Intelligence of Things: A spatial context-aware control system for smart devices (#119024)

First submission

Guidance from your Editor

Please submit by **8 Sep 2025** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

All review materials are strictly confidential. Uploading the manuscript to third-party tools such as Large Language Models is not allowed.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

14 Figure file(s)
4 Table file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

- 1. Basic Reporting
- 2. Study design
- 3. Validity of the findings
- 4. General Comments
- 5. Confidential notes to the editor

- You can also annotate the review pdf and upload it as part of your review (optional).

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

Article types: Research and AI Application

BASIC REPORTING

Include the appropriate criteria template based on the type variable
Clear and unambiguous, professional English used throughout.

The article must be written in English and must use clear, unambiguous, technically correct text. The article must conform to professional standards of courtesy and expression.

Literature references, sufficient field background/context provided.

The article should include sufficient introduction and background to demonstrate how the work fits into the broader field of knowledge. Relevant prior literature should be appropriately referenced.

Professional article structure, figures, tables. Raw data shared.

The structure of the article should conform to an acceptable format of ‘standard sections’ (see our Instructions for Authors for our suggested format). Significant departures in structure should be made only if they significantly improve clarity or conform to a discipline-specific custom.

Figures should be relevant to the content of the article, of sufficient resolution, and appropriately described and labeled.

All appropriate raw data have been made available in accordance with our Data Sharing policy.

Self-contained with relevant results to hypotheses.

The submission should be ‘self-contained,’ should represent an appropriate ‘unit of publication’, and should include all results relevant to the hypothesis.

Coherent bodies of work should not be inappropriately subdivided merely to increase publication count.

Formal results should include clear definitions of all terms and theorems, and detailed proofs.

EXPERIMENTAL DESIGN

Original primary research within Aims and Scope of the journal.
Research question well defined, relevant & meaningful. It is stated how research fills an identified knowledge gap.

The submission should clearly define the research question, which must be relevant and meaningful. The knowledge gap being investigated should be identified, and statements should be made as to how the study contributes to filling that gap.

Rigorous investigation performed to a high technical & ethical standard.

The investigation must have been conducted rigorously and to a high technical standard. The research must have been conducted in conformity with the prevailing ethical standards in the field.

Methods described with sufficient detail & information to replicate.

Methods should be described with sufficient information to be reproducible by another investigator.

VALIDITY OF THE FINDINGS

Impact and novelty not assessed. Meaningful replication encouraged where rationale & benefit to literature is clearly stated.

Decisions are not made based on any subjective determination of impact, degree of advance, novelty or being of interest to only a niche audience. We will also consider studies with null findings. Replication studies will be considered provided the rationale for the replication, and how it adds value to the literature, is clearly described. Please note that studies that are redundant or derivative of existing work will not be considered. Examples of "acceptable" replication may include software validation and verification, i.e. comparisons of performance, efficiency, accuracy or computational resource usage.

All underlying data have been provided; they are robust, statistically sound, & controlled.

The data on which the conclusions are based must be provided or made available in an acceptable discipline-specific repository. The data should be robust, statistically sound, and controlled.

Conclusions are well stated, linked to original research question & limited to supporting results.

The conclusions should be appropriately stated, should be connected to the original question investigated, and should be limited to those supported by the results. In particular, claims of a causative relationship should be supported by a well-controlled experimental intervention. Correlation is not causation.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Intelligence of Things: A spatial context-aware control system for smart devices

Sukanth Kalivarathan [1] , **Muhmmad Abrar Raja Mohamed** [1] , **Aswathy Ravikumar** Corresp., [2] , **Harini Sriraman** Corresp. [1]

[1] Vellore Institute of Technology University, Chennai, India

[2] Senior Consultant - AI, Sustainable Living Lab, Chennai, India

Corresponding Authors: Aswathy Ravikumar, Harini Sriraman
Email address: aswathyravi2290@gmail.com, harini.s@vit.ac.in

Background. The swift advancement of Internet of Things (IoT) technology has revolutionized smart home settings; the prevalent automation systems are limited by their need on specific device identification and rigid rule-based configurations. These constraints impede natural human-device interaction, especially in dynamic or communal environments where spatial context is more instinctive than predetermined naming conventions. Current solutions frequently neglect spatial reasoning and multimodal inputs, resulting in heightened cognitive demands and diminished accessibility. The proposedwork develops a spatial context-aware control system aimed at facilitating intuitive, vision-driven, and language-based interaction with smart devices to overcome these problems.

Methods. The proposed model modular, multimodal framework that integrates computer vision, natural language processing, and spatial inference for context-aware smart device control. The system comprises six core components: (i) an Onboarding Inference Engine for extracting device information via natural language input, (ii) Zero-Shot Device Detection using OWL-ViT for object identification without prior training, (iii) Metadata Refinement and Filtering for structured annotation and disambiguation, (iv) a Geospatial Device Visualizer for annotated visual feedback, (v) Spatial Topology Inference using GPT-4o for reasoning about physical layouts, and (vi) Intent-Based Command Synthesis with Gemini Flash to generate precise, executable control commands. The final Agentic Execution Module interfaces with the Tuya Smart Device API, ensuring vendor-agnostic actuation. The system supports multilingual input and adapts to various environmental contexts including smart homes and assisted living facilities.

Results. A user study involving 15 participants (aged 18–80, diverse educational backgrounds) was conducted to evaluate the effectiveness of proposed method in comparison to the Google Home Assistant. Quantitative findings demonstrate a statistically significant reduction in cognitive workload, with NASA Task Load Index (TLX) scores decreasing by an average of 13.17 points (p = 0.0013, Cohen’s d = 1.0381). Participants rated the proposed method higher in terms of ease of use (mean = 4.67) compared to Google Home (mean = 3.8) on a 5-point Likert scale. Qualitative feedback highlighted the intuitive nature of spatial context commands, reduced cognitive burden due to elimination of device name memorization, and enhanced accessibility via support for regional languages. 93.3% of users preferred the proposed method over the baseline system. These results affirm the feasibility and user-centric benefits of integrating vision-language models for context-aware smart device control.

Intelligence of Things: A Spatial Context-Aware Control System for Smart Devices

Sukanth Kalivarathan[1], Muhmmad Abrar Raja Mohamed[1], Aswathy Ravikumar[2], Harini Sriraman[1]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India-600127

[2]AI/ML Consultant, Sustainable Living Lab, India

Corresponding Author:

Harini Sriraman
Kelambakkam - Vandalur Rd, Rajan Nagar, Chennai, Tamil Nadu 600127

Email address: harini.s@vit.ac.in

Intelligence of Things: A Spatial Context-Aware Control System for Smart Devices

Sukanth Kalivarathan[1], Muhmmad Abrar Raja Mohamed[1], Aswathy Ravikumar[2], Harini Sriraman[1]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India-600127

[2]AI/ML Consultant, Sustainable Living Lab, India

Corresponding Author:
Harini Sriraman
Kelambakkam - Vandalur Rd, Rajan Nagar, Chennai, Tamil Nadu 600127
Email address: harini.s@vit.ac.in

Abstract

Background. The swift advancement of Internet of Things (IoT) technology has revolutionized smart home settings; the prevalent automation systems are limited by their need on specific device identification and rigid rule-based configurations. These constraints impede natural human-device interaction, especially in dynamic or communal environments where spatial context is more instinctive than predetermined naming conventions. Current solutions frequently neglect spatial reasoning and multimodal inputs, resulting in heightened cognitive demands and diminished accessibility. The proposed work develops a spatial context-aware control system aimed at facilitating intuitive, vision-driven, and language-based interaction with smart devices to overcome these problems.

Methods. The proposed model modular, multimodal framework that integrates computer vision, natural language processing, and spatial inference for context-aware smart device control. The system comprises six core components: (i) an Onboarding Inference Engine for extracting device information via natural language input, (ii) Zero-Shot Device Detection using OWL-ViT for object identification without prior training, (iii) Metadata Refinement and Filtering for structured annotation and disambiguation, (iv) a Geospatial Device Visualizer for annotated visual feedback, (v) Spatial Topology Inference using GPT-4o for reasoning about physical layouts, and (vi) Intent-Based Command Synthesis with Gemini Flash to generate precise, executable control commands. The final Agentic Execution Module interfaces with the Tuya Smart Device API, ensuring vendor-agnostic actuation. The system supports multilingual input and adapts to various environmental contexts including smart homes and assisted living facilities.

Results. A user study involving 15 participants (aged 18–80, diverse educational backgrounds) was conducted to evaluate the effectiveness of proposed method in comparison to the Google Home Assistant. Quantitative findings demonstrate a statistically significant reduction in cognitive workload, with NASA Task Load Index (TLX) scores decreasing by an average of 13.17 points (p

= 0.0013, Cohen's d = 1.0381). Participants rated the proposed method higher in terms of ease of use (mean = 4.67) compared to Google Home (mean = 3.8) on a 5-point Likert scale. Qualitative feedback highlighted the intuitive nature of spatial context commands, reduced cognitive burden due to elimination of device name memorization, and enhanced accessibility via support for regional languages. 93.3% of users preferred the proposed method over the baseline system. These results affirm the feasibility and user-centric benefits of integrating vision-language models for context-aware smart device control.

Introduction

The widespread adoption of smart devices, particularly in residential settings, has fueled growing interest in intelligent home automation systems. These devices, ranging from basic appliances like lights and fans to more complex systems, are now embedded into daily life, enhancing convenience, energy efficiency, and overall quality of living. Despite advancements in IoT infrastructure, NLP, and computer vision, the most commercially available smart home solutions continue to depend on non-intuitive interfaces. Users are often required to issue rigid, explicitly formatted commands or remember device-specific names, which hampers seamless interaction especially in dynamic or shared spaces such as hotels, restaurants, and assisted living facilities.

The rapid advancement of technology has led to the emergence of spatial context-aware control systems, which represent a significant evolution in the domain of smart devices. These systems use the principles of automation and IoT to enhance device interactions and decision-making processes by interpreting spatial data. As smart home technologies become increasingly prevalent, the need for systems that can adapt to varying contexts and user needs has never been more important.

Spatial reasoning is a fundamental aspect of human cognition, enabling individuals to navigate and interact with their environments using spatial references. Integrating this capability into IoT control systems can significantly enhance their intelligence and usability. Systems that support spatial awareness can interpret user intent based on environmental context, allowing natural, indirect references to devices without the need for explicit identifiers. This is especially valuable in scenarios such as elder care and assistive living, where users may face cognitive or physical challenges. For instance, a resident with dementia may not recall the name of a device but can still refer to it using spatial cues. Embedding spatial awareness into smart home frameworks allows for adaptive, human-like understanding, making technology more accessible and inclusive.

Conventional IoT control systems present several key limitations. They typically rely on static rule-based frameworks, predefined device identifiers (e.g., names or UUIDs), and voice commands without accounting for spatial context or user familiarity. This becomes especially problematic in multi-user environments or setups with multiple identical devices. The absence of spatial awareness restricts users from issuing natural, indirect commands such as "turn on the light near the window" or "switch on the fan beside the table," making interactions more cognitively demanding and less intuitive.

To address these challenges, this work proposes a novel spatial context-aware control framework for smart environments that combines computer vision, natural language understanding, and real-time IoT actuation. The system begins with a one-time onboarding process that uses zero-shot object detection to visually identify and annotate devices in a scene. Users confirm these annotations, which are then used to infer spatial relationships and environmental layout. This spatial model allows the system to interpret natural language commands based on contextual cues enabling users to interact using spatial references rather than explicit device names.

The proposed framework is platform-agnostic and can be integrated into diverse IoT ecosystems without requiring hardware modifications. A user study involving participants from varied demographic and educational backgrounds demonstrated the system's superiority over existing solutions. Users reported significantly improved experience, citing easier interaction, reduced memorization, and better accessibility including support for indirect spatial references and regional languages. Quantitative analysis, including NASA-TLX scores, revealed a substantial reduction in cognitive workload during task execution. These findings underscore the potential of spatially aware interaction models to enhance the usability, inclusivity, and intelligence of smart home systems, paving the way for more adaptive and user-friendly automation technologies.

Background

Spatial context-aware systems signify a transformative leap in technology, enabling smart devices to dynamically adapt their operations by interpreting environmental and spatial data. This capability enhances IoT device performance, ensuring optimal functionality in diverse settings such as smart homes and industrial environments. The IoT forms the backbone of these systems, comprising interconnected devices that facilitate efficient data-driven operations essential in today's technology landscape (Baby, 2014; Harini & Ravikumar, 2020; Shi et al., 2022; Ravikumar & Sriraman, 2023a). Control systems within spatial context-aware environments use adaptive algorithms and predefined rules to manage device operations, ensuring stability and optimal resource utilization. These systems are pivotal in precise data handling, particularly in synchronizing communication parameters (Shi et al., 2022). Smart devices equipped with advanced sensors, actuators, and communication capabilities enable important functionalities like IoT traffic analysis and fault detection, maintaining network integrity and performance. The integration of these components fosters real-time data processing, enhancing system efficacy.
Emerging technologies like non-orthogonal multiple access (NOMA) and virtual multiple input Multiple-output (MIMO) are important for understanding spatial context-aware systems, facilitating efficient resource allocation and adaptability (Shi et al., 2022). Sensor networks, consisting of distributed nodes with sensors, are vital for environmental data monitoring and

informed decision-making. Energy harvesting IoT (EH-IoT) technologies offer sustainable energy solutions by autonomously harnessing energy from environmental sources, reducing reliance on traditional batteries and addressing maintenance challenges in battery-operated IoT infrastructures. Advancements in EH-IoT include efficient energy harvesting methods, wireless power transfer systems, and innovative communication techniques optimizing power management under unpredictable conditions (Ma et al., 2020; Schulthess et al., 2022)

Smart home technology exemplifies IoT and sensor network integration in residential environments, aiming to improve user convenience, security, and energy efficiency. Recent research highlights sophisticated behavior-modeling methods for detecting irregularities in user interactions, leveraging real-time data from home IoT sensors for accurate security threat detection (Yamauchi et al., 2021). Spatial context aware systems automate household operations based on contextual data, streamlining tasks. Given communication vulnerabilities in smart home systems, secure mutual authentication protocols are important for user safety and data integrity.

Context-aware computing uses contextual information like location, time, and user activity to deliver tailored services, foundational for spatial context-aware systems' operations. Concepts such as deep learning models, microcontroller units, memory management, and segment-level control optimize communication costs and enhance system efficiency (Zheng et al., 2024). The integration of multivariate IoT data streams, event detection, and event correlation is pivotal for these systems, highlighting the interplay between technological components and their collective impact on spatial context-aware operations.

