Mobile Communication Network-Guided Vision-Language Navigation in Multi-Agent Systems

Mohammad Bariq Khan*†, Daniel Gordon*, Xueli An*, and Falko Dressler†
*AWTL, Munich Research Center, Huawei Technologies, Munich, Germany
†School for Electrical Engineering and Computer Science, TU Berlin, Berlin, Germany
{mohammad.bariq.khan1, daniel.gordon, xueli.an}@huawei.com, dressler@ccs-labs.org

***Abstract*—The emergence of AI-native 6G networks introduces an agentic architectural paradigm where intelligent network entities can actively coordinate distributed systems. We leverage this capability to address scalability limitations in vision-and-language navigation (VLN) for robotic agents. Unlike existing VLN approaches designed for isolated operation, we propose a collaborative framework where a network agent orchestrates data collection from distributed robotic agents to construct a shared, query-able semantic map of the environment. By offloading computationally intensive mapping tasks from resource-constrained robots to the network, the framework improves operational efficiency. It also reduces redundancy and exploration cost by limiting each agent's search area. Furthermore, because semantic mapping is highly sensitive to data distortion, the network agent leverages its intrinsic access to communication metrics to guide robotic agents during data collection, minimizing transmission errors and ensuring robust map generation. Simulation results demonstrate that our off-board collaborative framework achieves mapping accuracy comparable to on-board individualistic methods, with negligible time overhead for participating robotic agents. Notably, it significantly lowers exploration costs for newly deployed agents, facilitating efficient adaptation to dynamic environments without compromising performance.**

***Index Terms*—AI-Native networks, Embodied AI, Large Language Models, Vision-Language Navigation, Collaborative Agents**

I. Introduction

The emergence of AI-native 6G networks, characterized by an agentic architectural paradigm [1], redefines communication networks as dynamic ecosystems of intelligent agents – autonomous entities that monitor network states and environmental conditions, optimize resource allocation through distributed decision-making, and coordinate with other agents for optimal service delivery. This paradigm shift transcends conventional connectivity, enabling networks to deliver advanced services to user terminals, which may themselves consist of intelligent agents. A key target application of this architecture is embodied AI systems, particularly networked robotics, where interconnected robotic agents execute coordinated operations. Such systems hold transformative potential across industries, enabling intelligent, collaborative, and context-aware task execution. Among the critical challenges in this domain is enabling robust autonomous navigation – a foundational capability for deploying scalable robotic solutions.

Traditionally, robotic navigation systems relied on deterministic algorithms for localization, mapping, and path planning [2], [3]. While these techniques work well for navigation tasks in static environments, they are far from creating truly intelligent systems that can learn through experience by interacting with the physical world, making decisions, and adapting to dynamic environments. Vision-and-language navigation (VLN) has emerged as a key enabling paradigm that aims to bridge this gap by integrating vision, language, and navigation into end-to-end systems [4]. More specifically, VLN tasks involve processing visual data and natural language instructions to generate navigational actions, enabling effective navigation through complex environments. A line of research in VLN involves simultaneous semantic mapping of the environment and grounding language to visual observations [5], [6]. These utilize pre-trained visual and language encoders to fuse visual-language features with a 3D reconstruction of the physical world, thereby generating a semantic map, commonly referred to as a Vision-Language Map. While state-of-the-art VLN frameworks excel in single agent operation, they impose significant limitations in practical deployments: each robot must independently explore environments, construct semantic maps, and execute navigation tasks – an inefficient paradigm for resource-constrained robots operating at scale.

A naive solution would offload compute-intensive semantic mapping, which usually involves 3D reconstruction, to edge servers but this introduces a critical vulnerability: the mapping process is highly sensitive to sensor data distortions (e.g., corrupted depth maps or point clouds) caused by transmission errors. The agentic, compute-native architecture of 6G networks fundamentally addresses this challenge by transforming the network itself into an intelligent mediator. Unlike passive edge servers, the network agent can leverage its intrinsic access to real-time network metrics, such as coverage maps and quality of service (QoS) profiles, to actively optimize robotic data collection and guarantee transmission reliability.

