0.077	41.09	4	0.039	34.88	7	0.055	37.93
Self Reflection (GPT-4.1)	16	0.112	42.90	4	0.040	36.20	3	0.042	36.33	12	0.052	41.95
Self Reflection (GPT-5)	22	0.168	48.14	8	0.079	41.48	5	0.045	37.84	13	0.050	41.64
AgentNav	92	0.557	75.27	17	0.101	43.67	32	0.182	49.95	22	0.080	44.05

Table 3: The table shows comparison results with different baselines. All the experiments here are performed using GPT-4.1. Self Reflection (GPT-4.1) means the agent used is GPT-4.1 and reflection is done with GPT-4.1. Self Reflection (GPT-5) agent used is GPT-4.1 and reflection is done with GPT-5.

error analysis (Sec. 5.5).

5.1 Implementation Details

For each run within every city, we set a maximum limit of 150 decision points for the agent to reach its destination before automatic termination. Additionally, we restrict the maximum number of graph node transitions to 2000. Rather than executing self-positioning at every decision point, we perform self-positioning every third decision point (a separate call to same MLLM). This deliberate choice introduces minor positional uncertainty, effectively testing the robustness and accuracy of the verbalized path by challenging the agent to reason with slightly imprecise localization.

5.2 MLLMs and Reasoning Baselines

We evaluate our method using a selection of strong closed-source and open-source multimodal foundation models to effectively probe their internal world knowledge and reasoning capabilities. Specifically, our evaluation includes: GPT-4o (Achiam et al., 2023a), a widely-used baseline model in multimodal research; GPT-4.1, which is expected to demonstrate enhanced geographical reasoning capabilities (Grainge et al., 2025); Gemini-2.5 Flash (Team et al., 2024), serving as an additional closed-source comparison; GPT-5 (thinking), known for its advanced reasoning abilities; and Qwen-2.5VL-32B (Bai et al., 2025), a powerful open-source counterpart. Our focus on these sophisticated multimodal models is motivated by our goal of *coaxing* out the latent world knowledge embedded within large-scale, web-trained models.

To systematically evaluate reasoning effectiveness, we employ standard reasoning baselines, including Chain-of-Thought (CoT) (Wei et al., 2022), Self-Reflection (GPT-4.1) (Shinn et al., 2023), and Self-Reflection (GPT-5). Here, the labels GPT-4.1 and GPT-5 indicate the specific model used to perform the reflective reasoning step, revisiting and refining the initial reasoning output.

5.3 Quantitative Results

Significant Performance Across 5 MLLMs. Table 2 demonstrates that VoP consistently delivers strong improvements across all five evaluated MLLMs. In high-density urban settings such as New York, our approach achieves significantly higher success rates and SPL scores compared to the base configurations of each model. This indicates that the Verbalization of Path (VoP) mechanism effectively *coaxes* latent world knowledge from MLLMs, enabling more reliable decision-making in complex navigation tasks. Importantly, similar performance gains are observed across all four cities, highlighting the generality of VoP.

Limitations of Existing Reasoning Techniques. Table 3 highlights the limitations of standard reasoning strategies. While these techniques have proven effective for static reasoning tasks, they fail to achieve competitive performance on our long-range, embodied navigation benchmark. Their inability to elicit sufficiently structured internal knowledge from MLLMs leads to lower success rates and SPL scores.

Challenges Arising from the Diversity of CityNav. Despite the substantial gains delivered by AgentNav, Table 2 also reveals the inherent difficulty of long-horizon sequential decision-making tasks. Even with VoP, absolute success rates and SPL scores remain modest for Tokyo, Vienna and Sao Paulo, reflecting the persistent gap in reasoning abilites of current MLLMs and reasoning methods (as we see in Table 3).

5.4 Ablations and Analysis

To systematically analyze the contributions of individual components of our method, we perform an ablation study summarized in Table 4. We start with the baseline GPT-4.1 model, which attains a success rate of 15%. We incrementally add components to evaluate their individual effects. Initially,

Methods	Success(%)	SPL	D.A.(%)
GPT-4.1	15	0.097	42.27
GPT-4.1 + Markovian Memory	23	0.162	47.19
GPT-4.1 + Decision History	29	0.228	55.63
GPT-4.1 + Previous Visit	35	0.298	56.67
GPT-4.1 + Partially Verbalized Path	66	0.469	63.48
AgentNav	**92**	**0.557**	**75.2755**

Table 4: Performance breakdown showing contributions of different components of the AgentNav.