The evolution of spatial context-aware control systems in smart homes is marked by advancements in IoT, edge computing, and sensor networks. IoT technologies have progressed from basic RFID systems to complex interconnected platforms, addressing efficiency and reliability demands in smart homes (Shi et al., 2022). This transition from centralized cloud-based systems to decentralized Internet of Federated Things (IoFT) systems offers enhanced scalability and reduced latency (Kontar et al., 2021). Voice-controlled devices have transformed user interactions, enhanced convenience but introducing vulnerabilities like spoofing attacks (Baumann et al., 2019). This duality necessitates robust security measures, evolving from central authority reliance to more cost-efficient solutions (Dang & Tran, 2019). Wearable technologies integrated with IoT revolutionize personalized healthcare, providing context aware insights for individualized treatment plans (Khan & Alam, 2020). Efficient data management techniques, like data stream processing and complex event processing, have evolved to meet IoT-driven application demands (Qin et al., 2014). Fog computing solutions enhance data management efficiency by distributing computing resources closer to the data source, mitigating latency and improving responsiveness in smart homes (Mihai et al., 2019). The technological evolution in spatial context-aware control systems reflects efforts to control new technologies for enhanced efficiency, security, and user satisfaction. Innovative applications and scalable system architectures have the potential to significantly enhance home automation, safety, and energy efficiency. With projections of 50 billion interconnected devices in the next 5 to 10 years, advanced integration methods like the Context-Aware Dynamic Discovery of Things (CADDOT) model will be important for seamless communication between diverse sensor technologies and cloud-based IoT platforms, transforming interactions with living spaces (Maghsoudi et al., 2023).

Spatial context-aware control systems signify a transformative leap in the realm of smart devices, leveraging automation and IoT to enable adaptive management of device operations through spatial data interpretation. These systems exploit the advanced computational capabilities of Edge IoT devices, which have transitioned from basic low-power units to sophisticated configurations equipped with FPGAs and AI accelerators, thus facilitating real-time data processing and intelligent decision making. The integration of ML within these systems introduces novel security challenges, necessitating a thorough understanding of potential threats to ensure secure and reliable operations (Liu et al., 2024).

Important components of spatial context-aware systems are their ability to prioritize data based on the UoI, a context-driven metric that evaluates the nonlinear significance of status information, thereby optimizing decision-making processes(Zheng, Zhou & Niu, 2020). This capability is particularly essential in environments such as smart factories, where the construction of digital shop floor representations demands the seamless integration of heterogeneous production modules (Bader & Maleshkova, 2019). Prioritization of information not only enhances operational efficiency but also ensures that important data is addressed promptly, thereby reducing the risk of system failures.

The deployment of spatial context-aware systems effectively addresses latency and data processing challenges inherent in cloud-centric IoT applications, especially given the geographically distributed nature of IoT data (AlMahamid, Lutfiyya & Grolinger, 2022). These systems play a pivotal role in the integration of wearable technology with IoT, enhancing personalized healthcare by providing context-aware insights vital for individualized treatment plans (Khan & Alam, 2020). This integration exemplifies how spatial context-aware systems can bridge the gap between physical and digital realms, fostering a more interconnected and responsive environment. Furthermore, spatial context-aware systems enable innovative applications such as WiFi-based crowd monitoring, utilizing existing infrastructure to conduct real-time monitoring and predictive analysis of crowd dynamics, demonstrating the versatility and applicability of these systems across diverse domains (Mu, 2020). The emerging paradigm of Wireless Information and Energy Transfer (WIET) further exemplifies the dual functionality of these systems, amalgamating data communication with wireless charging capabilities, particularly in the context of 6G networks (Psomas et al., 2024). Such advancements not only optimize resource utilization but also pave the way for more sustainable smart environments.

The efficiency of spatial context-aware systems is augmented by the ability to deploy deep learning models on microcontroller units with significantly constrained memory, emphasizing the necessity for effective memory management. Moreover, these systems facilitate precise IoT device identification from physical layer signals without relying on conventional cryptographic methods, underscoring their importance in maintaining secure and efficient IoT ecosystems (Liu et al., 2021b). The combination of these features positions spatial context-aware systems as essential components in the evolution of smart technologies(Ravikumar, Saritha & Chandra, 2013; S & Ravikumar, 2015; Ravikumar & Sriraman, 2023a,b).

Spatial context-aware control systems create a comprehensive framework that significantly improves the functionality, efficiency, and security of smart devices. By intelligently leveraging spatial data and incorporating advanced technological solutions, these systems enable dynamic discovery and configuration of IoT, allowing for seamless integration and communication among heterogeneous devices. Furthermore, they enhance real-time monitoring and control capabilities, utilizing innovative metrics such as UoI to prioritize timely status updates based on contextual relevance. This integrated approach not only optimizes energy consumption and extends the operational lifespan of devices but also facilitates collaborative intelligence and in-sensor analytics, ultimately leading to more effective and sustainable smart environments. The ongoing evolution of smart home technologies and other IoT applications underscores the indispensable role of these systems, offering a robust platform for future innovations (Chatterjee et al., 2020)

Automation and IoT

Automation and IoT are important to the development and functionality of smart home technology, serving as foundational elements that enable seamless device integration, efficient data processing, and real-time responsiveness. The interconnected nature of IoT systems, which includes devices, sensors, and actuators communicating over the internet, facilitates dynamic interactions within household environments (Masuduzzaman et al., 2019). This connectivity is essential for smart home devices to achieve common goals, such as enhanced user experience and operational efficiency. As the landscape of smart homes continues to evolve, the importance of automation and IoT becomes increasingly pronounced, necessitating a closer examination of their roles and impacts. The rapid proliferation of IoT technologies has led to the emergence of numerous platforms, presenting challenges for organizations in selecting the most appropriate solutions for their specific needs (Ullah et al., 2020). These challenges are further compounded by the necessity to maintain the integrity of computations performed in edge computing environments, where automation plays a vital role in verifying outsourced computations (Ullah et al., 2020). Automation is also important in optimizing IoT data management, particularly in latency-sensitive applications where cloud-based systems may struggle with inefficiencies. The integration of automation within IoT frameworks not only enhances performance but also ensures that systems can adapt to changing conditions and user preferences. In industrial settings, automation and IoT are indispensable for achieving faster conversion rates and implementing data-driven maintenance strategies, as evidenced in smart factories. The integration of IoT technology in these environments necessitates robust automation solutions to address the complexities of rapid sensor deployment in unstructured settings (Mihai et al., 2019). Moreover, the ability to analyze data in real-time enables organizations to make informed decisions that enhance operational efficiency and reduce downtime, thereby maximizing productivity. Additionally, automation and IoT are pivotal in developing personalized healthcare solutions, meeting the growing demand for patient-centric health management. By leveraging data from wearable devices and other IoT-enabled technologies, healthcare providers can offer tailored interventions that improve patient outcomes. This shift towards personalized care highlights the potential of automation and IoT to transform traditional healthcare models, fostering a more holistic approach to health management. As IoT networks continue to expand, automation remains a key factor in addressing the complexities and challenges inherent in smart home environments, ensuring efficient operation, enhanced user experience, and robust privacy and security measures. The intersection of automation and IoT not only facilitates the development of innovative solutions but also lays the groundwork for future advancements in smart home technologies.

Related Works

Context-aware computing in smart homes facilitates intelligent decision-making and interaction by leveraging IoT devices to create responsive and personalized environments. This approach enables smart home systems to adapt to user preferences and environmental changes, enhancing user experience and operational efficiency. The growing prevalence of IoT devices underscores the importance of context-aware technologies, which improve user satisfaction and the effectiveness of smart home systems.

Intelligent Decision-Making and Interaction

Context-aware computing significantly enhances intelligent decision-making and interaction in smart homes by utilizing IoT data to provide personalized experiences. Mixed reality avatars, as explored by Morris et al., improve user interaction by representing IoT devices in an engaging manner, facilitating more intuitive decision-making processes (Morris et al., 2020). (Liu et al., 2021a) zero-bias deep learning enabled method enhances decision-making by using zero-bias DNNs as performance-assured abnormality detectors. (AlQahtani, Alamleh & Smadi, 2022) demonstrate effective proximity authentication for IoT devices, ensuring secure interactions within smart home environments. (Han & Huang, 2016) WP-BC network optimizes resource usage by using contextual information for energy harvesting and data transmission decisions. (Sun, Wu & Wang, 2021) improve data collection efficiency with their compressive data collection method, enhancing decision-making accuracy. (Zambonelli, 2016) software engineering methodology supports robust IoT system development, improving decision-making through structured design. Recent advancements like CADDOT, ISA, and CI further enhance decision-making and interaction by dynamically integrating IoT devices and optimizing energy consumption.

Adaptive and Predictive Systems

Adaptive and predictive systems in smart homes optimize functionality by leveraging contextual data for personalized user experiences. Edge-ICN technology enhances IoT communications by offering multicast and anycast capabilities, improving data forwarding efficiency (Fotiou et al., 2017). Predictive systems use data-driven approaches to anticipate user needs, with hybrid techniques offering superior results by addressing individual method limitations (Achiluzzi et al., 2022). These systems dynamically adjust operations based on real-time data, improving responsiveness and satisfaction. The integration of adaptive and predictive systems enhances smart home functionality by delivering customized services, optimizing energy consumption, and ensuring privacy and security (Sayed et al., 2022).

Energy Management and Efficiency

Context-aware computing enhances energy management in smart homes by optimizing resource usage and reducing consumption. The ACMCA algorithm improves reconstruction accuracy and energy efficiency, exemplifying adaptive data processing (Salehi & DeMara, 2019). (Jiang et al., 2021) hybrid mesh network offers improvements in power consumption and communication range, contributing to efficient energy management. LiPI's data aggregation strategy enhances latency and energy efficiency, outperforming existing methods (Goyal, Kodali & Saha, 2022). (Kaplan, Vieira & Larsson, 2024) reduce power requirements for signal processing through direct link interference suppression. (Homssi et al., 2020) framework captures energy consumption patterns, aiding context-aware systems in implementing optimal strategies. (Wisy, 2021) trust metric improves sensor network reliability, supporting efficient energy utilization. These advancements facilitate the creation of sustainable smart home environments by enhancing energy efficiency, user convenience, and addressing environmental concerns.

LLM-Orchestrated Flexible Smart Home Control

Recent advancements in Large Language Models (LLMs) have enabled more flexible and intuitive smart home control by allowing systems to interpret under-specified and context-dependent commands, such as “make it cozy,” without requiring explicitly named devices. For instance, approaches like IoT Smart Home (Rivkin et al., 2025) demonstrate the ability to control visual or contextual cues to generate appropriate device actions. Systems such as SAGE (Spandan & Iqbal, 2024) integrate LLMs with tools for direct device interaction, persistent monitoring, and flexible prompting, significantly outperforming standard LLM baselines in structured task benchmarks. Similarly, frameworks like Sasha (King et al., 2024) and LLM Home (King et al., 2023) emphasize the generation of action plans from vague user intents. However, these systems face limitations in disambiguating spatial references and recovering from execution failures.

While these methods exhibit promising capabilities in interpreting naturalistic commands, their spatial reasoning tends to be implicitly driven by the language model itself, lacking explicit spatial calculus or topological modeling. This highlights an important gap in formal spatial environment and interaction modeling, necessary for accurate interpretation of spatially grounded commands in dynamic or unfamiliar smart home environments.

Spatial Environment and Interaction Modeling

ProxeGraph (Spandan & Iqbal, 2024) introduces proxemics-aware scene graphs to enhance spatial modeling by incorporating non-verbal cues such as gestures and eye tracking. While its primary focus lies in improving scene understanding for HCI, it lacks direct integration with natural language parsing or downstream device actuation mechanisms.

QueSTMaps (Mehan et al., 2024) constructs semantic and topological 3D representations of environments to support spatial language queries (e.g., "a place to cook"). Although effective for robotic navigation and semantic localization, it is not designed for smart home device control or interaction resolution based on user commands.

Contextual and Goal-based Approaches

Graph-based and personalized systems (Li & Wu, 2022) use lightweight NLP and inference over contextual graphs to map user goals to specific room-level actions or automation scenarios. These systems are adept at recognizing user preferences and environmental context but typically do not handle fine-grained spatial references in natural language.

Location- and gesture-driven approaches (Mehan et al., 2024) enable users to select devices by physically pointing at them using a mobile device, estimating spatial relationships through localization techniques. However, these systems do not process or interpret spatial cues conveyed through natural language, limiting their adaptability to verbal instructions in dynamic or unfamiliar settings.

Spatial Topology Inference Methods

Spatial reasoning is an important component of AI-driven smart home automation, allowing intelligent systems to interpret and interact meaningfully with their physical environments. Although various methodologies have been proposed to improve spatial understanding ranging from scene analysis and visual question answering to industrial spatial intelligence, most existing approaches are designed for analytical purposes rather than enabling real-time automation in diverse and dynamic room settings.

ROOT, a vision-language model scene understanding system, uses an iterative object perception algorithm to detect and annotate objects within indoor environments (Wang et al., 2024a). While effective at generating structured spatial representations, its primary utility lies in static scene interpretation rather than in dynamic smart home control. Similarly, Spatial VLM (Chen et al., 2024) facilitates large-scale 3D spatial reasoning by training on Internet-scale datasets, enhancing capabilities in VQA and robotics. However, it falls short in supporting real-time adaptability and automation tasks in general-purpose home environments.

Additional progress has been made through spatial relation modeling in vision-language frameworks. These models use techniques such as object position regression and spatial relation classification to enhance visual commonsense reasoning (Yang et al., 2023). While they improve performance in structured language-vision tasks, their application in real-world automation remains limited. Industrial spatial intelligence research (Wang et al., 2024b) has focused on generating scene graphs for predefined factory environments. Despite excelling in structured and controlled settings, these approaches lack the flexibility required for adapting to dynamic and heterogeneous residential scenarios.

Recent advancements highlight the potential of LLMs in IoT applications. For example, IoT-LLM (An et al., 2024) demonstrates how LLMs can enhance task reasoning in domains such as human sensing and indoor localization. While effective in interpreting sensor data, this approach does not incorporate vision-language integration necessary for spatial disambiguation or dynamic device referencing. The SAGE framework (Rivkin et al., 2024) controls LLMs within a fixed prompt tree, utilizing pre-registered static images to resolve device ambiguity through manually updated spatial mappings. Although SAGE improves over prior LLM baselines, it still relies on static inputs and lacks adaptability to real-time spatial changes.

Device Onboarding and Management

Foundational work in IoT device onboarding has laid the groundwork for efficient detection and interaction. AIDE (Zhang et al., 2019) offers an augmented onboarding experience by leveraging received signal strength profiles to associate physical devices with their digital counterparts. However, it lacks deeper contextual awareness and does not incorporate user intent. In contrast, our system supports multi-modal inputs and enhances onboarding accuracy by integrating LLMs for structured data extraction. (Meyuhas, Bremler-Barr & Shapira, 2024) introduced a hybrid labeling strategy that combines string-matching for vendor identification with a RoBERTa-based model for functional classification. Though effective in network-based labeling, it does not use computer vision. Our approach advances this by applying computer vision to visually identify, label, and map devices in the environment.