Motivated by this paradigm, we propose a collaborative mapping framework where multiple robotic agents collaborate with a network entity to construct a shared semantic map of the environment (Figure 1). Each robot performs distributed exploration while maintaining minimal onboard localization (2D position awareness), streaming its sensor observations (RGB-D, LiDAR) to the network entity for integration. The network entity – conceptually an autonomous AI agent but implemented here as an advanced network function operating across both control and data planes – fuses these multi-agent observations into a globally consistent semantic map. This

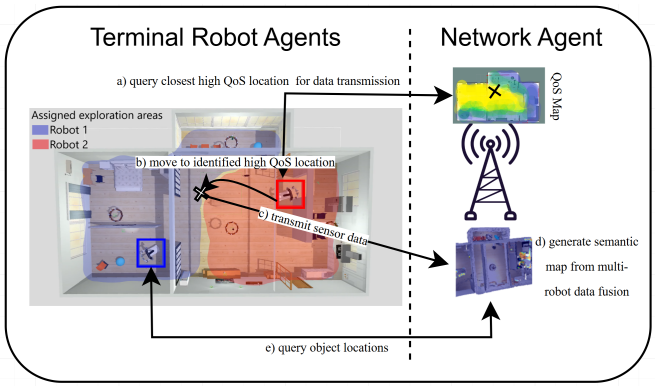


Fig. 1: Framework overview: The network agent generates a shared semantic map through multi-robot data fusion and optimizes data collection by directing robots to high QoS locations based on its QoS map. The shared map can then be queried for object locations (e.g. "fridge", "sofa").

allows robotic agents to dynamically query object locations, including those not yet observed in their individual exploration. While this framework can theoretically enable dynamic QoS negotiation where the robotic agents can request tailored QoS parameters (e.g., high throughput for sensor streaming or low latency for urgent queries), this study focuses primarily on validating the feasibility of the core collaborative mapping capability. The proposed framework offers three key advantages: (1) enhanced resource efficiency through offloading of compute-intensive semantic mapping to the network entity; (2) scalable mapping via distributed data collection that prevents redundant exploration by robotic-agents; and (3) improved robustness through optimized data transmission that minimizes sensor data distortion.

The key contributions of this work are threefold:

- We propose a novel network-assisted collaborative mapping framework where a network agent fuses multi-robot sensor data into a unified vision-language map, enabling robotic agents to query unseen object locations through coordinated language interactions.
- We develop a QoS-guided data collection protocol that leverages the network agent's global awareness of coverage conditions to optimize transmission reliability.
- We implement and evaluate the system in simulated environments, demonstrating its viability and potential benefits over individualistic and network-agnostic approaches.

II. Related Work

Recent breakthroughs in robotic navigation tasks have been significantly propelled by the use of pre-trained foundation models [4], largely due to their ability to provide rich spatial representations, transfer learning and multi-modal integration. For example, Clip-on-Wheels [7] leverages open-vocabulary models such as CLIP to propose a language-driven zero-shot object navigation (L-ZSON) task, which benchmarks the ability of agents to search for objects in unfamiliar environments. Recent work has focused on creating open-vocabulary semantic maps that allow natural language indexing, enabling more intuitive interaction with the environment [5], [6]. For instance, VLMaps [5] constructs environment maps using pre-collected offline datasets, while OVL-Maps [6] performs mapping during navigation, offering real-time adaptability. Despite these advancements, these approaches are designed for standalone single robot systems, leaving the potential of multi-robot collaboration largely unexplored.

A number of works have considered the problem of communication-aware navigation [8]–[11]. For example, Lindhe et al. [8] consider a robot with a predefined trajectory, and propose a periodic and controlled stopping policy based on measured SNR, to increase the average throughput. Luo et al. [10] introduce a system designed to maximize long-term throughput by concurrently optimizing the downlink transmission power at the access point (AP) and the motion trajectories of the robots. However, this approach primarily operates by steering robots away from regions of low communication quality, making it less suitable for dynamic exploration and map-building tasks. A more closely related study to our approach is ACHORD [11], wherein the authors propose a sophisticated multi-layer networking solution that intricately couples network architecture with high-level decision-making processes to enhance communication performance. ACHORD employs bandwidth prioritization and supports timely data transmission even in scenarios characterized by intermittent connectivity. In contrast to these prior works, our approach uniquely integrates network-guided data collection with dynamic semantic mapping of the unexplored environment, enabling the creation of a queryable, semantic map through seamless multi-robot collaboration.