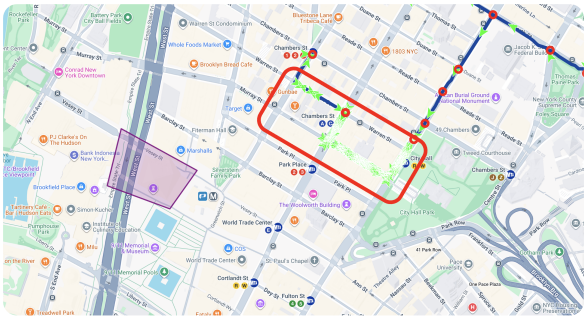

Figure 4: Failure case, the red box highlights the path in which AgentNav gets suck in a loop.

we introduce different memory components: incorporating *Markovian Memory* raises the success rate to 23%. The subsequent addition of *Decision History* further improves these metrics, achieving 29% success. Integrating *Previous Visit Tracking* leads to additional improvement, elevating performance to 35%.

Next, we further dissect the Verbalization of Path (VoP) component by breaking it down into two stages to clearly identify its effect. In the *Partially Verbalized Path* scenario, we instruct the agent to explicitly ground its reasoning solely based on the final destination, prompting it to write the destination's address at the beginning of its reasoning. This targeted grounding notably enhances the success rate to 66%. Subsequently, employing the complete verbalization—where the agent explicitly generates detailed walking directions from its current location to the target destination—our full *AgentNav* model reaches the highest performance with a 92% success rate. This stepwise analysis underscores how each introduced stage of verbalization progressively grounds and coaxes richer world knowledge from GPT-4.1, significantly improving the agent's capability to reason and navigate reliably in complex, real-world scenarios.

5.5 Error Analysis & Failure Cases

We find three key failure modes:

Unexpected Street View Linkage. One notable failure mode arises from inaccuracies inherent to the Street View panograph, specifically unexpected or misleading panorama linkages. Occasionally, a panorama indicated to face eastward (90°) may, in reality, curve unexpectedly northward (45°), causing inconsistencies between the visual cues provided and the actual navigation outcome. These discrepancies yield unreliable visual input to the MLLM, impairing its directional reasoning. For instance, an agent may repeatedly attempt to proceed southward toward its intended destination, yet find itself redirected northward.

Grounding Challenges in Complex Linguistic Environments. Another occasional failure mode occurs in cities featuring complex linguistic environments, such as Tokyo. Here, visual signage predominantly comprises intricate Kanji characters, which, combined with the limited resolution of input images, occasionally hinders effective visual grounding and accurate self-positioning. As a result, the self-positioning of the agent sometimes fails to correctly interpret critical spatial cues, causing inaccurate localization.

Over-Reliance on Initial Plans. A further failure mode emerges when the agent occasionally demonstrates excessive adherence to its initial navigation plan. In such cases, the agent sometimes prioritizes its original intended route without adequately reconsidering or revising it in response to new observations or emerging evidence. This occasional rigidity may result in biased or suboptimal decisions despite the availability of more effective alternative paths.

6 Conclusion

In this work, we introduced CityNav, a comprehensive benchmark designed to rigorously assess MLLMs on real-world, long-range urban navigation tasks. Through extensive experiments across diverse global cities, we demonstrated that our proposed Verbalization of Path mechanism, complemented by strategic memory components, effectively coaxes the intrinsic world knowledge from MLLMs, resulting in significant performance improvements over existing reasoning methods. These findings underline current limitations of MLLMs for embodied sequential decision-making tasks, emphasizing the need for continued research in robust, adaptive reasoning frameworks.

7 Limitations

While CityNav and our proposed AgentNav framework represent a significant step forward in evaluating MLLMs on real-world long-range navigation tasks, several limitations remain. First, despite explicit verbalization and memory mechanisms, absolute success rates and SPL scores remain modest for certain models and cities, highlighting the persistent gap between current MLLM capabilities and the demands of large-scale, embodied reasoning tasks.

Second, our evaluation relies on Google Street View imagery and its underlying graph structure which occasionally contain structural inconsistencies such as missing reverse links or misleading panorama orientations. Although our preprocessing mitigates many of these issues, residual anomalies may still introduce unintended noise into the agent's input and affect its performance.