Multimodal IoT Systems: Advancements in Spatially Aware Automation

Emerging research explores the use of LLMs like GPT-3 for contextual smart home control. (King et al., 2023) demonstrates that high-level user intents can be translated into actionable device commands. While effective in mapping textual commands, this system lacks real-time visual scene interpretation and does not incorporate spatial relationships between devices, relying solely on linguistic cues.

(Zong et al., 2025) further demonstrates the potential of LLMs in IoT ecosystems, showing that these models can interpret complex data streams, facilitate predictive maintenance, and support natural language interactions for intuitive control. Their work highlights the significance of prompt engineering and device interoperability but does not focus on real-time scene understanding or spatial adaptability.

Objectives

To address these gaps, this research introduces a spatial context-aware control system that integrates computer vision, VLMs, and agentic natural language processing to revolutionize human-IoT interaction. The main objectives of the study are:

- To develop an AI-driven architecture capable of performing spatial reasoning and natural language understanding for smart device control.

- To build a multimodal dataset and interaction pipeline that supports autonomous decision-making based on visual and linguistic cues.

- To enable indirect spatial referencing in user commands, thereby reducing cognitive load and improving the intuitiveness of device interaction.

- To enhance usability for non-technical users and individuals with accessibility needs by eliminating dependence on explicit device naming and structured commands.

Gaps Identified

Based on the comprehensive review of the literature, the identified gaps are summarized in Table 1.

Methodology

The primary goal of the proposed system as shown in Figure1 is to develop an intelligent, context-aware automation framework for smart home devices. By leveraging Vision Language Models and modular architecture, the system ensures seamless interaction, precise control, and adaptive automation with minimal human intervention. The proposed model is shown in Fig 1.

- Onboarding Inference Engine: This module serves as the initial point of user interaction, collecting information about IoT devices present in the environment. It processes natural language inputs, enabling users to provide device details effortlessly. The extracted information is converted into a structured device inventory, which forms the basis for all subsequent modules.

- Zero-Shot Device Detection: This module identifies and localizes IoT devices in each image. Using OWL-ViT it performs zero-shot object detection, enabling the system to recognize previously unseen device types. The generated metadata provides vital attributes for each detected device, essential for precise control and automation.

- Metadata Refinement and Filtering: To improve the accuracy of the system, this module processes the raw metadata generated by the detection module. It assigns unique identifiers and filters data based on user inputs and model confidence scores. This ensures that only relevant and high-confidence detections are retained for further use.

- Geospatial Device Visualizer: This component overlays bounding boxes and labels onto the input image based on the refined metadata. It provides users with an intuitive understanding of the device layout, supporting more effective automation decisions.

- Spatial Topology Inference: This module analyzes the spatial configuration of devices by inferring their positions relative to room features and other IoT devices. Contextual spatial relationships are extracted to support intelligent automation strategies, ensuring optimal device coordination within the environment.

- Intent-Based Agentic Command Synthesis: By combining spatial metadata with user intent, this module synthesizes precise control commands. It interprets real-time user instructions and environmental cues from the Spatial Topology Inference module to generate adaptive automation commands for responsive smart home interaction.

- Agentic Actuation & Execution Module: Serving as the final operational stage, this module interfaces with Tuya Smart Device API a smart home management platform. It executes the control commands generated by the system while handling validation and potential errors, ensuring smooth integration within the IoT ecosystem.

AI Models Used

- Qwen-2.5-32B: Used for onboarding and interpreting natural language descriptions of IoT devices.
- OWL-ViT (OWL2): Responsible for automatic image-based annotation through zero-shot object detection.
- GPT-4o: Extracts spatial relationships and topology from annotated device data.
- Gemini 2.0 Flash: Processes user commands.

Onboarding Inference Engine

Onboarding Inference Engine serves as the initial module in the proposed system. Its primary function is to facilitate user onboarding by collecting information regarding the number and types of IoT devices present in each environment from the user. This ensures that the system is aware of the available devices before proceeding with subsequent detection and control processes.

- Operational Mechanism: The operational mechanism of the Onboarding Inference Engine begins with user input collection, where the system prompts the user to provide details about the IoT devices present in their environment. The input can be provided in natural language and supports both text and voice modalities.

- Prompt Used for Device Extraction: To effectively extract the number and type of IoT devices from user input, a predefined prompt is used by the Onboarding Inference Engine. The prompt ensures that the system can consistently identify and quantify devices, regardless of the input format.

Prompt: You are an AI assistant responsible for onboarding users into a smart IoT control system. Your task is to extract the number and type of IoT devices mentioned by the user in natural language input.

Rules:

1) Identify the device type
2) Extract the quantity of each device.
3) Ignore unrelated information and return only the structured device data.
4) Store the output as a JSON dictionary with device types as keys and their counts as values.

Zero-Shot Device Detection:

The Zero-Shot Device Detection Module constitutes the core vision-based component for identifying IoT devices from environmental imagery without requiring prior task-specific training. Leveraging OWL-ViT (OpenAI et al., 2024), a state-of-the-art zero-shot object detection framework developed by Google, this module enables the recognition of unseen object classes based on natural language prompts. By eliminating the need for retraining, the system can detect a broad range of smart devices within diverse real-world environments. The Onboarding Inference Engine provides a structured list of smart device types. Subsequently, the Zero-Shot Detection Module transforms these textual device types into object detection prompts, applying them directly to the input scene for visual localization. The output generated by this module serves as the annotated foundation for downstream spatial reasoning and command generation modules, enabling robust interaction and control. The complete annotation workflow is shown in figure 2.

The architecture of the Zero-Shot Device Detection Module follows a deterministic, multi-stage pipeline, described as follows:

- Device List Ingestion

The module ingests a predefined list of device classes, extracted from the onboarding scenario or user-provided instruction. Each device class is transformed into a natural language prompt tailored for OWL2 inference.

- Zero-Shot Inference Using OWL2

The OWL2 model performs inference by embedding both the visual features of the input image and the text embeddings of the device class prompts.

Algorithm: Matching Process

Input: Scene Image, List of Device Class Prompts (text descriptions)

Output: Set of Detected Objects with Bounding Box, Class Label, and Confidence Score

Algorithm Steps:

Embedding Generation:

1. For each device class prompt in the list, compute its text embedding using OWL2's language encoder.
2. Compute visual embeddings for regions within the input scene image using OWL2's vision encoder.

Feature Alignment and Matching:
1. For each region in the visual embedding:
2. Compare all text embeddings corresponding to the device class prompts.
3. Measure similarity between visual region embeddings and device class text embeddings.
Detection and Output Generation:
1. If similarity score exceeds the predefined threshold:
2. Record the following for the matched region:
3. Bounding Box Coordinates: (x_1, y_1, x_2, y_2)
4. Class Label: Corresponding device type
5. Confidence Score: Similarity score between 0 and 1 indicating detection confidence.
Result Compilation:
Aggregate all detected instances into a structured output set.

- Metadata Structuring

Following detection, the raw outputs are structured into a standardized metadata format compatible with subsequent modules.

Each metadata entry includes:

- Device Type: The detected class label
- Bounding Box: The spatial coordinates (x_1, y_1, x_2, y_2) of the detected device.
- Confidence Score: The associated model confidence level.

Metadata Refinement and Filtering:

The Metadata Refinement and Filtering Module is responsible for enhancing and structuring the raw detection outputs generated by the Zero-Shot Device Detection Module. This important step ensures that only the most relevant, accurate, and consistently formatted device data are forwarded to subsequent modules. Through rigorous filtering, application of user-specific criteria, and systematic metadata structuring, this module significantly improves the reliability and usability of downstream spatial reasoning processes. Operating as a quality assurance layer, the Metadata Refinement and Filtering Module refines the preliminary detection results produced by the OWL2 model. It ensures that the device metadata passed to later stages is precise, contextually appropriate, and properly labeled. Key processes include assigning UUIDs to each detected device, filtering out irrelevant or low-confidence detections, enforcing standardized naming conventions, and prioritizing detections based on confidence scores. The structured and refined metadata output strengthens the system's spatial awareness and enhances decision-making accuracy.

Algorithm: Metadata Refinement and Filtering

Input:
DetectedDevices: List of raw device detections, each with class label, bounding box coordinates, and confidence score.
UserTargetDevices: List of device types specified by the user's onboarding or interaction input.

```
Steps:
  1. Assign Unique Identifiers (UUIDs)
       o For each device in DetectedDevices, generate and assign a Universally Unique Identifier (UUID) to maintain device consistency and traceability throughout the pipeline.
  2. Apply Structured Naming
       o Label each device using a standardized naming convention that incorporates spatial positioning and contextual attributes to improve clarity and downstream interpretability.
  3. Apply Spatial Ordering
       o Arrange device labels based on positional hierarchy:
           ▪ Horizontal Ordering: Sort devices from left to right along the horizontal axis.
           ▪ Vertical Ordering: When multiple devices align horizontally, sort them from top to bottom vertically.
       o This ensures that identical device types are uniquely distinguishable based on spatial location.
  4. Filter Devices Based on User Input
       o Initialize an empty list FilteredDevices.
       o For each device in DetectedDevices:
           ▪ If device.type matches any type in UserTargetDevices, add it to FilteredDevices.
  5. Rank and Select Devices by Confidence Score
       o For each device type in FilteredDevices:
           ▪ Sort devices in descending order based on their ConfidenceScore.
           ▪ Select the top N devices, where N matches the quantity requested by the user.
           ▪ Discard devices below a predefined confidence threshold (e.g., 0.5) to maintain output reliability.
  6. Generate Structured Metadata Output
       o Initialize an empty list MetadataOutput.
       o For each device in the selected subset:
           ▪ Create a structured metadata entry containing:
               ▪ DeviceLabel: The class/type of the device.
               ▪ BoundingBox: The spatial extent of the device ($x_1$, $y_1$, $x_2$, $y_2$).
               ▪ ConfidenceScore: Detection confidence value between 0 and 1.
               ▪ UUID: Assigned unique identifier.
       o Append the structured metadata to MetadataOutput.
Output:
  ▪ MetadataOutput: A refined, filtered, and structured list of devices, ready for use by the Geospatial Device Visualizer and Spatial Topology Inference modules.
```

The refined metadata is subsequently passed to the Geospatial Device Visualizer and the Spatial Topology Inference module to support further spatial reasoning, visualization, and control operations.

Geospatial Device Visualizer:

The Geospatial Device Visualizer module is responsible for transforming the refined detection metadata into a human-interpretable visual representation. By overlaying bounding boxes and labels directly onto the input imagery, the module provides an immediate spatial understanding of the detected environment. This visualization acts as an important bridge between raw device detection and higher-level spatial reasoning, enabling both validation of detection accuracy and meaningful analysis of device relationships. This module consumes structured metadata and generates annotated images that illustrate detected devices along with their spatial arrangements. These annotated outputs not only facilitate subsequent AI-driven spatial analysis but also serve as essential tools for user validation, system debugging, and visual confirmation of detection outputs.

The module follows a structured multi-stage process to generate annotated visualizations from refined metadata:

Algorithm: Geospatial Device Visualization from Refined Metadata

```
Input:
    ▪ InputImage: Captured image from onboarding or command activation phase.
    ▪ RefinedMetadata: Structured list of detected devices including device type, bounding box
      coordinates, UUID, and optional confidence scores.
Steps:
    1. Load and Preprocess the Image
        o Load the InputImage into the system using OpenCV.
        o Convert the image format from BGR to RGB to ensure accurate color representation
          and compatibility with visualization libraries.
    2. Extract Object Metadata
        o Retrieve metadata for each detected device, including:
            ▪ DeviceType
            ▪ BoundingBoxCoordinates ($x_1$, $y_1$, $x_2$, $y_2$)
            ▪ UUID
            ▪ ConfidenceScore (optional)
        o Group the detected objects by DeviceType to maintain semantic structure and clarity.
    3. Draw Bounding Boxes and Labels
        o For each detected device:
            ▪ Render a bounding box at the corresponding BoundingBoxCoordinates.
            ▪ Attach a label indicating the DeviceType and optionally append UUID and
              ConfidenceScore.
        o Assign distinct colors dynamically to different DeviceType categories for clear visual
          differentiation.
        o Apply Spatial Labeling Order:
            ▪ Horizontal Ordering: Label devices left to right across the horizontal axis.
            ▪ Vertical Ordering: When multiple devices share the same horizontal
              alignment, label from top to bottom.
        o Ensure minimal label overlaps and high readability.
    4. Save the Annotated Image
        o Save the final annotated image in PNG or JPEG format to a designated output
          directory.
        o Utilize the annotated image for two main purposes:
```

```
                 Encode in Base64 and forward to the Spatial Topology Inference module for
                   AI-based spatial reasoning.
                 Optionally display to users for validation or developers for debugging.
    5.  Incorporate Error Tolerance and User Control
            o  Allow users to refresh the automatic annotation pipeline, triggering reprocessing of the
               input image.
            o  Provide a manual annotation interface enabling users to correct or adjust bounding
               boxes through a drag-and-drop GUI.
            o  These mechanisms ensure reliable annotation quality and foster user trust in system
               outputs.
Output:
      AnnotatedImage: A spatially contextualized, visually annotated image ready for
        downstream spatial topology analysis and user verification.
```

Users can trigger a refresh of the automatic annotation pipeline, prompting reprocessing of the original image. Alternatively, a manual annotation interface is provided, allowing users to adjust or redefine device bounding boxes through a drag-and-drop GUI.

These provisions ensure greater reliability of the final annotated visual output and enhance user trust in system-generated spatial representations. Through this comprehensive visual annotation workflow, the system develops a spatially contextualized and accurate understanding of the smart environment, establishing an important foundation for intelligent IoT device control and automation.

Spatial Topology Inference Engine:

The Spatial Topology Inference Module is an important component that extends beyond device detection to analyze the spatial arrangement of IoT devices within their environmental context. Rather than treating smart devices as isolated entities, this module infers relational information between devices and environmental features. This spatial understanding facilitates intelligent, context-aware decision-making, providing the foundation for human-like reasoning in smart environments. The Spatial Topology Inference Module serves as the bridge between perception and reasoning AND it transforms annotated images and structured metadata into rich spatial insights using GPT-4o, a state-of-the-art vision-language model. In this module it interprets the spatial layout and orientation of devices within the environment. Enables context-aware command synthesis by factoring real-world constraints. Supplies structured spatial descriptions to the Command Generation Module, informing precise device targeting and automation logic.

Intent Based Agentic Command Synthesis

The Command Generation Module serves as the cognitive core of the smart IoT control pipeline, synthesizing user intent and spatial device information into executable control instructions. By integrating natural language understanding and spatial metadata, the module enables precise, context-sensitive automation within dynamic environments.