III. Problem Formulation

We consider the challenge of scalable collaborative mapping for networked robotic systems operating under communication constraints. It involves N robotic agents $\{\mathcal{R}_1, \ldots, \mathcal{R}_N\}$ exploring an unknown environment which presents three core challenges:

Data Transmission Vulnerability: High-dimensional sensor streams (RGB-D/LiDAR) are susceptible to distortion when transmitted from regions of poor network quality, compromising map integrity.

Semantic Query Requirement: Agents must locate objects specified via textual descriptors (e.g., "fridge", "sofa") by querying a shared map, including objects not yet observed during their individual exploration trajectories.

Exploration Redundancy: Independent mapping by multiple agents leads to inefficient overlapping coverage and resource waste.

The goal is to enable: (1) reliable construction of a globally consistent semantic map $\mathcal{M}$ through multi-agent collaboration; and (2) accurate resolution of language queries for object coordinates $g_c = (x_g, y_g)$, despite dynamic network conditions. Crucially, the solution must ensure that language-

based navigation tasks remain viable when target objects fall outside an agent's local observation range.

IV. Methodology

The proposed system centers on network-mediated collaboration between robotic agents and the network agent which can be formalized through two components: (i) Network-guided data collection where it coordinates with robotic agents to ensure reliable sensor data transmission by directing them to high-QoS zones. (ii) Centralized map fusion where it integrates observations from multiple robotic-agents into a unified vision-language map.

The communication link quality between the network and robotic agents is characterized by key QoS metrics. In our proposed system, we assume that the network has initial knowledge of the environment's communication characteristics and maintains a coverage map, based on historical data and predictive models from NWDAF [12] or even estimates based on environmental features (e.g., open spaces are likely to have better signal strength). Furthermore, as the robots explore the environment, the coverage map can be continuously updated using the real-time QoS data shared by the robots. As such, the centralized QoS coordination by the network agent offers a critical advantage over terminal-only measurement by enabling predictive identification of optimal zones beyond a single robot's local perspective. By coordinating robotic agents through the network, the framework eliminates redundant exploration: each agent focuses on distinct sub-regions guided by the network's global perspective, collectively covering the environment without overlap. This spatial division of labor reduces per-agent exploration costs compared to single-agent systems. Furthermore, newly deployed agents inherit the network's shared map, bypassing solo exploration entirely and enabling immediate task execution.

A. QoS-Guided Data Collection

Each robot $\mathcal{R}_i$ is assigned an exploration region $\mathcal{S}_i$ based on its initial position p_i^{init}. A transmission event is triggered when the robot satisfies either a temporal threshold (T seconds elapsed since last transmission) or a data-volume threshold (e.g, F RGB-D frames acquired). Upon triggering, $\mathcal{R}_i$ requests the network agent for the optimal transmission positions:

1) **Trigger Condition**: For robot $\mathcal{R}_i$ at position p_i^{curr}:

$$\text{Transmit if } |\mathcal{I}_t^{(i)}| \geq F \text{ or } t - t_{\text{last}}^{(i)} \geq T \tag{1}$$

2) **QoS Target Selection**: The network agent constructs a Voronoi tessellation $\mathcal{V}$ over all high-QoS points $\{p_1^*, \ldots, p_n^*\}$ satisfying $\mathcal{Q}(p_k^*) \geq \theta_{\text{QoS}}$. Each Voronoi cell $\mathcal{C}_i$ is defined as:

$$\mathcal{C}_i = \left\{ p \in \mathcal{E} \,\middle|\, \|p - p_i^*\|_2 \leq \|p - p_j^*\|_2, \forall j \neq i \right\}, \quad \text{where } \mathcal{E} \text{ represents the entire exploration area} \tag{2}$$

For an agent at position p_i^{curr}, the target point p_i^* is:

$$p_i^* = \arg\min_{p \in \mathcal{P}} \|p_i^{\text{curr}} - p\|_2, \quad \text{where } \mathcal{P} = \{p \in \mathcal{E} \mid \mathcal{Q}(p) \geq \theta_{\text{QoS}}\} \tag{3}$$

3) **Navigation**: Robot $\mathcal{R}_i$ plans path $\mathcal{P}_i$ to p_i^*:

$$\mathcal{P}_i = \arg\min_{\mathcal{P}} \|\mathcal{P}\|_2 \quad \text{s.t. } \mathcal{P} \subset \mathcal{S}_i \tag{4}$$

Exploration terminates when the visual language map $\mathcal{M}$ achieves coverage ratio:

$$\frac{|\mathcal{M}_{\text{covered}}|}{|\mathcal{M}|} \geq \gamma \tag{5}$$

After the initial map construction, robotic agents continue to transmit incremental updates to maintain the global map's accuracy in dynamic environments. The update frequency can be adaptively adjusted based on environmental dynamics or operational requirements.

