

Plot-and-Ask: Multimodal Local LLMs in Smart Environments for Visual IoT Analytics

Aygun Varol*, Asif Shaikh*, Naser Hossein Motlagh[†], Mirka Leino[‡], and Johanna Virkki*

*Faculty of Information Technology and Communication Sciences, Tampere University, Finland

[†]Department of Computer Science, University of Helsinki, Finland

[‡]Faculty of Technology, Satakunta University of Applied Sciences, Finland

*{aygun.varol, asif.shaikh, johanna.virkki}@tuni.fi, [†]naser.motlagh@helsinki.fi, [‡]mirka.leino@samk.fi

Abstract—Smart environments generate continuous heterogeneous sensor data that require real-time processing to support occupant services and decision-making. Existing approaches either compromise data sovereignty through cloud dependencies or employ rigid rule-based systems that lack natural-language interaction. This paper presents a local multimodal analytics architecture for smart indoor environments that combines edge-resident Multimodal Large Language Models (MLLMs) with a Knowledge Graph (KG) policy layer. The proposed *plot-and-ask* workflow couples automatic time-series visualization with vision-enabled language interpretation, intended to let non-expert users explore indoor sensing data without programming expertise. The system supports natural-language-to-SQL querying, plot generation with KG-derived comfort bands, and multimodal plot interpretation while keeping all processing on local hardware. We evaluate the framework in a two-week indoor air quality deployment across office, hallway, and kitchen locations using four local models: Gemma 3 4B, Ministral 3 3B, Qwen 3 4B, and Granite 3 2B. All four achieved 100% accuracy for structured query synthesis and plot generation. For plot interpretation, Ministral 3 3B and Gemma 3 4B best balanced accuracy and latency, reaching 100% and 96.67% accuracy with 5.84s and 5.34s end-to-end latency, respectively. A KG-grounding ablation showed the largest gain for Qwen 3 4B (+25.0 pp), with latency reduced from 27.98s to 15.21s, and a supplementary indoor-scene safety check confirms that the same models support direct image understanding. Together, these results show that plot-based summarization and lightweight KG grounding improve local multimodal IoT analytics under edge constraints. Overall, hybrid KG-MLLM architectures can support privacy-preserving visual IoT analytics at the edge while being designed for interactive exploration by non-expert users.

Index Terms—Internet of Things, Sensor networks, Smart environments, Edge AI, Knowledge Graph, Visual analytics.

I. INTRODUCTION

THE ubiquitous deployment of Internet of Things (IoT) sensors in smart environments produces continuous streams of heterogeneous data that must be interpreted by building occupants, facility managers, and decision-makers. Traditional interfaces require technical expertise to query databases or read raw sensor values, creating barriers for non-expert users [1]. Large Language Models (LLMs) open new possibilities for natural-language interaction with IoT systems, letting occupants obtain actionable insights through conversational queries without programming knowledge and moving from rigid dashboards to flexible analytics [2]. However, leveraging LLMs for smart-environment analytics in-

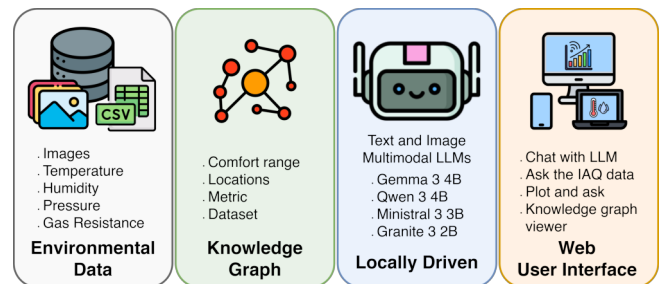


Fig. 1. Four-layer system architecture: data ingestion, Knowledge Graph reasoning, local multimodal LLM inference, and a web interface for natural-language queries and plot-and-ask.

roduces significant privacy concerns when sensor data—including occupancy patterns, thermal preferences, and behavioral indicators—is transmitted to proprietary cloud Application Programming Interfaces (APIs) for processing. Cloud-dependent systems compromise data sovereignty, introduce unpredictable costs, and create network dependencies that affect system responsiveness [3], [4]. To address these limitations, local edge inference is an attractive alternative as it keeps sensitive environmental data on-site, avoids external network calls, and supports predictable response latencies for building interaction and automation [5].