Algorithm: Command Generation for Context-Aware Smart IoT Control

```
Input:
```

- UserIntent: Parsed natural language command from the Onboarding Inference Engine.
- DeviceMetadata: List of devices with UUIDs, device types, and associated labels.
- SpatialDescriptions: Topological and contextual descriptions from the Spatial Topology Inference module.

Steps:

1. **Input Integration**
   - Merge UserIntent, DeviceMetadata, and SpatialDescriptions into a unified input space.
   - Ensure each device entry is associated with spatial cues and a UUID.
2. **Prompt Construction for Large Language Model (LLM)**
   - Prepare a structured prompt for Gemini Flash, incorporating:
     - **Explicit Intent Embedding**: Place the user's natural language command at the beginning.
     - **Device Enumeration with Spatial Context**: List available devices with their UUIDs and brief spatial descriptions.
     - **Contextual Emphasis**: Highlight important spatial landmarks (e.g., “leftmost light”, “fan near window”).
     - **Output Format Specification**: Instruct Gemini to return results in structured, machine-readable formats (e.g., JSON or key-value pairs).
3. **LLM-Based Command Synthesis**
   - Submit the constructed prompt to Gemini Flash LLM.
   - Receive a structured response that maps user intent to specific device actions based on spatial relevance.
4. **Handling Multi-Device Instructions**
   - If the user command targets multiple devices:
     - **Filter** candidate devices based on device type, spatial cues, and specified quantity.
     - **Rank** candidates by confidence scores and contextual relevance.
     - **Select** top N matching devices.
     - **Generate** a list of structured actionable instructions, one per device.
5. **Command Structuring**
   - Format the final output into a consistent schema for the Execution Module, typically containing:
     - UUID: Target device unique identifier.
     - Action: Intended operation (e.g., “switch_on”, “dim_light”, “increase_speed”).

Output:

- StructuredCommands: A list of machine-executable control instructions, ready for dispatch to the Execution Module.

Agentic Execution and Action Module

The Execution Module represents the final stage of the IoT control pipeline, responsible for translating structured control commands into real-world actions on smart devices. Acting as the operational backbone of the system, this module ensures that user instructions and system-generated commands are effectively and reliably executed through standardized communication protocols. It receives structured control instructions, interprets them, and initiates device-specific actions, thereby completing the loop from user intent to tangible automation.

The module ingests structured control commands, typically formatted in JSON or dictionary-like structures, containing the following important attributes:

- UUID: A Universally Unique Identifier specifying the target IoT device.
- Action: The specific control action to be performed

Device Communication via TuyaAPI

TuyaAPI provides a standardized platform for secure communication with a wide range of IoT devices over Wi-Fi. It abstracts the complexity of device interaction by handling authentication, command encoding, network messaging, and response management, enabling seamless device control.

Algorithm: IoT Device Command Execution via TuyaAPI

```
Input: Structured control command containing:
UUID (Universally Unique Identifier of the target device)
Action (Specific operation to be performed, e.g., "turn-on", "adjust-brightness")
Output: Successful execution of device action or appropriate error handling
Algorithm Steps:
    1. API Authentication
    o  Initiate authentication with TuyaAPI using secure credentials (API key, security token, or OAuth).
    o  If authentication is successful, proceed to Step 2.
    o  If authentication fails, log the error and terminate the process.
    2. Command Transmission
    o  Encode the structured command (UUID and Action) into a TuyaAPI-compliant request.
    o  Transmit the encoded command to the target IoT device via TuyaAPI.
    3. Command Execution by Device
    o  Upon receiving the command, the IoT device decodes the instruction.
    o  The device performs the specified action
    o  The device generates a response indicating the success or failure of the action.
    4. Response Handling
    o  Process the response received from the device:
           o  If the action is successful:
                  ▪  Log the successful execution event.
    o  If the action fails:
           o  Trigger error-handling mechanisms, which may include:
```

Retrying the command transmission.
Notifying the user or system administrator.
Escalating the error for manual intervention if necessary.

Implementation

This section covers the structured breakdown of the implementation for the proposed spatial context-aware smart device control system for the scenario shown in fig 4. The proposed system enables users to control smart home devices using natural language commands that reference spatial context. This is achieved through a modular pipeline comprising several components, each responsible for a specific function in the process.

Onboarding Inference Engine

The Onboarding Inference Engine serves as the initial interface between the user and the system. Users provide a natural language description of the devices present in their environment, such as "There are 4 lights and 1 fan in the room." This input is processed using a language model (e.g., Qwen 2.5) to extract structured information about device types and quantities. The output is a JSON object, for example, "light": 4, "fan": 1}, which informs subsequent modules about the devices to detect and control as shown in Fig 3.

Zero-Shot Device Detection

This module employs a zero-shot object detection model, such as OWL-ViT, to identify and localize devices within a room image without prior training on specific device types. By leveraging the device types obtained from the onboarding phase, the model generates prompts to detect corresponding objects in the image. The output includes bounding boxes, labels, and confidence scores for each detected device as shown in Fig 5.

Metadata Refinement and Filtering

After the device detection, the system refines the raw outputs to ensure accuracy and consistency. Each detected device is assigned a UUID, and detections with confidence scores below a predefined threshold are discarded. The remaining devices are sorted based on their spatial arrangement (e.g., left-to-right, top-to-bottom) to maintain a coherent structure. The resulting metadata includes device type, location, UUID, and confidence score, forming a reliable foundation for subsequent modules as shown in Fig 6.

The Geospatial Device Visualizer provides a visual representation of the devices detected within the room image. By overlaying bounding boxes and labels onto the original image, users can verify the accuracy of detections and understand the spatial distribution of devices as shown in Fig 7. This visualization aids in both user validation and as an input for spatial reasoning in the next module.

Spatial Topology Inference

Utilizing models like GPT-4o, this module analyzes the annotated image and metadata to infer spatial relationships between devices and other room elements. It generates textual descriptions detailing each device's position relative to landmarks (e.g., "Light1 is above the desk and near the wall clock"), enabling the system to comprehend spatial context and disambiguate user commands effectively as shown in Fig 8.

Intent-based Agentic Command Synthesis

When a user issues a natural language command, “Turn on every light” is precisely translated into executable instructions for each light in the environment, demonstrating the system’s ability to intelligently interpret, synthesize, and act on natural language commands as shown in Fig 9.

Agentic Actuation & Execution Module

The final module executes the generated commands by interfacing with smart home APIs, such as the Tuya Smart Device API. It authenticates with the API, transmits the control commands, and handles responses to confirm successful execution or manage errors. This module completes the control loop, translating user intent into physical actions within the smart home environment as shown in Fig 10.

Fig 11 represents the complete system workflow for the proposed Spatial Context-Aware Smart Device Control System for the given environment.

Experimental Setup and Result Analysis

Case Study Design

To simulate real-world deployment scenarios, participants were introduced into a smart home environment without prior knowledge of the device configurations or naming conventions. This setup emulated a typical user entering an unfamiliar smart space.

Before the main evaluation, participants received a demonstration highlighting the basic functionalities of both the Google Home Assistant and the proposed method, especially aimed at participants with no prior experience with smart home technologies. Google Home Assistant – uses Gemini as a LLM in the backend

During the Google Home Assistant session, participants were tasked with completing predefined operations by either:

- Consulting a device map,
- Requesting assistance from the researcher, or
- Recalling specific device ID labels (e.g., "Switch on light 4").

In the proposed method, participants issued natural language commands incorporating spatial context without needing to reference explicit device IDs.

For consistency, both assistants were activated via designated hotkeys:

- Space bar for the proposed system
- Microphone logo key for Google Home.

Voice-based wake-word activation was deliberately disabled to eliminate ambiguities and ensure uniform conditions across all participants.

Reference tasks provided included:

- "Switch on the light near the AC."
- "Switch on the light above the photo frame."
- "Turn on the light on the desk."
- "Switch on the leftmost light."
- "Turn on the fan."
- "Turn on lighting for studying or working."

Participants were also encouraged to issue open-ended commands based on their own interpretation of the environment, ensuring a balance between guided and exploratory interactions.

Participant Demographics

A total of fifteen participants were recruited for the study, with ages ranging from 18 to 80 years (Mean: 45.8 years, Median: 49 years, Standard Deviation: 19.08).
The gender distribution included:

- 8 females (53.3%)
- 7 males (46.7%).

Educational backgrounds varied significantly:

- One participant had education below 10th standard.
- One participant had completed senior secondary education (12th standard).
- One participant held a doctoral degree.
- The remainder held or were pursuing bachelor's degrees.

Prior experience with smart home systems was notably limited:

- Only two participants had actively used Amazon Echo devices.
- One participant reported past exposure.
- The remaining participants had no prior experience with smart home technologies.

Participants were introduced to device ID labels only before the Google Home interaction phase, ensuring that their experience with the proposed method Assistant remained unaffected and as naturalistic as possible.

Experimental Infrastructure

- **Hardware for Proposed model**:
  - Laptop with Intel® Core™ i7-10510U processor
  - 16 GB RAM
  - Windows 11 Operating System
  - Built-in microphone
- **Hardware for Google Home Assistant**:
  - Android smartphone configured for Google Home device integration.

Experience and Usability

Participants reported a higher ease of task completion when interacting with the proposed method compared to Google Home Assistant. On a five-point Likert scale, where 1 indicated "very hard" and 5 indicated "very easy," the Google Home Assistant achieved a mean usability score of 3.8 and a median of 4, whereas the proposed method attained a mean of 4.67 and a median of 4. When asked about difficulties in expressing commands, a majority (6 out of 15 participants) reported no issues, indicating growing confidence and fluency after a brief familiarization period. However, participants interacting with Google Home Assistant noted challenges such as difficulty recalling device names, confusion between "on" and "off" commands, and uncertainty arising from the reliance on numerical device identifiers. In contrast, the proposed method's natural language-based and spatially aware interaction model alleviated such issues, enabling participants to express commands more intuitively and confidently. Users reported greater cognitive load and self-consciousness when issuing commands through Google Home Assistant due to its rigid identifier-based syntax. Conversely, the proposed method allowed for more natural, free-form expressions, further improving user confidence and interaction fluidity.

Emotional Reactions

Approximately 40% of participants found Google Home Assistant to be easy and comfortable to use; however, 53% cited the need to remember device IDs as a significant drawback. Three participants specifically noted that commands often needed to be overly specific for successful execution.

Outcomes for Google Home Assistant were mixed:

- 6 participants (40%) reported a positive experience,
- 4 participants (26.7%) reported a neutral experience, and

- 5 participants (33.3%) reported a negative experience.

The proposed method was positively received by 73% of users. Participants appreciated its spatial context-awareness and the ability to interact without memorizing device names. Support for regional languages was also highlighted as a major accessibility advantage. Minor challenges were reported, including ambiguity in object references (e.g., differentiating between “photo,” “painting,” or “red board”) and time limitations during command issuance. Overall, 11 participants (73%) had a positive experience with the proposed method, 2 participants (13.3%) were neutral, and 2 participants (13.3%) had a negative experience. Furthermore, 80% of participants reported no confusion or frustration with either system. When confusion did occur, it was predominantly associated with device ID dependency and multi-command processing in Google Home Assistant, and ambiguity in visual object references in the proposed method. 14 out of 15 participants (93.3%) reported enjoying the proposed method experience, citing advantages such as automatic light mapping, the ease of delivering complex commands, image-based spatial recognition, and robust natural language processing capabilities.

User Preference and Future Adoption

A strong preference emerged for the proposed method, with 14 of 15 participants (93.3%) indicating that they would prefer it over Google Home Assistant for future use. Participants especially valued directional language support, such as "turn on the lights to my left," which operated seamlessly with the proposed method but was not possible with Google Home Assistant. One participant expressed concerns regarding image data privacy with the proposed method, although they acknowledged that the system addressed these concerns appropriately. Another participant favored Google Home Assistant due to its more refined mobile interface. In terms of system trust, 80% of users expressed confidence in the proposed method’s ability to control devices without relying on device names or identifiers. Suggestions for future improvements include enhancing the user interface and developing even more intuitive communication methods.

NASA-TLX Cognitive Load Assessment

Cognitive workload was assessed using the NASA-TLX following interaction with both systems. Results demonstrated a significant reduction in perceived workload when using the proposed method compared to the baseline Google Home Assistant condition.

Mean TLX score for Google Home Assistant (Condition 1): 35.24 (SD = 10.84)

Mean TLX score for the proposed method (Condition 2): 22.06 (SD = 7.97)

This reflects an average reduction of 13.17 points. The median TLX score also decreased from 35.71 to 19.05. A broad shift toward lower workload scores was observed across all percentiles, indicating consistent user experience improvements. Fig 12 -14 shows the cognitive load and user experience evaluation of your spatial context-aware control system (Condition 2) compared to a baseline system (Google Home Assistant, Condition 1) using the NASA Task Load Index (TLX) methodology. Figure 12 presents a boxplot comparison of NASA-TLX scores between the baseline system (Condition 1 – Google Home Assistant) and the proposed spatial context-aware control system (Condition 2). The average TLX score under Condition 1 was significantly higher (Mean = 35.24, SD = 10.84) compared to Condition 2 (Mean = 22.06, SD = 7.97), indicating a notable reduction in perceived cognitive workload. The interquartile range in Condition 2 is narrower, suggesting more consistent user experience and lower variability in perceived effort. This figure clearly highlights the improved usability of the proposed system across diverse users.

Figure 13 shows the density distribution of TLX scores under both conditions. The curve representing Condition 2 is strongly skewed toward the lower end of the workload spectrum, whereas Condition 1's distribution is wider and centered around a higher mean. The separation of the distributions further reinforces that participant consistently experienced lower cognitive load when using the spatial context-aware system. The non-overlapping peaks confirm the system's efficiency in minimizing user stress and mental demand.

Figure 14 compares the mean scores of the six TLX sub-dimensions across the two systems. The proposed system (Condition 2) performed better across all dimensions, particularly in terms of mental demand, temporal demand, and effort. Users also reported reduced frustration and improved performance, reflecting a more intuitive and fluid interaction experience. Table 2 shows the comparison of average TLX sub-dimension scores between the baseline system (Condition 1) and the proposed system (Condition 2), highlighting the difference and percentage reduction in workload. Table 3 shows the results of statistical analysis comparing both systems. It includes t-statistics, p-values, significance indicators, and effect size interpretations for each TLX subscale. Table 4 shows how individual participants rated the proposed system (Condition 2) compared to the baseline system (Condition 1), indicating whether they found it lower, equal, or higher in workload per dimension.