The above procedure is summarized in Algorithm 1.

Algorithm 1 QoS-Guided Visual Language Mapping

Input: Robots $\mathcal{R}_1, ..., \mathcal{R}_N$, QoS map $\mathcal{Q}(p)$, thresholds F, T, γ
Output: Visual language map $\mathcal{M}$
for each robot $\mathcal{R}_i \in \{\mathcal{R}_1, ..., \mathcal{R}_N\}$ **do**
    Initialize region $\mathcal{S}_i$
    $t_{\text{last}}^{(i)} \leftarrow 0$, $\mathcal{I}_t^{(i)} \leftarrow \emptyset$
while $\frac{|\mathcal{M}_{\text{covered}}|}{|\mathcal{M}|} < \gamma$ **do**
    for each robot $\mathcal{R}_i$ **do**
        Collect RGB-D frame $(I_r^{(i)}, I_d^{(i)})$ at position p_i^{curr}
        $\mathcal{I}_t^{(i)} \leftarrow \mathcal{I}_t^{(i)} \cup \{(I_r^{(i)}, I_d^{(i)}, p_i^{\text{curr}})\}$
        if $|\mathcal{I}_t^{(i)}| \geq F$ **or** $t - t_{\text{last}}^{(i)} \geq T$ **then**
            Network agent computes:
            $p_i^* \leftarrow \arg\max_{p \in \mathcal{B}(p_i^{\text{curr}}, R_s)} \mathcal{Q}(p)$
            Send p_i^* to $\mathcal{R}_i$
            $\mathcal{R}_i$ navigates to p_i^* via $\mathcal{P}_i \leftarrow \arg\min_{\mathcal{P} \subset \mathcal{S}_i} \|\mathcal{P}\|_2$
            Transmit $\mathcal{I}_t^{(i)}$ to network agent
            $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{I}_t^{(i)}$
            $\mathcal{I}_t^{(i)} \leftarrow \emptyset$, $t_{\text{last}}^{(i)} \leftarrow t$
    Network agent updates $\mathcal{M}_{\text{covered}}$
Return $\mathcal{M}$

B. Dynamic VLMap Construction

The VLMap framework constructs a globally consistent 3D visual-language map by incrementally fusing asynchronous, frame-wise RGB-D (RGB frames with depth) observations from distributed agents into a unified representation. The system processes each input frame by first reconstructing 3D points in the camera coordinate system via depth back-projection and intrinsic calibration. These points are then transformed into a shared global frame using estimated camera pose data transmitted along with the RGB-D data, enabling multi-agent data integration. The global map is structured as a dynamically expandable voxel grid, where each voxel stores semantic features (extracted via a visual-language encoder), RGB color values, and a confidence weight to quantify observation reliability. The voxel grid is updated incrementally per frame, enabling real-time adaptation to new

observations without global recomputation. The framework's modular design supports scalable deployment across multi-agent systems, as each agent independently processes its sensor stream while contributing to the shared global map through pose-synchronized transformations. We denote the RGB frame at time step t as $F_t \in \mathbb{R}^{m\times n\times 3}$, and the corresponding depth map as $D_t \in \mathbb{R}^{m\times n}$, where m and n represent the image height and width, respectively. For each pixel (u, v) in the image, the depth value $D_t(u, v)$ is used to back-project the 3D point $\mathbf{p}_t^{\text{cam}} = (x, y, z)^\top$ in the camera coordinate system through the inverse projection model:

$$\mathbf{p}_t^{\text{cam}} = \mathbf{K}^{-1} \begin{bmatrix} u \cdot D_t(u, v) \\ v \cdot D_t(u, v) \\ D_t(u, v) \end{bmatrix} \tag{6}$$