Third, our benchmark emphasizes four diverse cities but is not yet exhaustive of all possible urban environments, languages, or signage complexities. Models trained primarily on English-dominated corpora may still exhibit limited generalization to regions with highly complex scripts or low-visibility environments.

Finally, our approach focuses on zero-shot reasoning and prompt-based policies without fine-tuning the models on navigation-specific data. While this design isolates intrinsic world knowledge, it may underestimate the full potential of models that could be adapted or specialized for long-range, embodied navigation tasks. We view these limitations as promising directions for future work aimed at developing more robust, adaptive, and context-aware reasoning frameworks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joshua Achiam, Sarina Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, F. L. Aleman, Daniel Almeida, Jan Altenschmidt, Sam Altman, Shantanu Anadkat, et al. 2023b. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. 2010. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Zhe Chen, Weijian Wang, Yue Cao, Yufei Liu, Zizheng Gao, Enming Cui, Junyang Zhu, Shilong Ye, Hao Tian, Zheng Liu, Lianlong Gu, Xinlong Wang, Qirong Li, Yi Ren, Zhiyu Chen, Jian Luo, Jinghao Wang, Tianhe Jiang, Bin Wang, Changxin He, Bo Shi, Xiaoqing Zhang, Hanyu Lv, Yu Wang, Wenqi Shao, Peng Chu, Zhenguo Tu, Tao He, Zheng Wu, Haohua Deng, Jinrui Ge, Kehan Chen, Kai Zhang, Liang Wang, Mingsong Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Jiapeng Wu, Weijian Wang, Wensheng Su, Gongwei Chen, Sheng Xing, Mulin Zhong, Qihang Zhang, Xizhou Zhu, Liang Lu, Bo Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. 2024. Lota-bench: Benchmarking language-oriented task planners for embodied agents. *arXiv preprint arXiv:2402.08178*. ICLR 2024.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Oliver Grainge, Sania Waheed, Jack Stilgoe, Michael Milford, and Shoaib Ehsan. 2025. Assessing the geolocation capabilities, limitations and societal risks of generative vision-language models. *arXiv preprint arXiv:2508.19967*.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11773–11781.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023a. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023b. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. 2023. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205 of *Proceedings of Machine Learning Research*, pages 80–93. PMLR.

J. Li, A. Padmakumar, G. Sukhatme, and M. Bansal. 2024a. VLN-Video: Utilizing driving videos for outdoor vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18517–18526.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. 2024b. Embodied agent interface: Benchmarking llms for embodied decision making. *arXiv preprint arXiv:2410.07166*. NeurIPS 2024 Datasets & Benchmarks.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2025. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Haotian Liu, Chenyang Li, Yuhang Li, Bo Li, Yutong Zhang, Shiyu Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*. ICLR 2024.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, Jiadai Sun, Xinyue Yang, Yu Yang, Zehan Qi, Shuntian Yao, Xueqiao Sun, Siyi Cheng, Qinkai Zheng, Hao Yu, Hanchen Zhang, Wenyi Hong, Ming Ding, Lihang Pan, Xiaotao Gu, Aohan Zeng, Zhengxiao Du, Chan Hee Song, Yu Su, Yuxiao Dong, and Jie Tang. 2024b. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*.

Kumar Manas, Stefan Zwicklbauer, and Adrian Paschke. 2024. Cot-tl: Low-resource temporal knowledge representation of planning instructions using chain-of-thought reasoning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9636–9643. IEEE.

P. Mirowski, M. Grimes, M. Malinowski, K. M. Hermann, K. Anderson, D. Teplyashin, K. Simonyan, A. Zisserman, R. Hadsell, et al. 2018a. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*.

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. 2018b. Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Machel Reid, Nikolay Savinov, Dmitry Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

R. Schumann and S. Riezler. 2021. Generating landmark navigation instructions from maps as a graph-to-text problem. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 489–502.

R. Schumann and S. Riezler. 2022. Analyzing generalization of vision and language navigation to unseen outdoor areas. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7519–7532.