The proposed method greatly improves the usability of smart homes through visual understanding, it also naturally raises privacy concerns. Continuous video monitoring even when used purely for real-time reasoning can make users feel uneasy, especially in private spaces like bedrooms or living areas. During the user studies, one participant specifically voiced concern about the possibility of sensitive information being captured unintentionally. Although the proposed method does not store or learn from user data, it relies solely on pre-trained models it's clear that the presence of always-on cameras requires careful attention to privacy. It's important to ensure that all processing happens locally on the device, avoiding any need to transmit data externally. Other improvements, like automatically blurring sensitive parts of a room, setting strict deletion timelines for visual data, and using encrypted processing pipelines, will be important for building trust. Clear communication with users about what data is being processed and why will also be key to making the system feel safe and respectful.

Personalization and Adapting to Users

The proposed method already offers strong spatial awareness, the next step is making it even more user centered. Future development should focus on adapting to users' preferences and behaviors naturally without needing them to always give explicit commands. This could include recognizing users by their voice, adjusting lights or climate based on mood detected from speech, or remembering daily habits, like automatically turning on the lights at 6 a.m. Personalization could make the smart home experience feel seamless and intuitive. To balance personalization with privacy, techniques like federated learning where the system learns from data locally without sending it to central servers should be explored. As the system is introduced into more diverse homes and lifestyles, it will also need to become even more robust and adaptable. Building a system that can evolve with users over time will be essential to maintaining its usefulness and trustworthiness.

Conclusion and Future Scope

This work introduced a novel spatial context-aware control system for smart devices that fundamentally reimagines how users interact with IoT environments. By combining advanced computer vision, natural language processing, and spatial reasoning, the proposed method overcomes important limitations of traditional IoT control systems that depend heavily on device-specific identifiers and preconfigured setups. Our comprehensive user study demonstrated that the proposed method significantly outperforms conventional solutions like Google Home Assistant across multiple dimensions. In particular, the NASA-TLX assessment showed a substantial reduction in cognitive workload, with users reporting a mean score of 22.06 compared to 35.24 for Google Home Assistant. Furthermore, 93.3% of participants experienced a lower cognitive burden, and 87% expressed a clear preference for the proposed method due to its intuitive spatial context-aware commands, elimination of the need to memorize device IDs, and support for regional languages.

The system's modular architecture includes components such as the Onboarding Inference Engine, Zero-Shot Device Detection, Metadata Refinement, Geospatial Device Visualization, Spatial Topology Inference, and Intent-Based Command Synthesis enables dynamic, seamless adaptation to changing environments without the need for manual reconfiguration. This marks a significant advance over existing solutions that typically require static device labeling and rigid automation rules. However, several avenues remain for future enhancement. Key challenges include strengthening privacy safeguards during image-based processing, expanding device compatibility across a wider range of manufacturers, and optimizing the system to perform efficiently on resource-constrained edge devices. Additionally, exploring lightweight LLM architectures specifically tailored for IoT control could help maintain real-time responsiveness while reducing computational demands.

Another important direction for expansion involves integrating SLAM (Simultaneous Localization and Mapping) with smart glasses. By equipping users with wearable devices capable of mapping the environment in real time, the system could provide even more natural, hands-free spatial interactions. Commands such as "Turn on the light to my right" would dynamically adapt based on user orientation, enhancing autonomy and intuitive control especially for elderly individuals, people with disabilities, or users requiring continuous environmental awareness. The potential applications of the proposed method extend beyond traditional home automation. By removing the cognitive burden of remembering device names and enabling natural spatial language commands, the system creates a more accessible and empowering smart environment for diverse users, including those with cognitive or physical challenges. By bridging the gap between human spatial understanding and machine control, the proposed method lays the foundation for the next generation of smart environment spaces that adapt to human needs, instead of requiring humans to adapt to technological constraints.

References

Achiluzzi E, Li M, Georgy MFA, Kashef R. 2022. Exploring the Use of Data-Driven Approaches for Anomaly Detection in the Internet of Things (IoT) Environment. DOI: 10.48550/arXiv.2301.00134.

AlMahamid F, Lutfiyya H, Grolinger K. 2022. Virtual Sensor Middleware: Managing IoT Data for the Fog-Cloud Platform. In: *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. 41–48. DOI: 10.1109/CCECE49351.2022.9918499.

AlQahtani AAS, Alamleh H, Smadi BA. 2022. Technical Report-IoT Devices Proximity Authentication In Ad Hoc Network Environment. DOI: 10.48550/arXiv.2210.00175.

An T, Zhou Y, Zou H, Yang J. 2024. IoT-LLM: Enhancing Real-World IoT Task Reasoning with Large Language Models. DOI: 10.48550/arXiv.2410.02429.

Baby K. 2014. *Big Data: An Ultimate Solution in Health Care*.

Bader SR, Maleshkova M. 2019. Virtual Representations for Iterative IoT Deployment. DOI: 10.48550/arXiv.1903.00718.

Baumann R, Malik KM, Javed A, Ball A, Kujawa B, Malik H. 2019. Voice Spoofing Detection Corpus for Single and Multi-order Audio Replays. DOI: 10.48550/arXiv.1909.00935.

Chatterjee B, Seo D-H, Chakraborty S, Avlani S, Jiang X, Zhang H, Abdallah M, Raghunathan N, Mousoulis C, Shakouri A, Bagchi S, Peroulis D, Sen S. 2020. Context-Aware Collaborative-Intelligence with Spatio-Temporal In-Sensor-Analytics in a Large-Area IoT Testbed. DOI: 10.48550/arXiv.2005.13003.

Chen B, Xu Z, Kirmani S, Ichter B, Driess D, Florence P, Sadigh D, Guibas L, Xia F. 2024. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. DOI: 10.48550/arXiv.2401.12168.

Dang TK, Tran KTK. 2019. The Meeting of Acquaintances: A Cost-efficient Authentication Scheme for Light-weight Objects with Transient Trust Level and Plurality Approach. DOI: 10.48550/arXiv.1903.10018.

Fotiou N, Siris VA, Xylomenos G, Polyzos GC, Katsaros KV, Petropoulos G. 2017. Edge-ICN and its application to the Internet of Things. DOI: 10.48550/arXiv.1707.01721.

Goyal H, Kodali K, Saha S. 2022. LiPI: Lightweight Privacy-Preserving Data Aggregation in IoT. DOI: 10.48550/arXiv.2207.12197.

Han K, Huang K. 2016. Wirelessly Powered Backscatter Communication Networks: Modeling, Coverage and Capacity. In: *2016 IEEE Global Communications Conference (GLOBECOM)*. 1–6. DOI: 10.1109/GLOCOM.2016.7842391.

Harini S, Ravikumar A. 2020. Effect of Parallel Workload on Dynamic Voltage Frequency Scaling for Dark Silicon Ameliorating. In: *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. 1012–1017. DOI: 10.1109/ICOSEC49089.2020.9215262.

Homssi BA, Al-Hourani A, Chandrasekharan S, Gomez KM, Kandeepan S. 2020. On the Bound of Energy Consumption in Cellular IoT Networks. *IEEE Transactions on Green Communications and Networking* 4:355–364. DOI: 10.1109/TGCN.2019.2960061.

Jiang X, zhang H, Yi EAB, Raghunathan N, Mousoulis C, Chaterji S, Peroulis D, Shakouri A, Bagchi S. 2021. Hybrid Low-Power Wide-Area Mesh Network for IoT Applications. *IEEE Internet of Things Journal* 8:901–915. DOI: 10.1109/JIOT.2020.3009228.

Kaplan A, Vieira J, Larsson EG. 2024. Direct Link Interference Suppression for Bistatic Backscatter Communication in Distributed MIMO. *IEEE Transactions on Wireless Communications* 23:1024–1036. DOI: 10.1109/TWC.2023.3285250.

Khan S, Alam M. 2020. Wearable Internet of Things for Personalized Healthcare Study of Trends and Latent Research. DOI: 10.48550/arXiv.2005.06958.

King E, Yu H, Lee S, Julien C. 2023. “Get ready for a party”: Exploring smarter smart spaces with help from large language models. DOI: 10.48550/arXiv.2303.14143.

King E, Yu H, Lee S, Julien C. 2024. Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8:12:1-12:38. DOI: 10.1145/3643505.

Kontar R, Shi N, Yue X, Chung S, Byon E, Chowdhury M, Jin J, Kontar W, Masoud N, Noueihed M, Okwudire CE, Raskutti G, Saigal R, Singh K, Ye Z. 2021. The Internet of Federated Things (IoFT): A Vision for the Future and In-depth Survey of Data-driven Approaches for Federated Learning. *IEEE Access* 9:156071–156113. DOI: 10.1109/ACCESS.2021.3127448.

Li M, Wu Y. 2022. Intelligent control system of smart home for context awareness. *International Journal of Distributed Sensor Networks* 18:15501329221082030. DOI: 10.1177/15501329221082030.

Liu C, Chen B, Shao W, Zhang C, Wong KKL, Zhang Y. 2024. Unraveling Attacks to Machine-Learning-Based IoT Systems: A Survey and the Open Libraries Behind Them. *IEEE Internet of Things Journal* 11:19232–19255. DOI: 10.1109/JIOT.2024.3377730.

Liu Y, Wang J, Li J, Niu S, Song H. 2021a. Zero-bias Deep Learning Enabled Quick and Reliable Abnormality Detection in IoT. DOI: 10.48550/arXiv.2105.15098.

Liu Y, Wang J, Li J, Song H, Yang T, Niu S, Ming Z. 2021b. Zero-Bias Deep Learning for Accurate Identification of Internet-of-Things (IoT) Devices. *IEEE Internet of Things Journal* 8:2627–2634. DOI: 10.1109/JIOT.2020.3018677.

Ma D, Lan G, Hassan M, Hu W, Das SK. 2020. Sensing, Computing, and Communication for Energy Harvesting IoTs: A Survey. *IEEE Communications Surveys & Tutorials* 22:1222–1250. DOI: 10.1109/COMST.2019.2962526.

Maghsoudi M, Nourbakhsh R, Kermani MAM, Khanizad R. 2023. The Power of Patents: Leveraging Text Mining and Social Network Analysis to Forecast IoT Trends. DOI: 10.48550/arXiv.2309.00707.

Masuduzzaman M, Mahmud A, Islam A, Islam MM. 2019. Two Phase Authentication and VPN Based Secured Communication for IoT Home Networks. DOI: 10.48550/arXiv.1910.13625.

Mehan Y, Gupta K, Jayanti R, Govil A, Garg S, Krishna M. 2024. QueSTMaps: Queryable Semantic Topological Maps for 3D Scene Understanding. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 13311–13317. DOI: 10.1109/IROS58592.2024.10801814.

Meyuhas B, Bremler-Barr A, Shapira T. 2024. IoT Device Labeling Using Large Language Models. DOI: 10.48550/arXiv.2403.01586.

Mihai V, Hanganu CE, Stamatescu G, Popescu D. 2019. WSN and Fog Computing Integration for Intelligent Data Processing. DOI: 10.48550/arXiv.1903.09507.

Morris A, Guan J, Lessio N, Shao Y. 2020. Toward Mixed Reality Hybrid Objects with IoT Avatar Agents. In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 766–773. DOI: 10.1109/SMC42975.2020.9282914.

Mu M. 2020. WiFi-based Crowd Monitoring and Workspace Planning for COVID-19 Recovery. DOI: 10.48550/arXiv.2007.12250.

OpenAI, Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow AJ, Welihinda A, Hayes A, Radford A, Mądry A, Baker-Whitcomb A, Beutel A, Borzunov A, Carney A, Chow A, Kirillov A, Nichol A, Paino A, Renzin A, Passos AT, Kirillov A, Christakis A, Conneau A, Kamali A, Jabri A, Moyer A, Tam A, Crookes A, Tootoochian A, Tootoonchian A, Kumar A, Vallone A, Karpathy A, Braunstein A, Cann A, Codispoti A, Galu A, Kondrich A, Tulloch A, Mishchenko A, Baek A, Jiang A, Pelisse A, Woodford A, Gosalia A, Dhar A, Pantuliano A, Nayak A, Oliver A, Zoph B, Ghorbani B, Leimberger B, Rossen B, Sokolowsky B, Wang B, Zweig B, Hoover B, Samic B, McGrew B, Spero B, Giertler B, Cheng B, Lightcap B, Walkin B, Quinn B, Guarraci B, Hsu B, Kellogg B, Eastman B, Lugaresi C, Wainwright C, Bassin C, Hudson C, Chu C, Nelson C, Li C, Shern CJ, Conger C, Barette C, Voss C, Ding C, Lu C, Zhang C, Beaumont C, Hallacy C, Koch C, Gibson C, Kim C, Choi C, McLeavey C, Hesse C, Fischer C, Winter C, Czarnecki C, Jarvis C, Wei C, Koumouzelis C, Sherburn D, Kappler D, Levin D, Levy D, Carr D, Farhi D, Mely D, Robinson D, Sasaki D, Jin D, Valladares D, Tsipras D, Li D, Nguyen DP, Findlay D, Oiwoh E, Wong E, Asdar E, Proehl E, Yang E, Antonow E, Kramer E, Peterson E, Sigler E, Wallace E, Brevdo E, Mays E, Khorasani F, Such FP, Raso F, Zhang F, Lohmann F von, Sulit F, Goh G, Oden G, Salmon G, Starace G, Brockman G, Salman H, Bao H, Hu H, Wong H, Wang H, Schmidt H, Whitney H, Jun H, Kirchner H, Pinto HP de O, Ren H, Chang H, Chung HW, Kivlichan I, O’Connell I, O’Connell I, Osband I, Silber I, Sohl I, Okuyucu I, Lan I, Kostrikov I, Sutskever I, Kanitscheider I, Gulrajani I, Coxon J, Menick J, Pachocki J, Aung J, Betker J, Crooks J, Lennon J, Kiros J, Leike J, Park J, Kwon J, Phang J, Teplitz J, Wei J, Wolfe J, Chen J, Harris J, Varavva J, Lee JG, Shieh J, Lin J, Yu J, Weng J, Tang J, Yu J, Jang J, Candela JQ, Beutler J, Landers J, Parish J, Heidecke J, Schulman J, Lachman J, McKay J, Uesato J, Ward J, Kim JW, Huizinga J, Sitkin J, Kraaijeveld J, Gross J, Kaplan J, Snyder J, Achiam J, Jiao J, Lee J, Zhuang J, Harriman J, Fricke K, Hayashi K, Singhal K, Shi K, Karthik K, Wood K, Rimbach K, Hsu K, Nguyen K, Gu-Lemberg K, Button K, Liu K, Howe K, Muthukumar K, Luther K, Ahmad L, Kai L, Itow L, Workman L, Pathak L, Chen L, Jing L, Guy L, Fedus L, Zhou L, Mamitsuka L, Weng L, McCallum L, Held L, Ouyang L, Feuvrier L, Zhang L, Kondraciuk L, Kaiser L, Hewitt L, Metz L, Doshi L, Aflak M, Simens M, Boyd M, Thompson M, Dukhan M, Chen M, Gray M, Hudnall M, Zhang M, Aljubeh M, Litwin M, Zeng M, Johnson M, Shetty M, Gupta M, Shah M, Yatbaz M, Yang MJ, Zhong M, Glaese M, Chen M, Janner M, Lampe M, Petrov M, Wu M, Wang M, Fradin M, Pokrass M, Castro M, Castro MOT de, Pavlov M, Brundage M, Wang M, Khan M, Murati M, Bavarian M, Lin M, Yesildal M, Soto N, Gimelshein N, Cone N, Staudacher N, Summers N, LaFontaine N, Chowdhury N, Ryder N, Stathas N, Turley N, Tezak N, Felix N, Kudige N, Keskar N, Deutsch N, Bundick N, Puckett N, Nachum O, Okelola O, Boiko O, Murk O, Jaffe O, Watkins O, Godement O, Campbell-Moore O, Chao P, McMillan P, Belov P, Su P, Bak P, Bakkum P, Deng P, Dolan P, Hoeschele P, Welinder P, Tillet P, Pronin P, Tillet P, Dhariwal P, Yuan Q, Dias R, Lim R, Arora R, Troll R, Lin R, Lopes RG, Puri R, Miyara R, Leike R, Gaubert R, Zamani R, Wang R, Donnelly R, Honsby R, Smith R, Sahai R, Ramchandani R, Huet R, Carmichael R, Zellers R, Chen R, Chen R, Nigmatullin R, Cheu R, Jain S, Altman S, Schoenholz S, Toizer S, Miserendino S, Agarwal S, Culver S, Ethersmith S, Gray S, Grove S, Metzger S, Hermani S, Jain S, Zhao S, Wu S, Jomoto S, Wu S, Shuaiqi, Xia, Phene S, Papay S, Narayanan S, Coffey S, Lee S, Hall S, Balaji S, Broda T, Stramer T, Xu T, Gogineni T, Christianson T, Sanders T, Patwardhan T, Cunninghman T, Degry T, Dimson T, Raoux T, Shadwell T, Zheng T, Underwood T, Markov T, Sherbakov T, Rubin T, Stasi T, Kaftan T, Heywood T, Peterson T, Walters T, Eloundou T, Qi V, Moeller V, Monaco V, Kuo V, Fomenko V, Chang W, Zheng W, Zhou W, Manassra W, Sheu W, Zaremba W, Patil Y, Qian Y, Kim Y, Cheng Y, Zhang Y, He Y, Zhang Y, Jin Y, Dai Y, Malkov Y. 2024. GPT-4o System Card. DOI: 10.48550/arXiv.2410.21276.