where $\mathbf{K} \in \mathbb{R}^{3\times 3}$ is the camera intrinsic matrix, and (u, v) are the pixel coordinates of the depth map. The global map $\mathcal{M}$ is represented as a voxel grid with cell size c_s, where each grid cell $\mathbf{g}_{ijk} \in \mathcal{M}$ at coordinates (i, j, k) contains: Semantic feature vector $\mathbf{f}_{ijk} \in \mathbb{R}^d$ obtained from the LSeg encoder, RGB color $\mathbf{c}_{ijk} \in \{0, \ldots, 255\}^3$, Confidence weight $w_{ijk} \in \mathbb{R}^+$. The map is updated using an incremental fusion strategy. Let $\mathbf{T}_t^{\text{base}\leftarrow\text{cam}} \in SE(3)$ denote the camera pose transformation matrix at time t. The observed 3D point $\mathbf{p}_t^{\text{cam}}$ is transformed into the global base frame using:

$$\mathbf{p}_t^{\text{base}} = \mathbf{T}_t^{\text{base}\leftarrow\text{cam}} \mathbf{p}_t^{\text{cam}} \tag{7}$$

The corresponding grid cell indices (i, j, k) are computed as:

$$(i, j, k) = \left\lfloor \frac{\mathbf{p}_t^{\text{base}} - \mathbf{p}_{\text{min}}}{c_s} \right\rfloor \tag{8}$$

where $\mathbf{p}_{\text{min}}$ represents the origin of the map. A height map $\mathcal{H} \in \mathbb{R}^{m\times n}$ is maintained to track the maximum observed z-coordinate per ground grid cell, which helps in handling occlusions. The height map is updated as:

$$\mathcal{H}(i, k) = \max(\mathcal{H}(i, k), j) \tag{9}$$

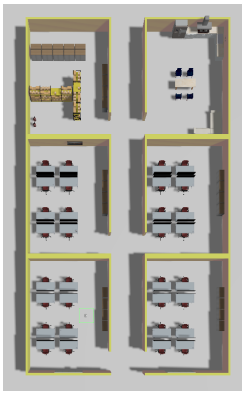
where j represents the vertical grid index, ensuring that higher objects or obstacles are prioritized.

To integrate new features, we follow the same strategy as in [5] and apply an exponentially decaying weight based on the radial distance from the sensor. The semantic features $\mathbf{f}_{ijk}$ and color values $\mathbf{c}_{ijk}$ are updated via weighted averaging.

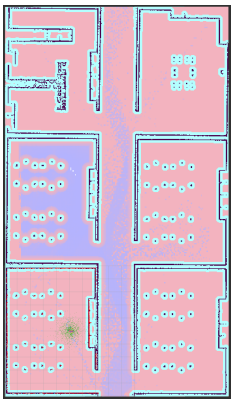
C. Communication Model

We obtain the wireless network coverage map for our simulated environment (Section V-A) using the ray-tracing capabilities of Sionna for an operating of frequency of 2.14 GHz. The generated coverage map, depicted in Figure 2b predicts the path loss at different grid locations in the given environment, which is subsequently used to estimate metrics such as Signal-to-Noise Ratio (SNR), Bit Error Rate (BER), and Packet Error Rate (PER). The received power at a given location is given by:

$$P_{\text{rx}}(\text{dBm}) = P_{\text{Tx}} + G_{\text{channel}}$$



(a) Simulated office scene in gazebo



(b) Network coverage map (coverage zones highlighted in purple)

Fig. 2: Simulated environment and coverage map.

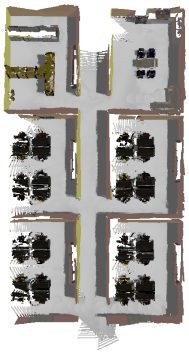


Fig. 3: Baseline (On-Board) VLMap.