R. Schumann, W. Zhu, W. Feng, T.-J. Fu, S. Riezler, and W. Y. Wang. 2024. VELMA: Verbalization embodiment of LLM agents for vision and language navigation in Street View. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18924–18933.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020a. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10740–10749.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020b. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*. ICLR 2021.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2022. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R. Devon Hjelm, and Alexander Toshev. 2023. Large language models as generalizable policies for embodied tasks. *arXiv preprint arXiv:2310.17722*. ICLR 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

H. Tian, J. Meng, W. S. Zheng, Y. M. Li, J. Yan, and Y. Zhang. 2024. Loc4plan: Locating before planning for outdoor vision and language navigation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4073–4081.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Sharan Batra, Prajjwal Bhargava, Gaurav Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Peiyi Wang, Shusheng Bai, Shuo Tan, Shiji Wang, Zhe Fan, Jun Bai, Kexin Chen, Xiang Liu, Jinghao Wang, Wenhai Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Weijia Xu, Andrzej Banburski-Fahey, and Nebojsa Jojic. 2023. Reprompting: Automated chain-of-thought prompt inference through gibbs sampling. *arXiv preprint arXiv:2305.09993*.

Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. 2025. Flame: Learning to navigate with multi-modal llm in urban environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9005–9013.

Ao Yang, Botao Yang, Bo Zhang, Bo Hui, Bing Zheng, Bowen Yu, Chaoqun Li, Dong Liu, Fei Huang, Hong Wei, et al. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. 2025. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yao Yao, Zuchao Li, and Hai Zhao. 2024. Got: Effective graph-of-thought reasoning in language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2901–2921.

Junpeng Yue, Xinrun Xu, Börje F Karlsson, and Zongqing Lu. 2024. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. *arXiv preprint arXiv:2410.03450*.

Qingbin Zeng, Qinglong Yang, Shunan Dong, Heming Du, Liang Zheng, Fengli Xu, and Yong Li. 2024. Perceive, reflect, and plan: Designing llm agent for goal-directed city navigation without instructions. *arXiv preprint arXiv:2408.04168*.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*.

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Eric Wang. 2022. Vlmbench: A compositional benchmark for vision-and-language manipulation. *arXiv preprint arXiv:2206.08522*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

Appendix

A Google Street View Policy Compliance Statement

Our study uses Google Street View imagery only at evaluation time and only via official Google-provided interfaces. The evaluation platform executes the agent directly on the Street View navigation graph and fetches the panorama views transiently for decision making; we do not scrape, bulk-download, mirror, or redistribute Street View imagery. The public release accompanying this paper consists of annotations (origin–destination pairs, destination polygons), code, and evaluation scripts. No images will be released. Human annotation was performed by the authors and is limited to drawing destination boundaries; no personally identifying information was collected or added. As documented in the paper (dataset description in Section 3.1; summary statistics in Table 1), imagery appears only as on-the-fly observations at intersections, multilingual scene text may be visible within those images, and all faces and license plates remain blurred as provided by Google, since we do not alter or post-process Street View content.

To align fully with Google's Street View and Maps Platform Terms of Service, we (i) access imagery solely through authorized APIs/viewers; (ii) preserve all provider attribution, blurring, and watermarks returned by the service; (iii) avoid storing or caching raw imagery beyond ephemeral runtime needs; (iv) release no derivative image dataset (only text/graph metadata and author-created annotations under our license); (v) require downstream users to supply their own API keys and to accept and comply with Google's Terms when reproducing our results; and (vi) forbid any use of our code or annotations to scrape, de-watermark, reverse-engineer, or otherwise circumvent Google's technical and policy safeguards. Our figures, where illustrative thumbnails are necessary for scholarly reporting, are minimal and strictly for explanation of results; they do not constitute redistributable imagery or a dataset. These measures, together with our annotations-only release and authors-only human labeling protocol, ensure that the work adheres to Google's Street View policy while enabling reproducible research on long-range, real-world navigation.