Psomas C, Ntougias K, Shanin N, Xu D, Mayer KM, Tran NM, Cottatellucci L, Choi KW, Kim DI, Schober R, Krikidis I. 2024. Wireless Information and Energy Transfer in the Era of 6G Communications. DOI: 10.48550/arXiv.2404.18705.

Qin Y, Sheng QZ, Falkner NJG, Dustdar S, Wang H, Vasilakos AV. 2014. When Things Matter: A Data-Centric View of the Internet of Things. DOI: 10.48550/arXiv.1407.2704.

Ravikumar A, Saritha R, Chandra V. 2013. Support vector machine based prognostic analysis of renal transplantations. In: *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. 1–6. DOI: 10.1109/ICCCNT.2013.6726819.

Ravikumar A, Sriraman H. 2023a. Computationally Efficient Neural Rendering for Generator Adversarial Networks Using a Multi-GPU Cluster in a Cloud Environment. *IEEE Access* 11:45559–45571. DOI: 10.1109/ACCESS.2023.3274201.

Ravikumar A, Sriraman H. 2023b. Real-time pneumonia prediction using pipelined spark and high-performance computing. *PeerJ Computer Science* 9:e1258. DOI: 10.7717/peerj-cs.1258.

Rivkin D, Hogan F, Feriani A, Konar A, Sigal A, Liu S, Dudek G. 2024. SAGE: Smart home Agent with Grounded Execution. DOI: 10.48550/arXiv.2311.00772.

Rivkin D, Hogan F, Feriani A, Konar A, Sigal A, Liu X, Dudek G. 2025. AIoT Smart Home via Autonomous LLM Agents. *IEEE Internet of Things Journal* 12:2458–2472. DOI: 10.1109/JIOT.2024.3471904.

S D, Ravikumar A. 2015. A Study from the Perspective of Nature-Inspired Metaheuristic Optimization Algorithms. *International Journal of Computer Applications* 113:53–56. DOI: 10.5120/19858-1810.

Salehi S, DeMara RF. 2019. Adaptive Non-Uniform Compressive Sensing using SOT-MRAM Multibit Crossbar Arrays. DOI: 10.48550/arXiv.1911.08633.

Sayed A, Himeur Y, Alsalemi A, Bensaali F, Amira A. 2022. Intelligent edge-based recommender system for internet of energy applications. *IEEE Systems Journal* 16:5001–5010. DOI: 10.1109/JSYST.2021.3124793.

Schulthess L, Villani F, Mayer P, Magno M. 2022. RF Power Transmission for Self-sustaining Miniaturized IoT Devices. In: *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. 1–4. DOI: 10.1109/ICECS202256217.2022.9970865.

Shi Z, Wang H, Fu Y, Yang G, Ma S, Hou F, Tsiftsis TA. 2022. Zero-Forcing Based Downlink Virtual MIMO-NOMA Communications in IoT Networks. DOI: 10.48550/arXiv.2209.11382.

Spandan DD, Iqbal R. 2024. ProxeGraph: Scene Graph Generation Utilizing Proxemics for Smart Homes. In: *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. 109–115. DOI: 10.1109/MIPR62202.2024.00024.

Sun P, Wu L, Wang Z. 2021. Towards Efficient Compressive Data Collection in the Internet of Things. DOI: 10.48550/arXiv.2106.00509.

Ullah M, Nardelli PHJ, Wolff A, Smolander K. 2020. Twenty-one key factors to choose an IoT platform: Theoretical framework and its applications. *IEEE Internet of Things Journal* 7:10111–10119. DOI: 10.1109/JIOT.2020.3000056.

Wang Y, Chen S-Y, Zhou Z, Li S, Li H, Zhou W, Li H. 2024a. ROOT: VLM based System for Indoor Scene Understanding and Beyond. DOI: 10.48550/arXiv.2411.15714.

Wang Z, Yan Z, Li S, Liu J. 2024b. Vlm-Based Scene Graph Generation for Industrial Spatial Intelligence.

Wisy S. 2021. Simple Trust Metric in a Low-Power Sensor Network. DOI: 10.48550/arXiv.2102.01041.

Yamauchi M, Tanaka M, Ohsita Y, Murata M, Ueda K, Kato Y. 2021. Smart-home anomaly detection using combination of in-home situation and user behavior. DOI: 10.48550/arXiv.2109.14348.

Yang C, Xu R, Guo Y, Huang P, Chen Y, Ding W, Wang Z, Zhou H. 2023. Improving Vision-and-Language Reasoning via Spatial Relations Modeling. DOI: 10.48550/arXiv.2311.05298.

Zambonelli F. 2016. Towards a General Software Engineering Methodology for the Internet of Things. DOI: 10.48550/arXiv.1601.05569.

Zhang H, Uddin M, Hao F, Mukherjee S, Mohapatra P. 2019. AIDE: Augmented Onboarding of IoT Devices at Ease. In: *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*. HotMobile '19. New York, NY, USA: Association for Computing Machinery, 123–128. DOI: 10.1145/3301293.3302354.

Zheng S, Chen R, Li M, Ye Z, Ceze L, Liang Y. 2024. vMCU: Coordinated Memory Management and Kernel Optimization for DNN Inference on MCUs. DOI: 10.48550/arXiv.2406.06542.

Zheng X, Zhou S, Niu Z. 2020. Urgency of Information for Context-Aware Timely Status Updates in Remote Control Systems. DOI: 10.48550/arXiv.2002.07987.

Zong M, Hekmati A, Guastalla M, Li Y, Krishnamachari B. 2025. Integrating large language models with internet of things: applications. *Discover Internet of Things* 5:2. DOI: 10.1007/s43926-024-00083-4.

Table 1(on next page)

Research Gaps

Area	Gaps Identified
Intelligent Decision-Making and Interaction	Limited formal modeling of spatial relationships; dependency on implicit LLM-driven reasoning without structured spatial calculus.
Adaptive and Predictive Systems	Lack of integration of real-time multimodal data for dynamic adaptability in smart home environments.
Energy Management and Efficiency	Insufficient incorporation of fine-grained user context and environmental variations for optimizing energy strategies.
LLM-Orchestrated Flexible Smart Home Control	Challenges in disambiguating spatial references and recovering from execution failures during real-time interactions.
Spatial Environment and Interaction Modeling	Focus mainly on static scene interpretation; limited real-time spatial reasoning and automation in dynamic settings.
Contextual and Goal-Based Approaches	Ineffective handling of fine-grained spatial references in user commands; primarily room-level actions only.
Spatial Topology Inference Methods	Designed for analytical tasks, not real-time smart home automation; limited flexibility for dynamic, heterogeneous environments.
Device Onboarding and Management	Inadequate contextual understanding and absence of multimodal (vision + language) integration for seamless device mapping.
Multimodal IoT Systems	Predominantly rely on linguistic cues without real-time visual scene interpretation or dynamic spatial adaptability.

1

Table 2(on next page)

NASA-TLX Dimension Comparison – Mean Scores and Percentage Change

Dimension	**Condition 1**	**Condition 2**	**Difference**	**% Change**
Mental	40.95	21.90	19.05	46.51%
Physical	18.10	15.24	2.86	15.79%
Temporal	37.14	17.14	20.00	53.85%
Performance	40.95	29.52	11.43	27.91%
Effort	50.48	20.95	29.52	58.49%
Frustration	23.81	27.62	-3.81	-16.00%
Overall	**35.24**	**22.06**	**13.17**	**37.39%**

Table 3(on next page)

Statistical Significance and Effect Size by TLX Dimension

Dimension	t-statistic	p-value	Significant	Effect Size
Mental	3.0054	0.0095	Yes	Medium
Physical	1.0000	0.3343	No	Small
Temporal	3.0725	0.0083	Yes	Medium
Performance	1.4446	0.1706	No	Small
Effort	4.4678	0.0005	Yes	Large
Frustration	-0.8446	0.4125	No	Small

Table 4(on next page)

Participant Response Patterns Across TLX Dimensions

Dimension	Lower in C2	Same in Both	Higher in C2	Dimension
Mental	11 (73.3%)	3 (20.0%)	1 (6.7%)	Mental
Physical	2 (13.3%)	12 (80.0%)	1 (6.7%)	Physical
Temporal	9 (60.0%)	5 (33.3%)	1 (6.7%)	Temporal
Performance	8 (53.3%)	4 (26.7%)	3 (20.0%)	Performance
Effort	11 (73.3%)	3 (20.0%)	1 (6.7%)	Effort
Frustration	2 (13.3%)	7 (46.7%)	6 (40.0%)	Frustration

Figure 1

Proposed Model System Architecture

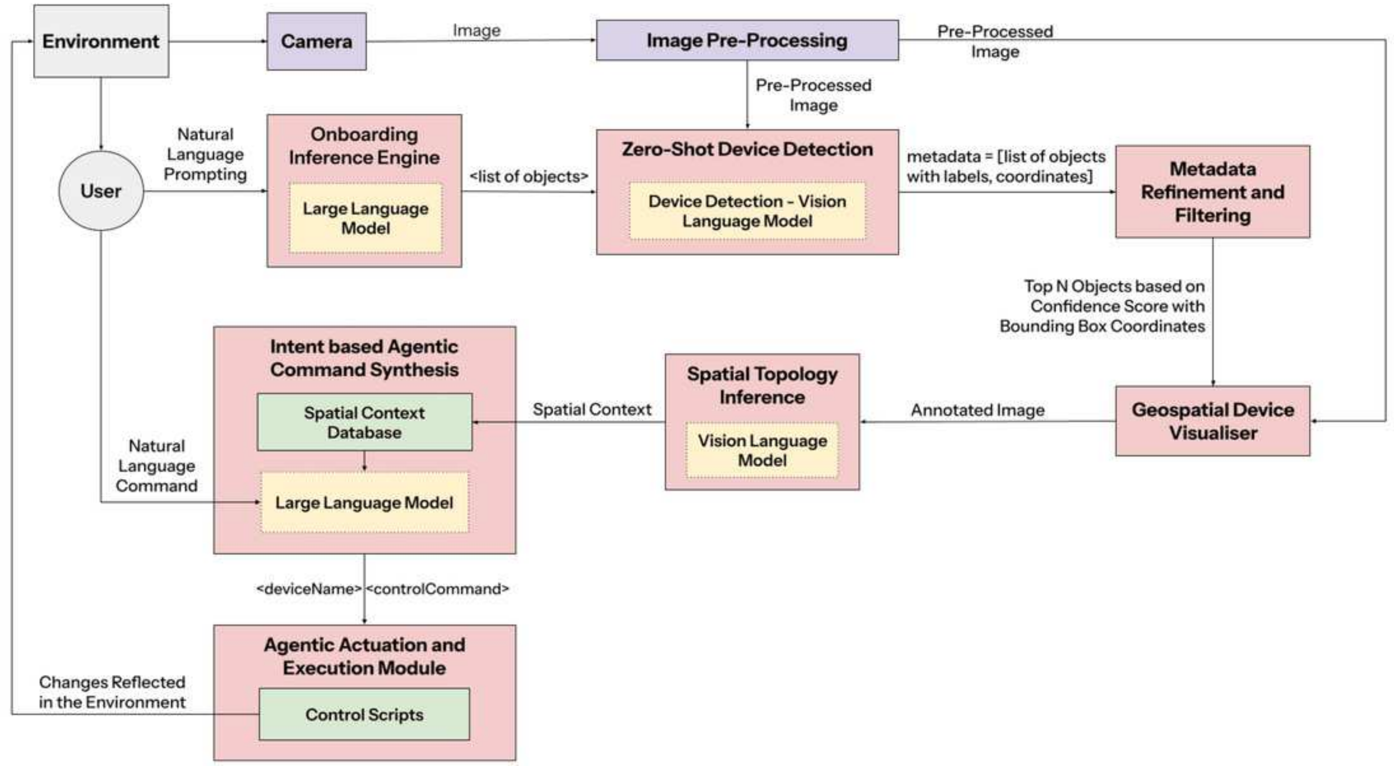

Environment
Camera
Image
Image Pre-Processing
Pre-Processed Image
Pre-Processed Image
User
Natural Language Prompting
Onboarding Inference Engine
Large Language Model
<list of objects>
Zero-Shot Device Detection
Device Detection - Vision Language Model
metadata = [list of objects with labels, coordinates]
Metadata Refinement and Filtering
Top N Objects based on Confidence Score with Bounding Box Coordinates
Geospatial Device Visualiser
Annotated Image
Spatial Topology Inference
Vision Language Model
Spatial Context
Intent based Agentic Command Synthesis
Spatial Context Database
Large Language Model
Natural Language Command
<deviceName> <controlCommand>
Agentic Actuation and Execution Module
Control Scripts
Changes Reflected in the Environment