Fig. 4: VLMap in network-agnostic scenario.

where G_{channel} is the path gain obtained from the radio map. The SNR_{DL} and SNR_{UL} would then be:

$$\text{SNR}_{\text{DL}}(\text{dB}) = P_{\text{rx}} - N_{\text{floor}}$$

$$\text{SNR}_{\text{UL}} = \text{SNR}_{\text{DL}} + (P_{\text{Tx,UL}} - P_{\text{Tx,DL}}) + \delta G,$$

where δG is a random variable modelling the difference in channel gain. For 256-QAM, the BER can be approximated as:

$$\text{BER} \approx \frac{4}{\log_2(M)} Q\left(\frac{3 \cdot \text{SNR}_{\text{linear}}}{M - 1}\right), \quad \text{with } M = 256$$

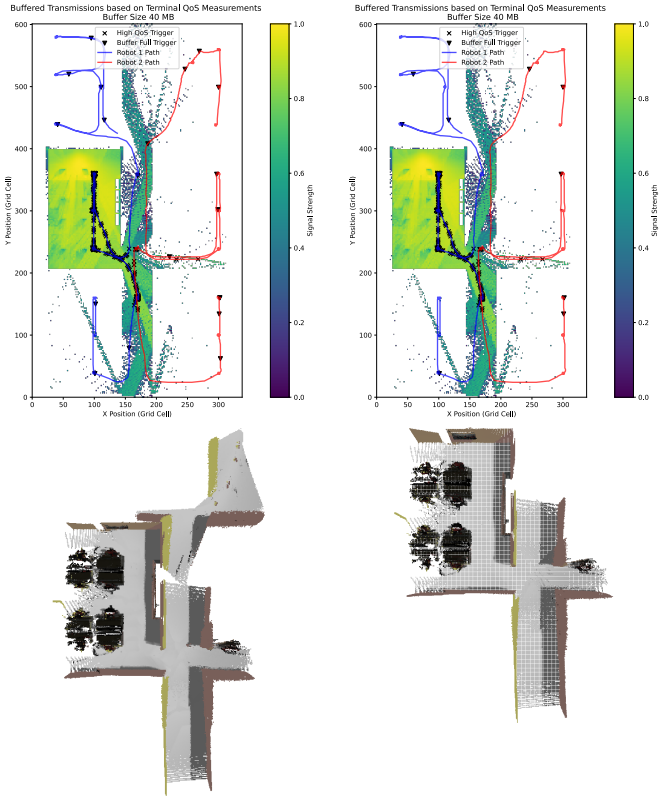
where $\text{SNR}_{\text{linear}} = 10^{\frac{\text{SNR (dB)}}{10}}$ and $Q(x)$ is the Gaussian Q-function. The Packet Error Rate (PER) is then computed as:

$$\text{PER} = 1 - (1 - \text{BER})^{L_{\text{packet}}\times 8}$$

V. Experiments

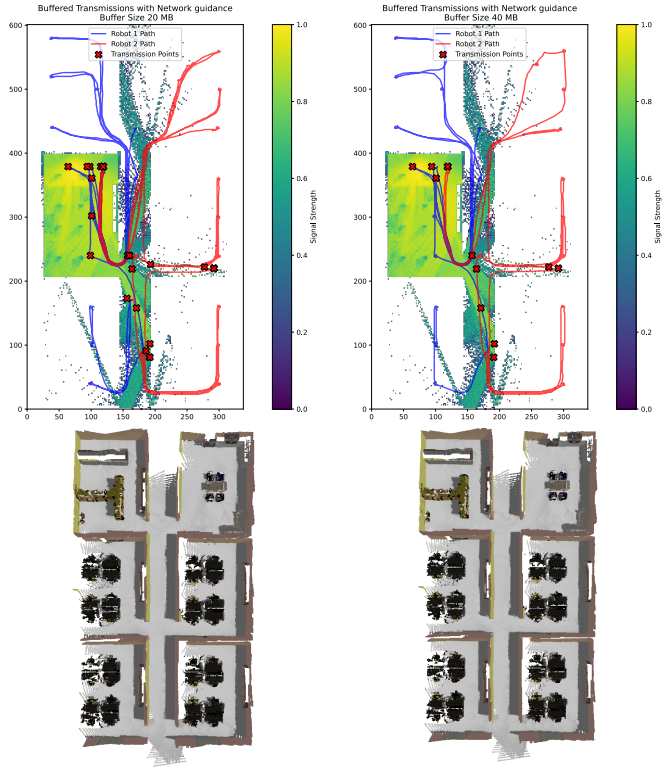
A. Experiment Setup

We evaluate the framework in a Gazebo-simulated office environment (17m × 30m × 2m) featuring corridors and six rooms. Two TurtleBot3 agents perform collaborative exploration with manually partitioned regions for experimental consistency. The QoS threshold θ_{QoS} is set to -70dB and the



(a) Buffer Size 20 MB (b) Buffer Size 40 MB

Fig. 5: Terminal agent determined transmission points and corresponding VLMaps for two buffer sizes.