B Examples

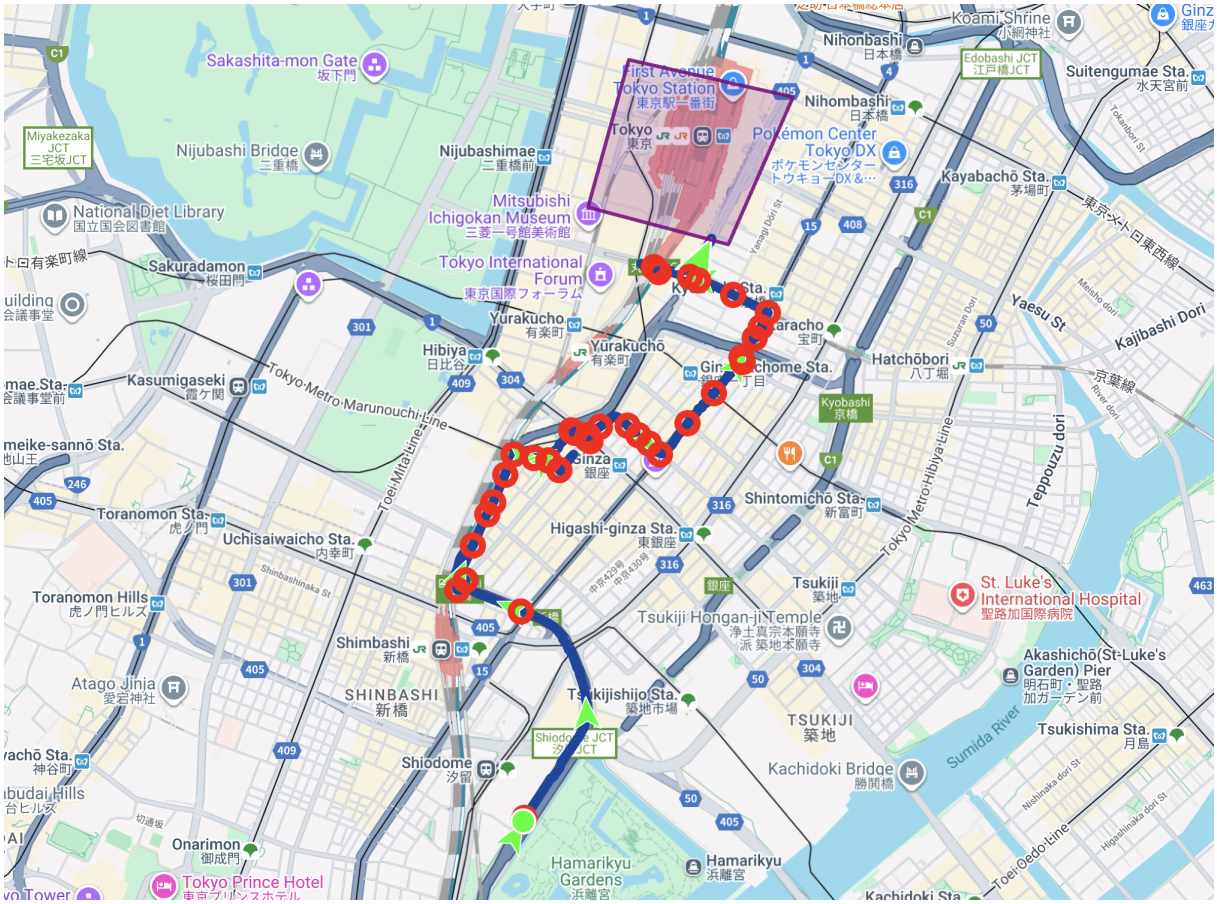

Figure 5: Navigation to Tokyo Station

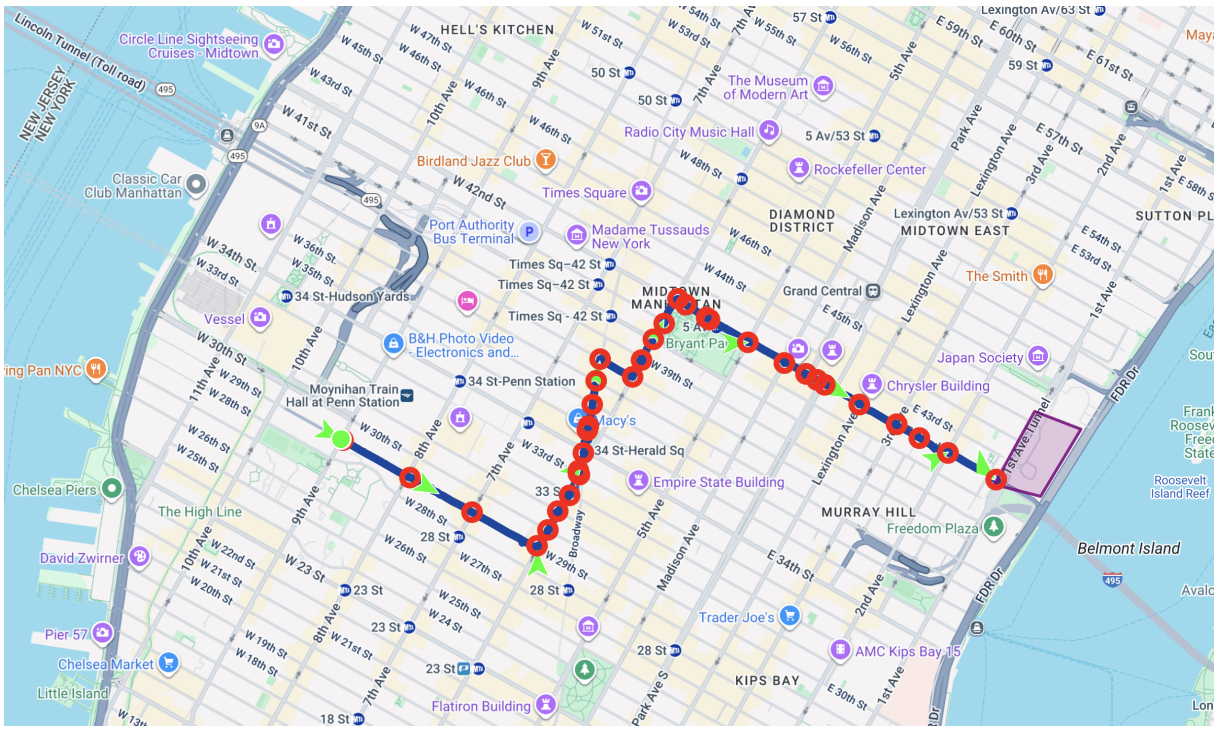