Figure 2

Annotation workflow

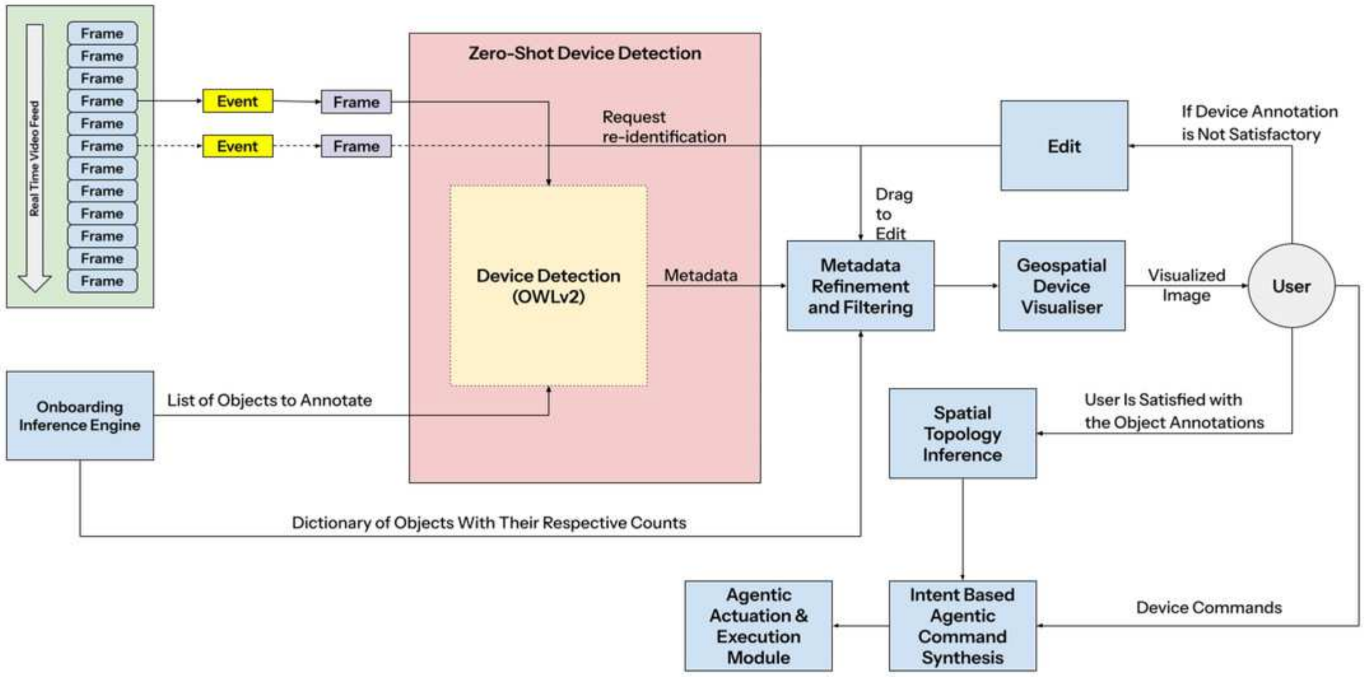
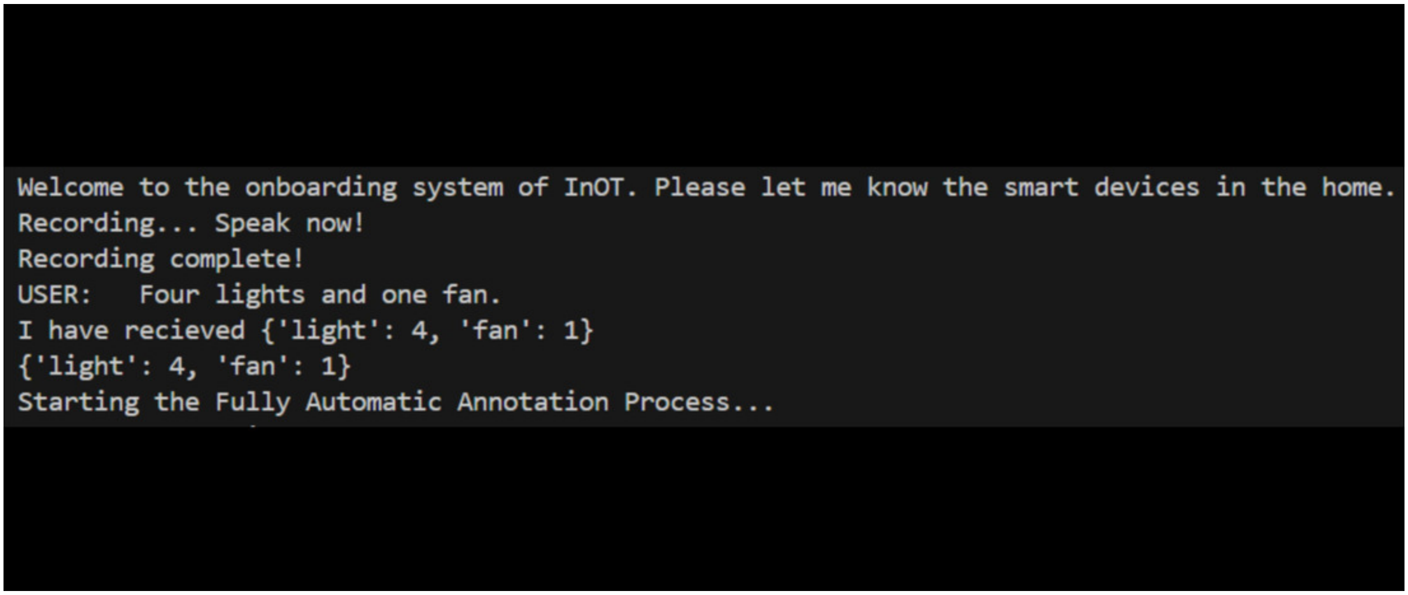


Figure 3

Devices onboarding



```
Welcome to the onboarding system of InOT. Please let me know the smart devices in the home.
Recording... Speak now!
Recording complete!
USER:   Four lights and one fan.
I have recieved {'light': 4, 'fan': 1}
{'light': 4, 'fan': 1}
Starting the Fully Automatic Annotation Process...
```