Beyond privacy, smart-environment outputs should be grounded in explicit domain knowledge: when LLMs recommend ventilation, heating, or occupancy actions, those outputs should rest on verifiable comfort policies rather than model priors alone. Knowledge Graphs (KGs) encode threshold constraints, domain semantics, and policy knowledge in machine-readable form, enabling grounded and traceable responses [6], [7]. Smart environments also span multiple modalities—time-series measurements, tabular data, and visual scenes—and Multimodal LLMs (MLLMs) process textual and visual inputs within one pipeline, unlike single-modality models that cannot interpret images or plots [8]. This paper addresses the lack of privacy-preserving, policy-grounded, and multimodal analytics for non-expert users in smart indoor environments. To this end, we present a local hybrid architecture that combines MLLMs with a KG-based policy layer (Fig. 1) and examine whether lightweight local MLLMs can generate accurate natural-language-to-SQL (Structured Query Language) queries

on commodity edge hardware, whether the plot-and-ask workflow supports accurate multimodal interpretation of time-series plots within interactive latency, and whether KG grounding improves the policy alignment of generated visualizations and status summaries.

This work makes three main contributions. First, it presents a local multimodal analytics framework that supports natural-language querying, plot generation, and plot interpretation entirely on edge hardware without external network calls. Second, it introduces a plot-and-ask workflow that uses plot-based visual summarization to reduce context load for local multimodal models. Third, it provides an empirical characterization of four local open-weight MLLMs across structured querying, visualization, and plot interpretation tasks in a two-week indoor air-quality deployment. The implementation is available as open-source software to support reproducibility¹.

II. RELATED WORK

Local and Edge-Resident Language Models: Recent work deploys language models at the edge to reduce cloud reliance and strengthen privacy. LLMEde [9] studies localized inference under resource constraints, EdgeShard [10] examines distributed execution for edge-resident models, and prompt-side IoT data preprocessing can narrow the local-to-cloud accuracy gap for environmental monitoring [11]. HoloLLM [8] demonstrates multimodal reasoning in smart homes across multiple sensing modalities. These studies advance edge-oriented design, IoT prompt preprocessing, and multimodal reasoning, but do not address interactive natural-language analytics for indoor sensing coupled with plot generation and policy-grounded visual feedback.

Knowledge-Graph Grounding and Retrieval-Augmented Reasoning: Knowledge graphs and retrieval-augmented generation (RAG) improve factual consistency and domain grounding. Combining symbolic knowledge with LLMs improves reasoning across heterogeneous standards [6], [7], and Telecom-RAG [12] guides technical question answering with retrieved knowledge. We adopt a lightweight design in which KG-derived policies extend beyond prompt grounding to support plot overlays and status-card classification.

Visual Analytics and Pervasive Displays: Research on pervasive displays and visual analytics has addressed display walls, situated analytics, and collaborative interaction [13]–[15], but typically relies on manually designed visualizations or fixed dashboards. By contrast, our system accepts a natural-language request, retrieves the relevant sensing data, generates a plot automatically, and passes it to a local vision-enabled model for interpretation.

Positioning of the Present Work: Table I positions this work against prior research, which addresses local inference, multimodal processing, and knowledge grounding individually. To our knowledge, prior work has not jointly evaluated edge-based multimodal inference, natural-language sensor querying, automatic plot generation, and KG-grounded interpretation within a smart-environment analytics workflow.

III. METHODOLOGY

Our system transforms IoT sensor data into plots and natural-language explanations on local hardware for pervasive-display contexts. It comprises a backend for local LLM inference, an in-memory database, a Knowledge Graph encoding domain policies, and a web interface with interaction panels. We deployed lightweight multimodal LLMs Gemma 3 4B [16], Ministral 3 3B [17], Qwen 3 4B [18], and Granite 3 2B [19] on an edge device using Ollama² as the local model-serving backend. The framework calls the local Hypertext Transfer Protocol (HTTP) API to run inference over sensor data, plots, and KG-enhanced prompts: it synthesizes SQL for sensor retrieval, injects KG-derived context into prompts, and generates comfort-grounded display cards, all exposed through HTTP endpoints and a four-panel web interface.