(a) Buffer Size 20 MB (b) Buffer Size 40 MB

Fig. 6: Network agent determined transmission points and corresponding VLMaps for two buffer sizes.

mapping phase terminates when each robotic agent has fully covered its assigned region. We compare four approaches:

1) **On-Board Mapping**: Each robot independently maps the entire environment (baseline).
2) **Network-Agnostic**: The robots immediately transmit all available sensor data without considering current network conditions at their location.
3) **Terminal-Agent-Controlled**: Robots monitor local network quality and buffer sensor data during exploration, transmitting it upon reaching high-QoS areas. If the buffer reaches capacity, data is sent immediately, regardless of network conditions.
4) **Network-Agent-Optimized**: Proposed QoS-guided transmission.

B. Evaluation Metrics

Performance is quantified using:

- **Task Success Rate**: To evaluate the semantic accuracy of the generated VLmaps, we test the maps on 9 navigation tasks: agents parse language instructions, query candidate object coordinates from the network agent, and navigate to these locations. A navigation task is deemed successful if at least one of the returned locations aligns with the ground-truth position of the target object.

TABLE I: Performance metrics for the four scenarios with different buffer sizes.

	Task Success Rate (%)			Normalized Map Coverage			Average Exploration Time per Robot (Minutes)		
Buffer Size (MB)	0	20	40	0	20	40	0	20	40
On-Board (Baseline)	77.8	-	-	1.0	-	-	24.2	-	-
Network-Agnostic	22.2	-	-	0.53	-	-	12.5	-	-
Terminal-Agent-Controlled	-	33.3	33.3	-	0.31	0.22	-	12.5	12.5
Network-Agent-Optimized	-	77.8	77.8	-	1.0	1.0	-	29.7	21.6

- **Normalized Map Coverage**: The proportion of the environment's total area that has been successfully mapped, quantified by the fraction of grid cells containing at least one semantic feature. Min-max normalization relative to the best-performing on-board baseline is applied, scaling the coverage metric between 0 (no coverage) and 1 (coverage equivalent to the baseline).
- **Average Exploration Time**: The duration required for each robot to traverse its assigned exploration region.

C. Results

a) Mapping Performance: The baseline vision-language map constructed by individual robotic agents, shown in Figure 3, achieves 85% environment coverage with an average exploration time of 24.2 minutes. As demonstrated in Table I, the network-assisted approach matches this coverage performance while enabling collaborative mapping, whereas the network-agnostic method suffers nearly 50% coverage

degradation evidenced by impaired reconstruction in Figure 4. Terminal-controlled transmission performs even worse, with large unmapped areas visible in Figure 5 due to substantial data loss in poor-coverage zones; a problem exacerbated by increased buffer sizes as shown in the transmission plots of Figure 5. However, the terminal-controlled method does observe lower distortion in successfully reconstructed areas compared to the network-agnostic approach.

b) Navigation Accuracy: As quantified in Table I, the network-assisted framework achieves 77.8% success rate on language-guided navigation tasks, matching the baseline performance and confirming minimal transmission-induced semantic distortion. This contrasts sharply with uncoordinated approaches where data loss and feature mislocalization cause significant degradation in query resolution accuracy. The results validate that network awareness is essential for maintaining both map coverage and semantic fidelity required for reliable resolution of language-based object queries.

c) Exploration Efficiency: In terms of the exploration costs, even two agents reduce exploration time considerably in comparison to the baseline, though residual overhead persists from the network agent's structural unawareness (e.g., undetected walls causing suboptimal transmission points in Figure 6). This overhead remains marginal versus solo mapping. Critically, any newly deployed agents can benefit from near-zero exploration costs, as they may directly query pre-existing maps from the network agent.