Figure 6: Navigation to United Nations Headquarters (New York)

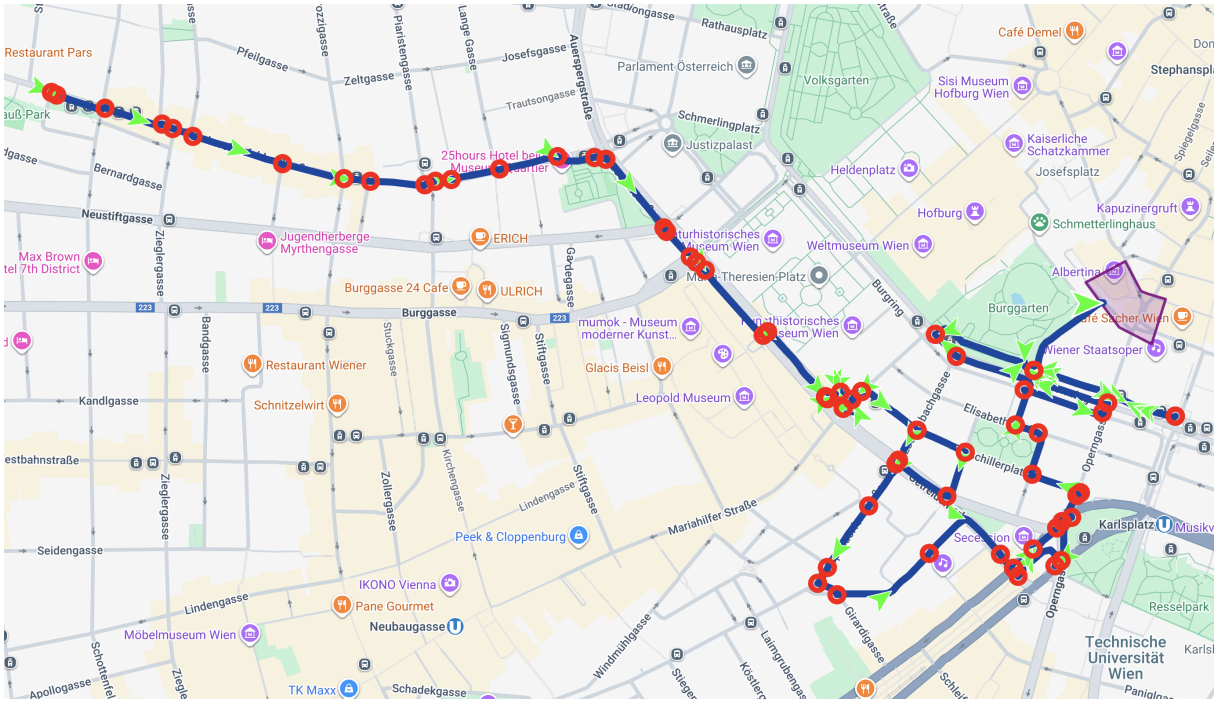

Figure 7: Navigation to Albertina Museum (Vienna)