Model Selection: We selected four lightweight multimodal LLMs (2–4B parameters) for edge deployability, multimodal capability, and architectural diversity. Compact models maintain interactive latency on constrained hardware [10], while plot-and-ask requires vision-language processing for plot interpretation [8]. The models span distinct lineages (Gemma 3, Ministral 3, Qwen 3, Granite 3), enabling comparison of performance-latency trade-offs.

Design Principles: Three principles guide the architecture. *Verifiable grounding* [6] makes model outputs traceable to data sources, KG facts, or user inputs. *Local execution* [9], [20] keeps inference, data processing, and knowledge reasoning on the deployment host. *Transparent operation* [21] exposes intermediate representations—generated SQL, generated SPARQL (SPARQL Protocol and RDF Query Language), and retrieved knowledge context—to support trust calibration.

System Architecture: Fig. 1 illustrates the four-layer architecture. The Environmental Data Layer loads timestamped measurements from tabular comma-separated values (CSV) files into an in-memory DuckDB relation. The dataset contains minutely averaged indoor air quality (IAQ) metrics—temperature, humidity, pressure, and gas resistance—across kitchen, office, and hallway locations. The Knowledge Graph Layer defines classes for *Metric*, *Location*, *PervasiveDisplay*, and *Policy*, with properties encoding recommended ranges (e.g., *recommendMin*, *recommendMax*, *atLocation*). Temperature is associated with a recommended range of 20–26 °C and humidity with 30–60%. These thresholds are used for plot shading and display-card classification. The Local LLM Orchestration Layer communicates with Ollama through its local HTTP API. Multi-turn conversations invoke the chat endpoint, while task-specific prompts are used for SQL synthesis and plot interpretation. The Web User Interface Layer exposes endpoints for chat, data querying, plot generation, KG access, and display-card generation, presented through four coordinated panels.

Natural Language to SQL Workflow: Natural-language questions are translated into executable SQL through a prompt-only approach (Fig. 2). The backend builds a constrained

¹<https://github.com/AygunVarol/multimodal-local>

²<https://ollama.com/>

TABLE I
COMPARISON BETWEEN RELATED WORK AND THE PROPOSED SYSTEM. ✓ = YES, ● = PARTIAL, – = NO.

Approach	Local Inference	Multimodal	KG / RAG Grounding	Natural-Language Sensor Querying	Auto Plot Generation	Smart-Env. Deployment
Edge / local LLM systems [9]–[11]	✓	●	–	–	–	●
KG-grounded LLM systems [6], [12]	–	–	✓	●	–	–
Pervasive display / visual analytics [13]–[15]	–	●	–	–	●	✓
This work	✓	✓	✓	✓	✓	✓

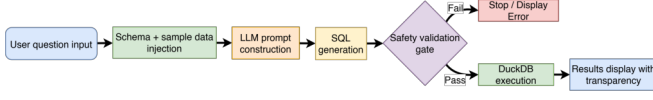


Fig. 2. Natural-language-to-SQL workflow with transparent query generation, validation, and execution.

prompt with (i) the table name (*iaq*), (ii) schema information from the DataFrame, and (iii) a JavaScript Object Notation (JSON) row sample to align column naming, and instructs the model to output exactly one DuckDB *SELECT* statement without prose or markdown. The backend then strips code and applies a safety gate that rejects non-*SELECT* statements and data-manipulation keywords (*INSERT*, *UPDATE*, *DELETE*, *DROP*) before executing the query against the in-memory relation. The response returns the generated SQL, column headers, and result rows, letting users verify question–query alignment while preventing inadvertent data mutation.

Plot-and-Ask Workflow: Local multimodal LLMs have limited context windows, constraining how much raw time-series data can be supplied. The plot-and-ask workflow instead turns longer sensing periods into a plot that summarizes temporal structure visually (Fig. 3). The *PLOT* endpoint parses requests such as “plot temperature for 2025-08-19 in office”, extracting metric, location, and date, then retrieves matching rows and renders a time-series plot. When the KG defines a recommended range for the metric, that band is shaded in the plot. The image is returned to the client and can be attached to a follow-up prompt such as “explain the trend” or “do you see any anomaly?”, with text and image forwarded to the local MLLM for interpretation. Compared with raw minute-level tables, the plot is a compact representation of one-day sensing data that preserves trends, threshold crossings, and anomalies, and the same image can be reused across follow-up questions rather than re-serializing long tables into the prompt.