VI. Limitations and Future Work

This work presents a simulation-based proof-of-concept for network-assisted multi-robot vision-language mapping under simplified communication assumptions. The primary limitation lies in the use of a static signal strength map as the sole QoS indicator. While signal strength effectively captures spatial variations in link quality and allows us to model the essential communication effects (sensor data loss or distortion), it does not account for dynamic network phenomena such as link-layer contention or traffic-driven congestion. These factors are particularly relevant under high traffic conditions, but are not modeled in our current setup. Additionally, the total mission time is not directly reported due to the sequential simulation of each robot's trajectory. Although BER and PER simulate the impact of packet loss, delays due to queuing and processing latency at the network agent remain unaccounted for.

To address these gaps, we plan to validate our approach on a physical multi-robot setup over a 5G testbed. This setup will enable empirical measurement of latency and throughput, and allow us to evaluate system performance under realistic wireless conditions as well as account for additional uncertainties such as actuation drift and localization errors.

On the algorithmic front, future work will focus on designing a centralized coordination algorithm that jointly optimizes communication and exploration. Specifically, we aim to assign exploration regions to robots in a way that maximizes overall coverage while simultaneously minimizing time overhead due to transmission detours.

VII. Conclusions

This paper presented a network-assisted collaborative framework for vision-language navigation that addresses the scalability limitations of individual robotic mapping. By leveraging a network agent with global awareness of communication conditions, our approach enables reliable construction of shared semantic maps through QoS-guided data collection, while supporting language-based object queries for unseen targets. Experimental validation in simulated environments demonstrated that the framework achieves map coverage and task navigation accuracy parity with on-board mapping, in contrast to the uncoordinated baselines, thus confirming the viability of network-mediated collaboration for scalable VLN systems.

References

[1] "Experiential Networked Intelligence (ENI); Study on AI Agents based Next-generation Network Slicing," ETSI, Sophia Antipolis, France, Standardization Group Report GR ENI 051 V4.1.1, Feb. 2025, pp. 1–32. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/ENI/001_099/051/04.01.01_60/gr_ENI051v040101p.pdf.

[2] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual Simultaneous Localization and Mapping: A Survey," *Artificial Intelligence Review*, vol. 43, pp. 55–81, Jan. 2015.

[3] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic Roadmaps for Path Planning in High-dimensional Configuration Spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, Aug. 1996.

[4] W. Wu, T. Chang, X. Li, Q. Yin, and Y. Hu, "Vision-Language Navigation: A Survey and Taxonomy," *Neural Computing and Applications*, vol. 36, no. 7, pp. 3291–3316, Mar. 2024.

[5] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual Language Maps for Robot Navigation," in *IEEE ICRA 2023*, London, United Kingdom: IEEE, May 2023, pp. 10 608–10 615.

[6] S. Wen, Z. Zhang, Y. Sun, and Z. Wang, "OVL-MAP: An Online Visual Language Map Approach for Vision-and-Language Navigation in Continuous Environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3294–3301, Apr. 2025.

[7] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "CoWs on Pasture: Baselines and Benchmarks for Language-driven Zero-shot Object Navigation," in *IEEE/CVF CVPR 2023*, Vancouver, Canada: IEEE, Jun. 2023, pp. 23 171–23 181.

[8] M. Lindhe and K. H. Johansson, "Using Robot Mobility to Exploit Multipath Fading," *IEEE Wireless Communications*, vol. 16, no. 1, pp. 30–37, Feb. 2009.

[9] S. Caccamo, R. Parasuraman, L. Freda, M. Gianni, and P. Ögren, "RCAMP: A Resilient Communication-aware Motion Planner for Mobile Robots with Autonomous Repair of Wireless Connectivity," in *IEEE/RSJ IROS 2017*, Vancouver, Canada: IEEE, Sep. 2017, pp. 2010–2017.

[10] R. Luo, T. Hui, and W. Ni, "Communication-Aware Path Design for Indoor Robots Exploiting Federated Deep Reinforcement Learning," in *IEEE PIMRC 2021*, Virtual Conference: IEEE, Sep. 2021, pp. 1197–1202.

[11] M. Saboia, L. Clark, V. Thangavelu, et al., "ACHORD: Communication-Aware Multi-Robot Coordination With Intermittent Connectivity," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 184–10 191, Oct. 2022.

[12] "Architecture enhancements for 5G System (5GS) to support network data analytics services," 3GPP, Sophia Antipolis, France, Technical Specification 23.288 V19.2.0, Mar. 2025. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3579.