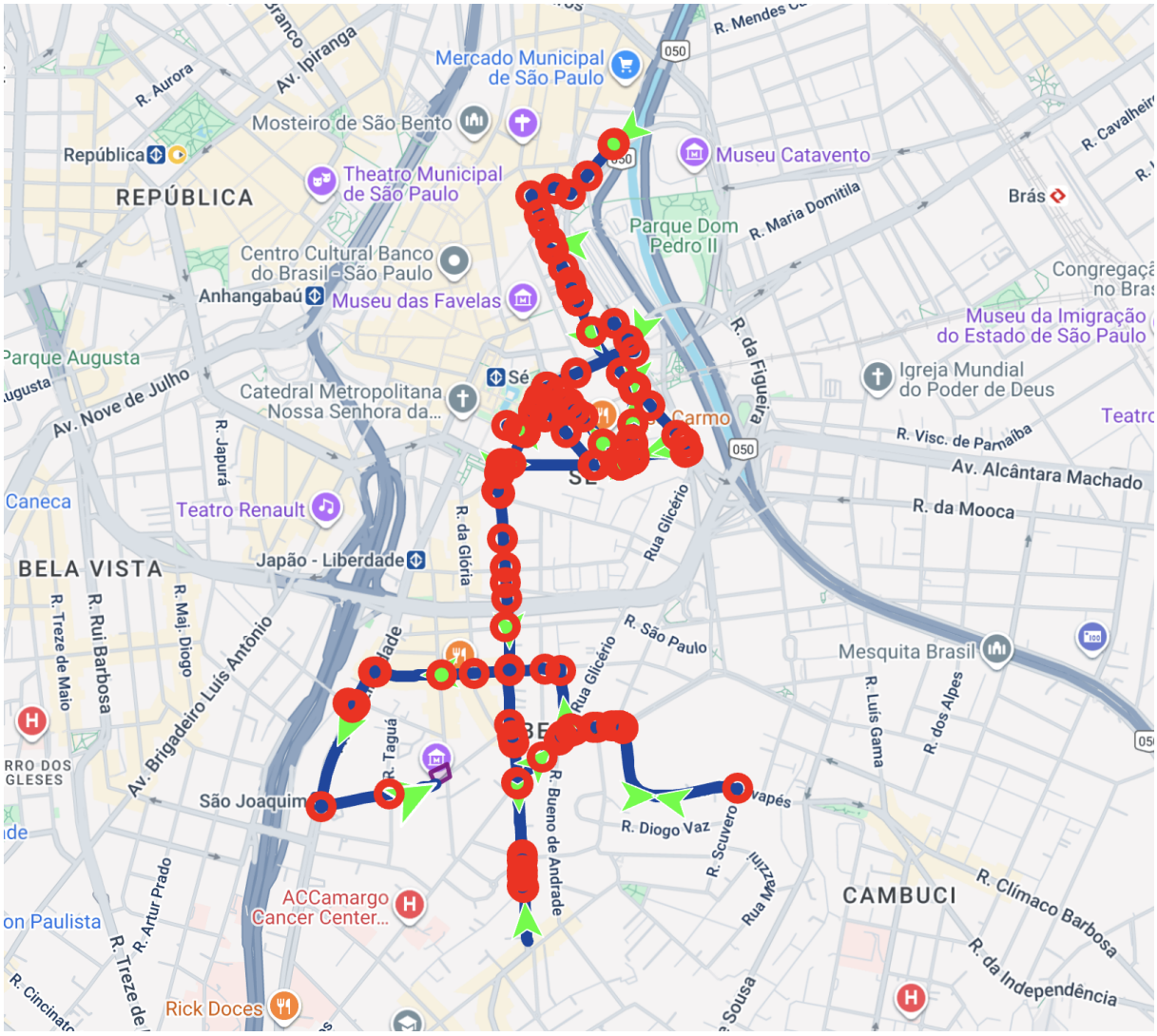

Figure 8: Museu da Imigracao Japonesa (Sao Paulo)