Knowledge Graph and Display Card Generation: The Knowledge Graph (i) provides semantic context for prompt augmentation [12], (ii) supplies recommended ranges for visualization and display cards, and (iii) enables natural-language KG querying. For chat, a lightweight keyword-matching mechanism extracts KG labels overlapping the user message and serializes the matching statements into the prompt. The KG can be queried two ways: expert users submit raw SPARQL *SELECT* queries through the API, while non-expert users ask in natural language and receive both the generated SPARQL and the result bindings. The display-card endpoint computes aggregates for given dates and locations and assigns one

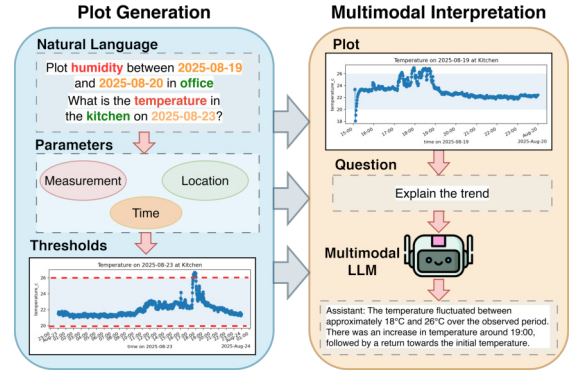


Fig. 3. Plot-and-ask workflow. Natural-language plot requests generate time-series visualizations with KG-derived comfort range shading (left). Users then attach the plot to follow-up questions, forwarding image and text to vision-enabled models for interpretation (right).

of three states—*ok*, *warn-low*, or *warn-high*—by comparing the value against KG-defined ranges, letting the symbolic policy layer participate directly in output generation rather than serving as a passive source.

Interface Design: The interface presents four adjacent panels to preserve context during sensemaking (Fig. 4): (i) a chat column with image attachments, (ii) an IAQ data panel showing generated SQL and result tables, (iii) a plot-and-ask panel, and (iv) a KG panel, letting users compare numeric summaries, visualizations, and explanations without losing history.

IV. EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

We deployed the system in a real-world setting spanning three locations: a shared kitchen, an office, and a hallway. IAQ sensors collected minutely averaged temperature, humidity, pressure, and gas-resistance measurements over two weeks in August 2025, capturing diurnal patterns, occupancy-driven variations, and ventilation events. Evaluations ran on commodity hardware (Intel Core i7-10750H, 32 GB RAM, RTX 2060 6 GB GPU); the system operates entirely locally with open-weight models.

A. Evaluation Tasks

We evaluated three core workflows aligned with the research questions—(i) natural-language-to-SQL synthesis, (ii) display-card generation, and (iii) visual IoT analytics via plot-and-ask—and additionally probed the models’ general image understanding through a supplementary indoor-scene safety classification task.

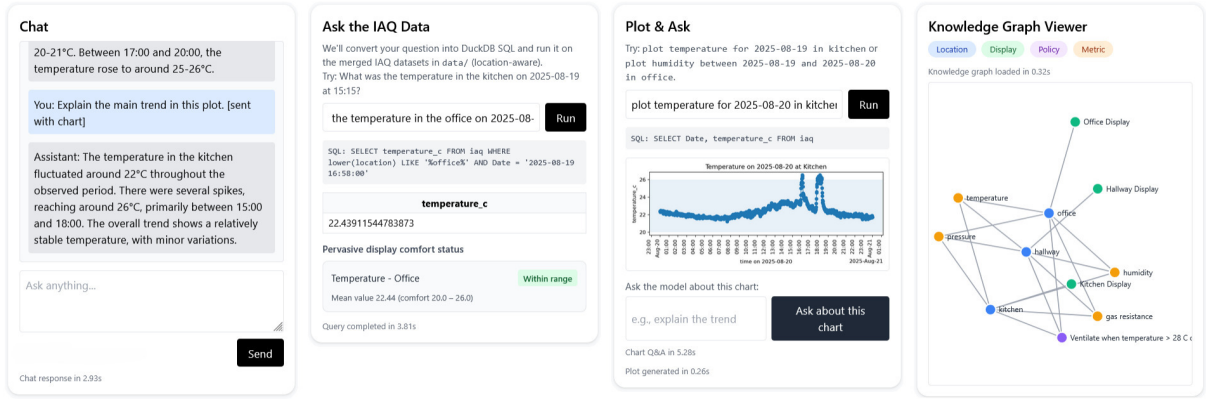


Fig. 4. Conversational interface supporting multimodal input with automatic Knowledge Graph context injection and transparency metadata.

B. Accuracy Measurement Protocol

We used task-specific accuracy criteria. **Natural-language-to-SQL.** Correct if the model produced a valid `SELECT` that passed the safety gate, executed without error, and returned the dataset-computed expected result. **Display cards.** Correct if the output matched a rule-based oracle from the data and KG thresholds, in both the numerical aggregate and the state label (*ok*, *warn-low*, *warn-high*). **Plot generation.** Correct if the system extracted the intended metric, location, and date range and rendered the expected plot with the correct KG comfort band when applicable. **MLLM interpretation.** Responses were scored against a researcher-defined rubric assessing whether they identified the main trend, recognized threshold crossings or anomalies visible in the plot, and avoided fabricated numerical claims unsupported by the visual input. A response was correct only if it met all three; e.g., “temperature rises through the afternoon and briefly exceeds the upper comfort limit” was accepted when consistent with the plot, whereas a fluent but unsupported claim (a peak value absent from the chart) was rejected. A single researcher graded all responses against this rubric. For all workflows, we measured end-to-end latency from user request to output.

C. KG-Grounding Ablation

To isolate the contribution of the KG policy layer, plot-interpretation responses were evaluated under two conditions using a paired within-model design. In the **KG-On** condition, KG-derived recommended ranges were injected into the prompt and rendered as comfort-band overlays on the generated plot. In the **KG-Off** condition, KG access was disabled via a backend flag, removing both the textual threshold context and the visual overlay, while the model received the same chart and question and had to infer the safe range from parametric knowledge alone. The ablation used temperature and humidity as the evaluated metrics, since the KG defines explicit recommended ranges for these two variables (temperature: 20–26 °C; humidity: 30–60%). Cases were constructed across three locations (hallway, kitchen, office) and two dates, yielding 12 paired cases per condition

per model. Each case passed the same natural-language plot request and the following fixed safety-limit decision prompt: “Based on this day’s [metric] observations, did the readings stay within the safe limit throughout the entire day? Answer with within safe limit or outside safe limit, then briefly explain using the plot.” An example model response is “within safe limit. The chart shows temperature readings consistently between 20.0°C and 26.0°C (the recommended safe range per KG facts), with no data points exceeding 26.0°C or dropping below 20.0°C.” Ground truth was computed directly from the underlying sensor measurements by comparing the observed daily minimum and maximum against the KG-defined recommended range. Responses were graded as correct only when both the binary safe-limit decision and the supporting explanation were consistent with the plotted observations and the ground truth.

V. RESULTS AND DISCUSSION

Table II summarizes the accuracy and end-to-end latency of all models across representative queries, including aggregates, date/location/time lookups, and one- or two-day plots.

Natural Language Data Analysis: We evaluated the natural-language-to-SQL workflow with 30 queries such as “What is the average temperature of hallway?”. All models reached 100% accuracy. Granite 3 2B was fastest (2.66 s), followed by Ministral 3 3B (2.99 s) and Gemma 3 4B (3.35 s), while Qwen 3 4B was notably slower (12.66 s). Under the proposed schema, local MLLMs reliably bridge natural language and structured sensor retrieval.

Pervasive Display Card Generation: Display cards are generated from structured queries and show the state of the selected measurement. Using 30 date-location-time queries such as “What is the temperature in the kitchen on 2025-08-20 at 15:15?”, Granite 3 2B, Gemma 3 4B, and Ministral 3 3B achieved 100% accuracy at 2.90 s, 3.80 s, and 3.19 s, respectively, while Qwen 3 4B reached 96.67% at 61.18 s.

Visual IoT Analytics with MLLM: We evaluated plot-and-ask in two stages: plot generation and vision-enabled interpretation. With 30 requests such as “plot temperature

TABLE II
PERFORMANCE EVALUATION ACROSS INTERACTION WORKFLOWS USING FOUR MULTIMODAL LLM MODELS.

Workflow	Queries	Gemma 3 4B		Ministral 3 3B		Qwen 3 4B		Granite 3 2B	
		Lat. (s)	Acc. (%)	Lat. (s)	Acc. (%)	Lat. (s)	Acc. (%)	Lat. (s)	Acc. (%)
NL to SQL	30	3.35	100.0	2.99	100.0	12.66	100.0	2.66	100.0
Display Cards	30	3.80	100.0	3.19	100.0	61.18	96.67	2.90	100.0
Plot Generation	30	0.25	100.0	0.29	100.0	0.26	100.0	0.24	100.0
MLLM Interpretation	30	5.10	96.67	5.55	100.0	32.46	90.00	6.04	63.33
Plot-and-Ask	30	5.34	96.67	5.84	100.0	32.72	90.00	6.27	63.33

for 2025-08-19 in kitchen”, all four models achieved 100% plot-generation accuracy at low latency (0.24–0.29 s). For interpretation, we used 30 prompts such as “explain the trend” and “do you see any anomaly?”. Ministral 3 3B reached 100% accuracy (5.55 s) and Gemma 3 4B 96.67% (5.10 s), while Qwen 3 4B reached 90% (32.46 s) and Granite 3 2B 63.33% (6.04 s). End-to-end plot-and-ask latency ranged from 5.34 s (Gemma 3 4B) to 32.72 s (Qwen 3 4B).

Knowledge-Graph Ablation Experiment: Table III reports plot-interpretation accuracy under KG-On and KG-Off conditions. These ablation scores are not directly comparable to the plot-and-ask workflow: the ablation task requires a precise binary safe-limit decision with a threshold-grounded explanation, whereas the main interpretation task evaluates trend identification and anomaly detection against a broader rubric. The results reveal model-dependent sensitivity to KG grounding. Qwen 3 4B showed the largest accuracy gain, +25.0 percentage points (pp), rising from 50.0% to 75.0%, and was the only model whose KG-On latency decreased (27.98 s to 15.21 s), consistent with policy context reducing its tendency toward verbose or exploratory outputs. Ministral 3 3B improved by 16.7 pp (25.0% to 41.7%), indicating that without grounding it frequently fabricated threshold values absent from the plot. Gemma 3 4B showed a modest accuracy decrease (−8.3 pp), which we attribute to the additional prompt length introduced by KG context occasionally shifting its focus from trend analysis toward threshold commentary. Granite 3 2B was unaffected in either direction, consistent with its lower baseline visual reasoning (63.33% plot-interpretation accuracy). Latency effects were likewise mixed: KG-On increased latency for Gemma (5.32 s to 6.96 s), Ministral (3.87 s to 4.99 s), and Granite (7.52 s to 8.55 s), decreasing only for Qwen. The effect of KG grounding is therefore model-dependent: it improved accuracy for Qwen 3 4B and Ministral 3 3B, slightly reduced it for Gemma 3 4B, left Granite 3 2B unchanged, and reduced latency only for Qwen 3 4B. These patterns point to task- and model-aware deployment, enabling grounding for stronger interpreters such as Qwen 3 4B and Ministral 3 3B while applying lighter context injection for Gemma 3 4B.

Indoor Safety Classification: Beyond plot interpretation, the same local models can interpret indoor scene images directly. As a supplementary capability check, we evaluated indoor safety classification on a balanced set of 120 images—60 *safe* scenes from the Indoor Scene Recognition dataset [22] and 60 *unsafe* scenes from the Indoor Fire Smoke

TABLE III
KG-GROUNDING ABLATION FOR PLOT INTERPRETATION ACROSS FOUR MLLMs (Δ = KG-ON MINUS KG-OFF ACCURACY).

Model	Condition	Acc. (%)	Lat. (s)	Δ Acc. (pp)
Gemma 3 4B	KG-Off	75.0	5.32	−8.3
	KG-On	66.7	6.96	
Ministral 3 3B	KG-Off	25.0	3.87	+16.7
	KG-On	41.7	4.99	
Qwen 3 4B	KG-Off	50.0	27.98	+25.0
	KG-On	75.0	15.21	
Granite 3 2B	KG-Off	33.3	7.52	0.0
	KG-On	33.3	8.55	



Fig. 5. Example indoor scenes from the safety classification task, with safe and unsafe cases presented to the local models.

dataset [23]—using the fixed prompt “Is this place safe? Answer with exactly one word: SAFE or UNSAFE” (Fig. 5). As reported in Table IV, Qwen 3 4B attained the highest overall accuracy (90.83% at 5.63 s mean latency), followed by Gemma 3 4B (87.50%), while Granite 3 2B and Ministral 3 3B showed complementary class biases—Granite toward *safe* and Ministral toward *unsafe* scenes. Lightweight local MLLMs can thus perform non-trivial scene-level classification without cloud support, though accuracy varies by model family.

Resilience: Because each interpretation is anchored in multiple cross-checkable sources—time-series data, the rendered plot, and KG-derived thresholds—a misleading or adversarial signal on any single channel is less likely to drive the final answer than in a single-modality pipeline [24].

Limitations and Future Work: The compact 2–4B models trade some task fidelity for low latency relative to larger cloud models, and the current evaluation uses 30 queries per workflow drawn from common analytic patterns. Extending it to ambiguous, multi-day, and out-of-domain requests, comparing plot-and-ask against simpler tabular or rule-based baselines,

TABLE IV
INDOOR SAFETY CLASSIFICATION PERFORMANCE ACROSS MLLMs.

Metric		Gemma 3 4B	Ministral 3 3B	Qwen 3 4B	Granite 3 2B
Accuracy	Overall (%)	87.50	67.50	90.83	74.17
	Safe (%)	80.00	35.00	98.33	100.00
	Unsafe (%)	95.00	100.00	83.33	48.33
Latency	Mean (s)	4.00	2.96	5.63	4.56
	Median (s)	3.99	2.71	4.40	4.97
	P95 (s)	4.11	4.31	10.51	7.58

and supporting streaming ingestion such as Message Queuing Telemetry Transport (MQTT) and larger knowledge graphs beyond the current CSV are natural next steps.

VI. CONCLUSION

This paper presented a privacy-preserving architecture for visual IoT analytics that integrates locally deployed multimodal LLMs with a semantic Knowledge Graph. Through transparent SQL/SPARQL synthesis, KG-grounded prompt augmentation, and automated visualization with comfort-threshold overlays, it enables natural-language interaction with sensor data without external dependencies. Across a two-week indoor air quality deployment, all four evaluated local models achieved 100% accuracy for structured query synthesis and 100% accuracy for plot generation. For plot interpretation, Ministral 3 3B achieved 100% accuracy with 5.84s end-to-end plot-and-ask latency, while Gemma 3 4B provided a strong balance at 96.67% accuracy and 5.34s latency. The KG-grounding ablation revealed model-dependent effects: policy-layer context improved interpretation accuracy for Qwen 3 4B (by up to 25.0 pp) and Ministral 3 3B and lowered Qwen's latency. These results indicate that local multimodal analytics can support interactive indoor sensemaking when a plot-based representation is paired with knowledge-graph grounding.

ACKNOWLEDGMENTS

The research was supported by the Ministry of National Education of the Republic of Türkiye, the Jane and Aatos Erkko Foundation EVIL-AI project, and the Research Council of Finland with grant number 362594.

The authors used generative AI tool (Claude) for language editing and manuscript refinement. We are grateful to the reviewers for their constructive comments.

REFERENCES

- [1] N. H. Motlagh, M. A. Zaidan, L. Lovén, P. L. Fung, T. Hänninen, R. Morabito, P. Nurmi, and S. Tarkoma, "Digital twins for smart spaces—beyond IoT analytics," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 573–583, 2023.
- [2] A. Varol, N. H. Motlagh, M. Leino, S. Tarkoma, and J. Virkki, "Creation of AI-driven smart spaces for enhanced indoor environments—a survey," *Internet of Things*, p. 101876, 2026.
- [3] M. Tawfik, A. H. Abdelhalim, and I. S. Fathi, "Transforming IoT security through large language models: A comprehensive systematic review and future directions," *Statistics, Optimization & Information Computing*, vol. 14, no. 2, pp. 1018–1044, 2025.
- [4] A. Sangwan, P. Sen, A. Singh, S. Zaman, Y. Zhou, and M. Medley, "From cloud to edge: Enabling offline IoT with small-scale language models," *IEEE Internet of Things Magazine*, 2025.
- [5] A. Varol, N. H. Motlagh, M. Leino, and J. Virkki, "Performance of large language models across edge and cloud platforms in smart spaces," in *2025 10th International Conference on Smart and Sustainable Technologies (SpliTech)*. IEEE, 2025, pp. 1–6.
- [6] K. Mohsenzadegan, V. Tavakkoli, W. V. Kambale, and K. Kyamakya, "A hybrid AI framework integrating ontology learning, knowledge graphs, and large language models for improved data model translation in smart manufacturing and transportation," in *2024 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2024, pp. 1–6.
- [7] K. Mohsenzadegan, V. Tavakkoli, W. V. Kambale, and K. Kyamakya, "Towards seamless data translation based on data models: A hybrid AI framework for smart transportation and manufacturing," in *2024 19th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)*. IEEE, 2024, pp. 68–73.
- [8] C. Zhou and J. Yang, "HoloLLM: Multisensory foundation model for language-grounded human sensing and reasoning," *arXiv preprint arXiv:2505.17645*, 2025.
- [9] P. P. Ray and M. P. Pradhan, "LLMEdge: A novel framework for localized LLM inferencing at resource-constrained edge," in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*. IEEE, 2024, pp. 1–8.
- [10] M. Zhang, X. Shen, J. Cao, Z. Cui, and S. Jiang, "EdgeShard: Efficient LLM inference via collaborative edge computing," *IEEE Internet of Things Journal*, vol. 12, no. 10, pp. 13 119–13 131, 2024.
- [11] A. Varol, K. Kołodziej, Łukasz Sobczak, M. Romaszewski, P. Głomb, N. H. Motlagh, M. Leino, and J. Virkki, "Enabling cloud-level accuracy in edge ai through iot data preprocessing," 2026. [Online]. Available: <https://arxiv.org/abs/2606.22496>
- [12] G. M. Yilma, J. A. Ayala-Romero, A. Garcia-Saavedra, and X. Costa-Perez, "TelecomRAG: Taming telecom standards with retrieval-augmented generation and LLMs," *ACM SIGCOMM Computer Communication Review*, vol. 54, no. 3, pp. 18–23, 2025.
- [13] M. Klapperstueck, T. Czaundera, C. Goncu, J. Glowacki, T. Dwyer, F. Schreiber, and K. Marriott, "Contextuwall: Multi-site collaboration using display walls," *Journal of Visual Languages & Computing*, vol. 46, pp. 35–42, 2018.
- [14] M. K. Konkel, B. Ullmer, O. Shaer, and A. Mazalek, "Envisioning tangibles and display-rich interfaces for co-located and distributed genomics collaborations," in *Proceedings of the 8th ACM International Symposium on Pervasive Displays*, 2019, pp. 1–8.
- [15] S. K. Badam, E. Fisher, and N. Elmqvist, "Munin: A peer-to-peer middleware for ubiquitous analytics and visualization spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 2, pp. 215–228, 2014.
- [16] Gemma Team, "Gemma 3," Technical report, Google DeepMind, 2025, <https://goo.gle/Gemma3Report>.
- [17] A. H. Liu, K. Khandelwal *et al.*, "Ministral 3," 2026. [Online]. Available: <https://arxiv.org/abs/2601.08584>
- [18] S. Bai, Y. Cai *et al.*, "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.
- [19] IBM Granite Team, "Granite 3.0 language models," October 2024. [Online]. Available: <https://github.com/ibm-granite/granite-3.0-language-models/>
- [20] C. Tian, X. Qin, K. Tam, L. Li, Z. Wang, Y. Zhao, M. Zhang, and C. Xu, "CLONE: Customizing LLMs for efficient latency-aware inference at the edge," *arXiv preprint arXiv:2506.02847*, 2025.
- [21] S. S. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, and J. W. Vaughan, "'i'm not sure, but...': Examining the impact of large language models' uncertainty expression on user reliance and trust," in *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, Rio de Janeiro, Brazil, 2024, pp. 822–835.
- [22] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 413–420.
- [23] A. K. Putra, "Indoor fire smoke dataset," Mar. 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.15826133>
- [24] A. Shaikh, A. Varol, and J. Virkki, "From prompts to motors: Man-in-the-middle attacks on llm-enabled vacuum robots," *IEEE Access*, 2025.