Figure 4

Initial environment

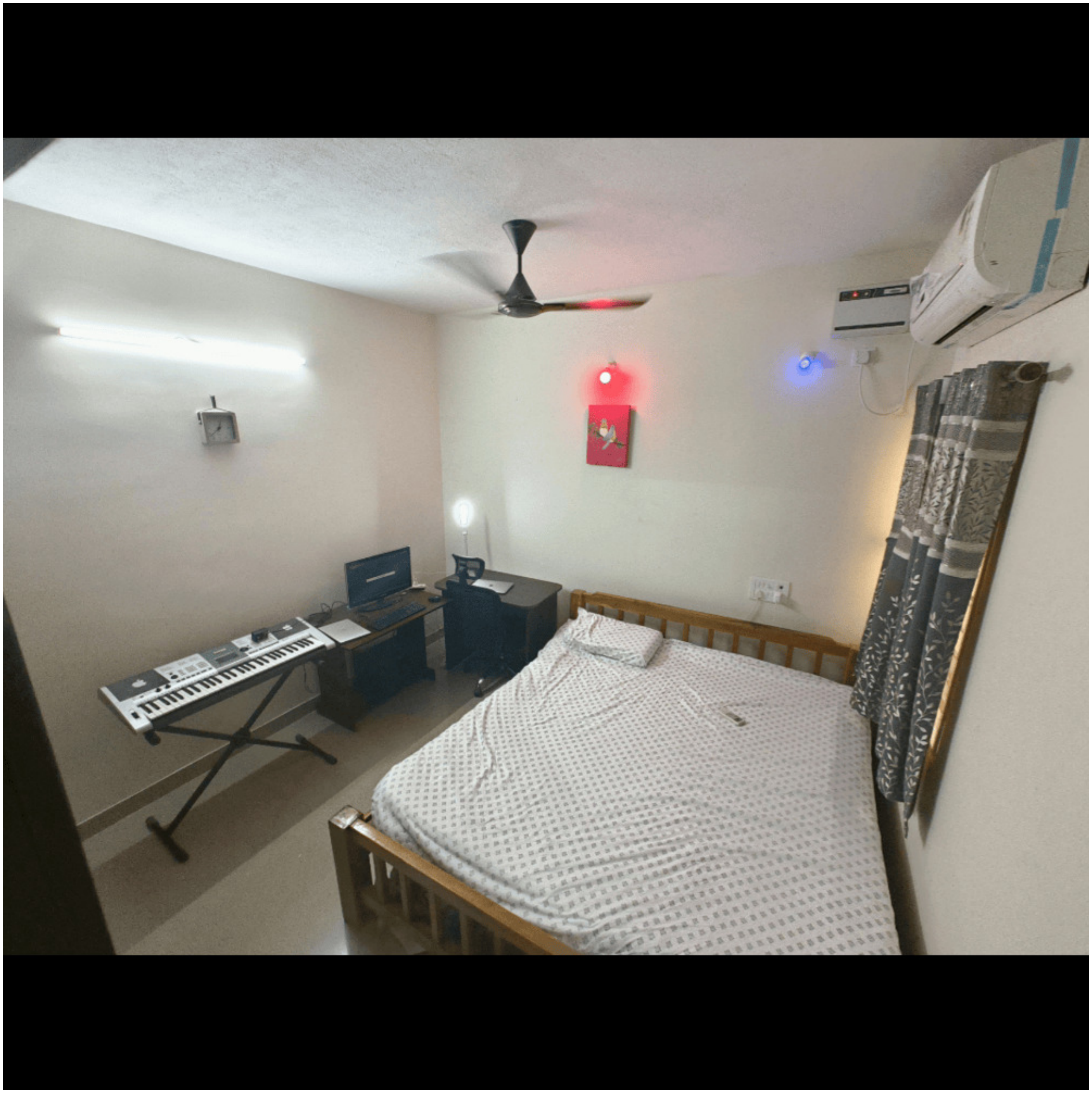


Figure 5

Device Detection

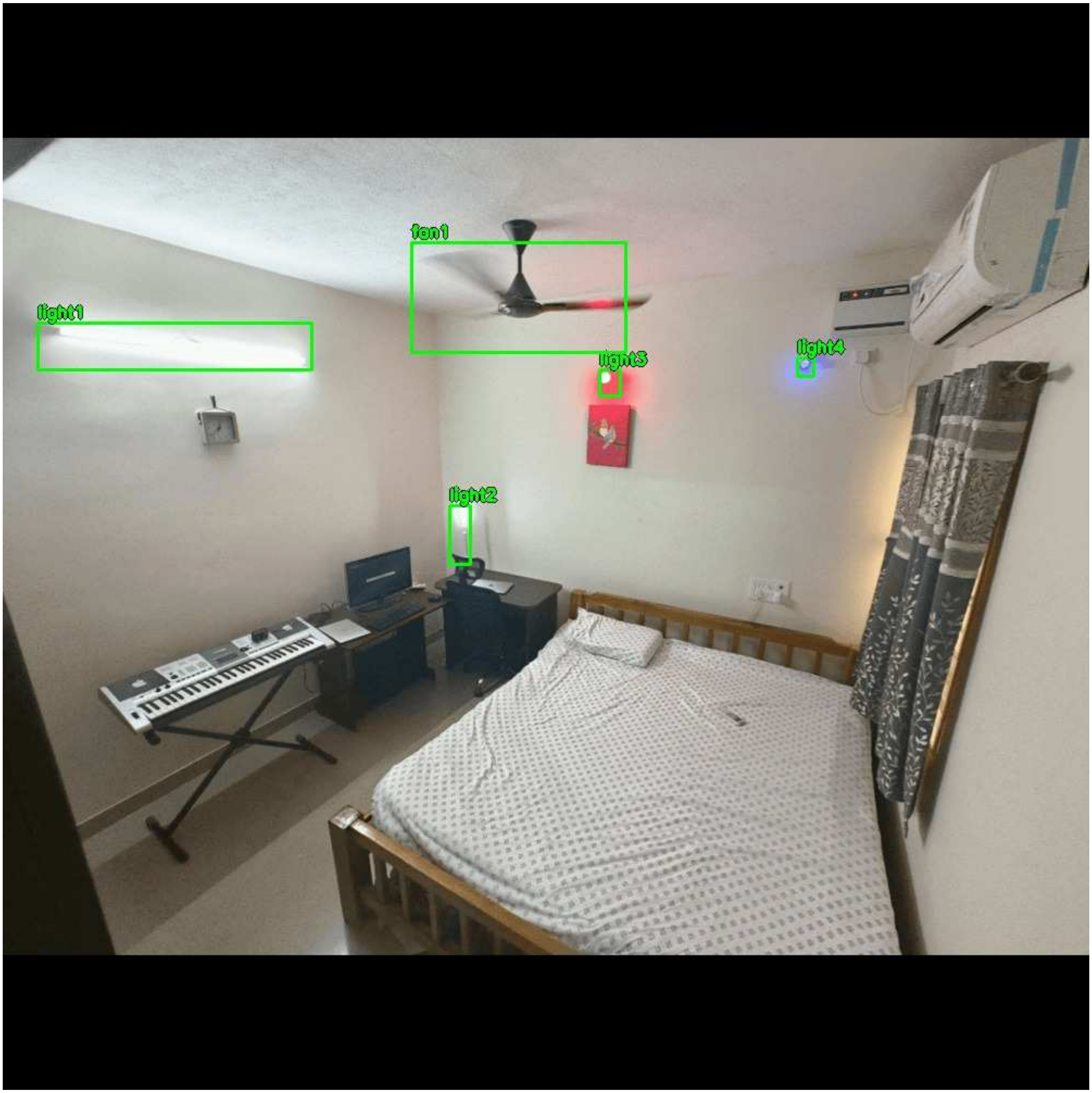


Figure 6

Metadata Refinement and Filtering

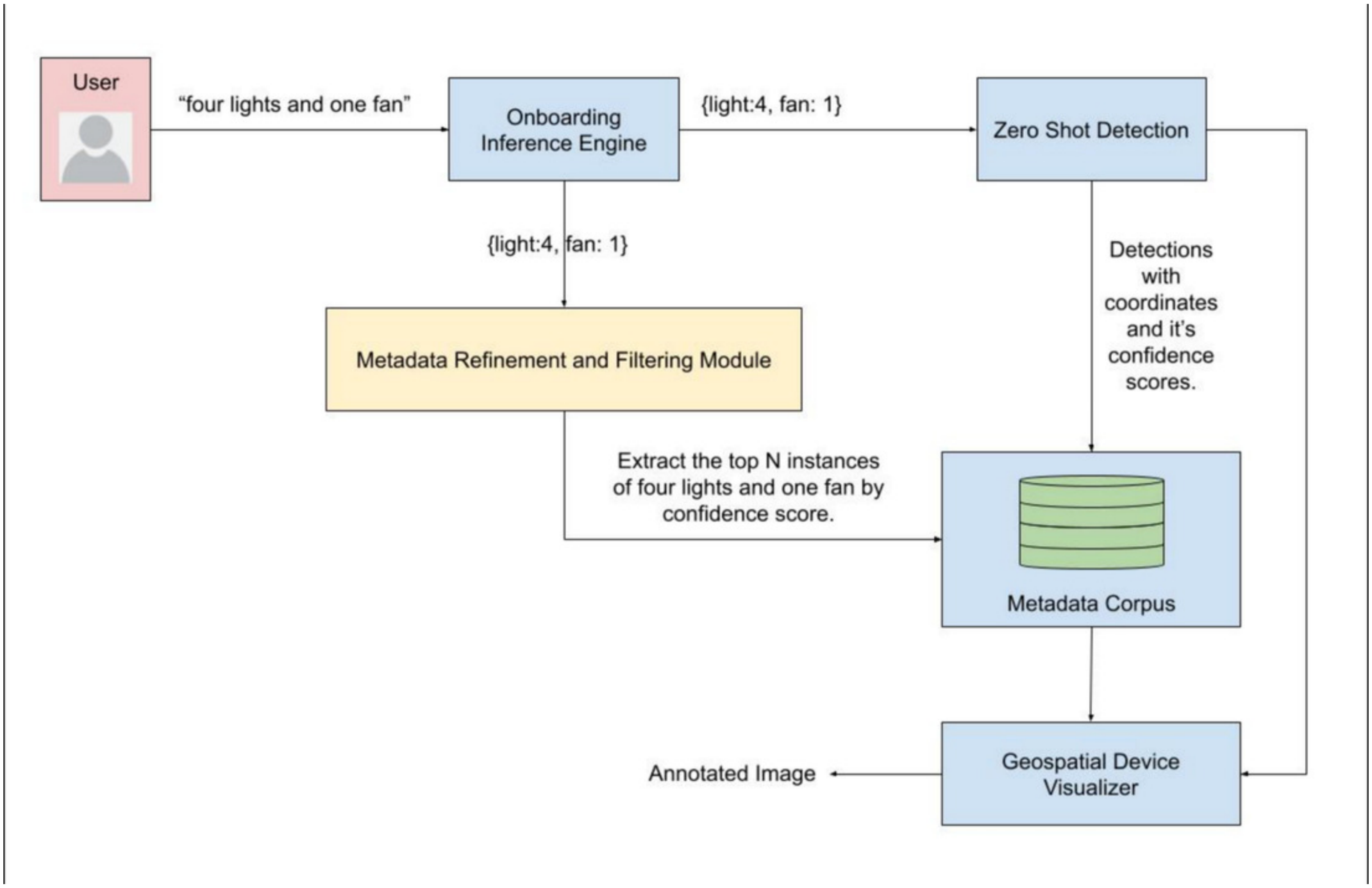
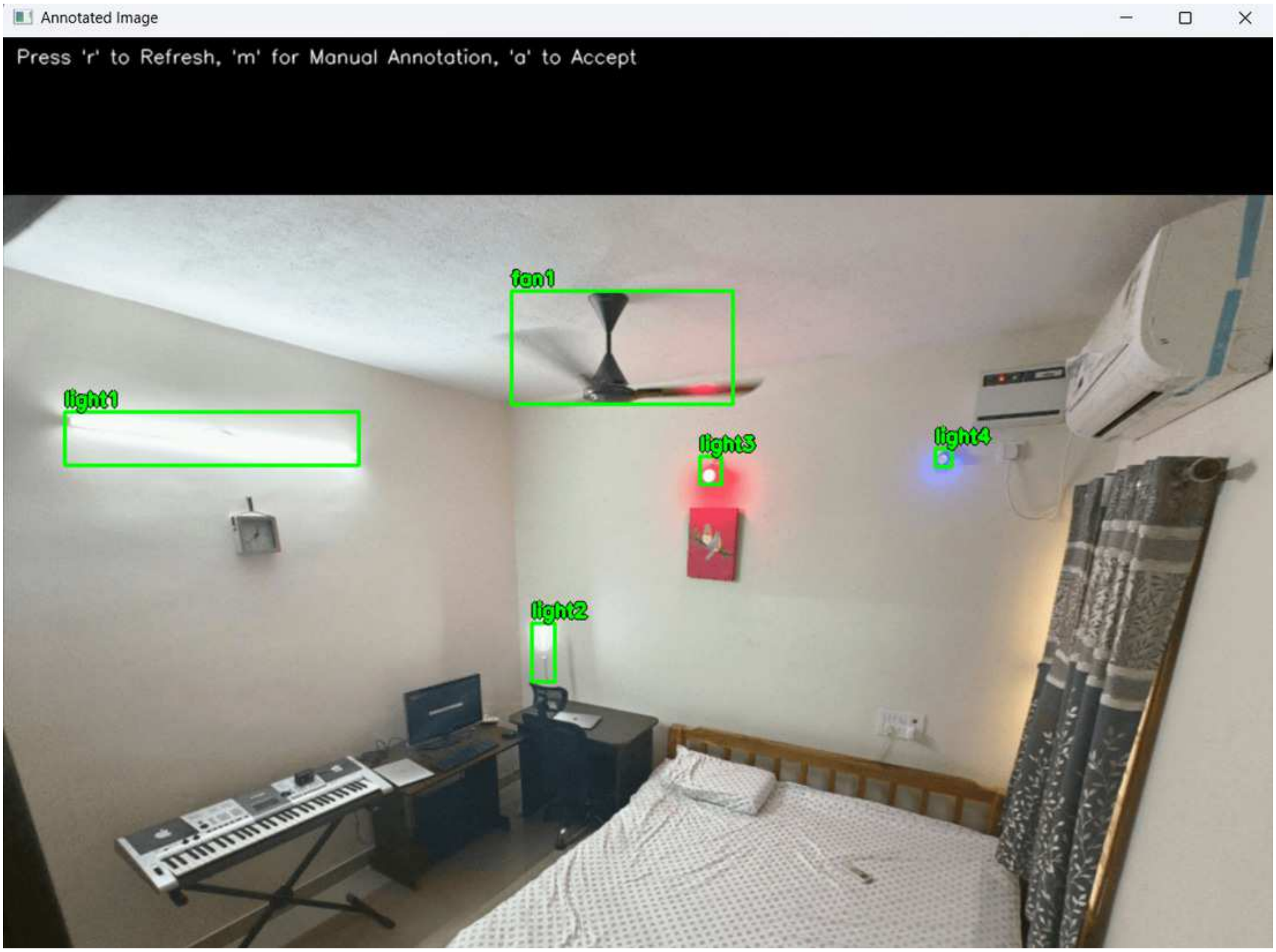

User
"four lights and one fan"
Onboarding Inference Engine
{light:4, fan: 1}
Zero Shot Detection
{light:4, fan: 1}
Metadata Refinement and Filtering Module
Detections with coordinates and it's confidence scores.
Extract the top N instances of four lights and one fan by confidence score.
Metadata Corpus
Annotated Image
Geospatial Device Visualizer

Figure 7

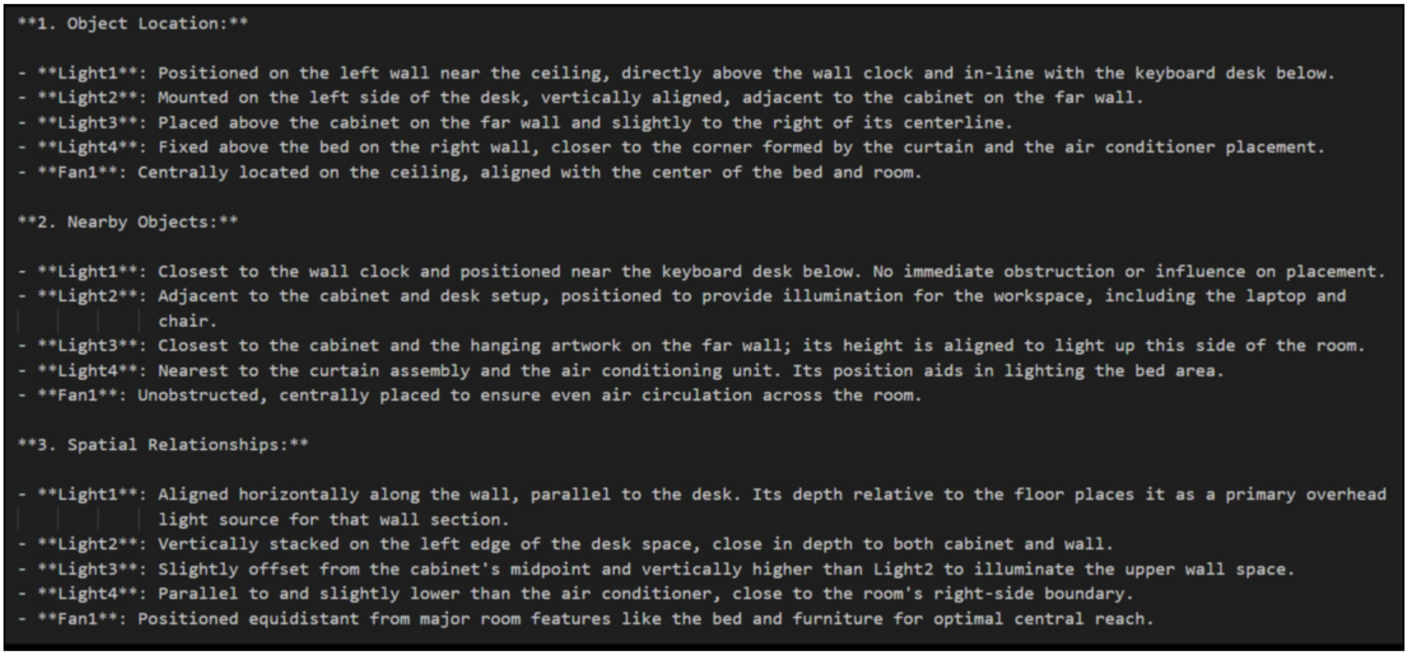
Geospatial Device Visualizer



Press 'r' to Refresh, 'm' for Manual Annotation, 'a' to Accept

Figure 8

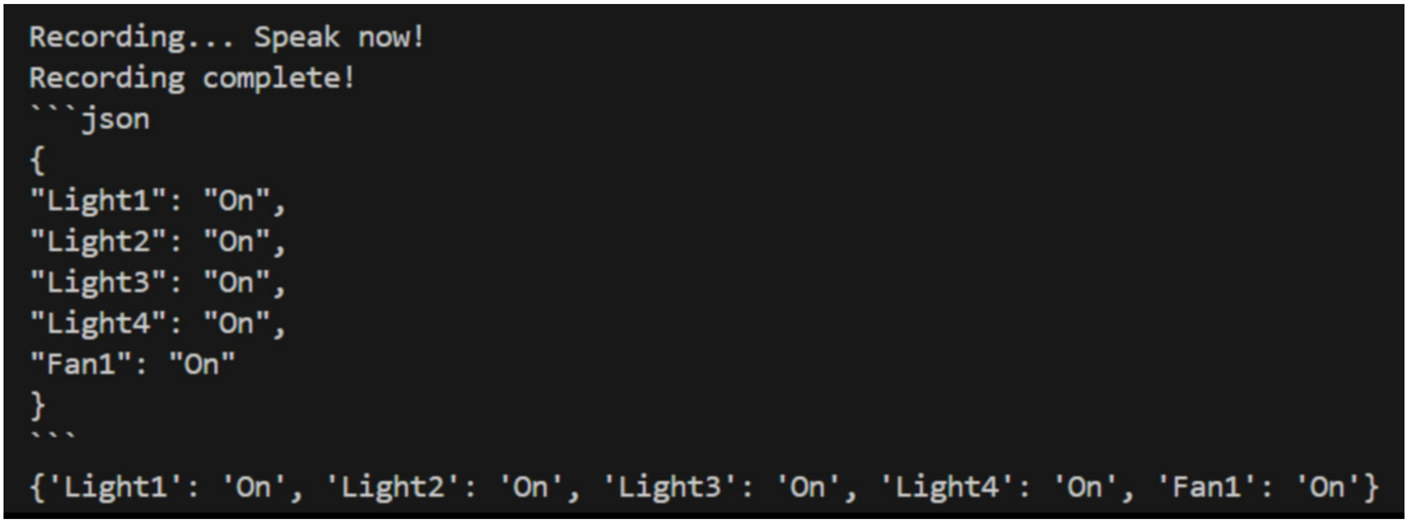
Spatial Topology Inference



```
**1. Object Location:**

- **Light1**: Positioned on the left wall near the ceiling, directly above the wall clock and in-line with the keyboard desk below.
- **Light2**: Mounted on the left side of the desk, vertically aligned, adjacent to the cabinet on the far wall.
- **Light3**: Placed above the cabinet on the far wall and slightly to the right of its centerline.
- **Light4**: Fixed above the bed on the right wall, closer to the corner formed by the curtain and the air conditioner placement.
- **Fan1**: Centrally located on the ceiling, aligned with the center of the bed and room.

**2. Nearby Objects:**

- **Light1**: Closest to the wall clock and positioned near the keyboard desk below. No immediate obstruction or influence on placement.
- **Light2**: Adjacent to the cabinet and desk setup, positioned to provide illumination for the workspace, including the laptop and
             chair.
- **Light3**: Closest to the cabinet and the hanging artwork on the far wall; its height is aligned to light up this side of the room.
- **Light4**: Nearest to the curtain assembly and the air conditioning unit. Its position aids in lighting the bed area.
- **Fan1**: Unobstructed, centrally placed to ensure even air circulation across the room.

**3. Spatial Relationships:**

- **Light1**: Aligned horizontally along the wall, parallel to the desk. Its depth relative to the floor places it as a primary overhead
             light source for that wall section.
- **Light2**: Vertically stacked on the left edge of the desk space, close in depth to both cabinet and wall.
- **Light3**: Slightly offset from the cabinet's midpoint and vertically higher than Light2 to illuminate the upper wall space.
- **Light4**: Parallel to and slightly lower than the air conditioner, close to the room's right-side boundary.
- **Fan1**: Positioned equidistant from major room features like the bed and furniture for optimal central reach.
```

Figure 9

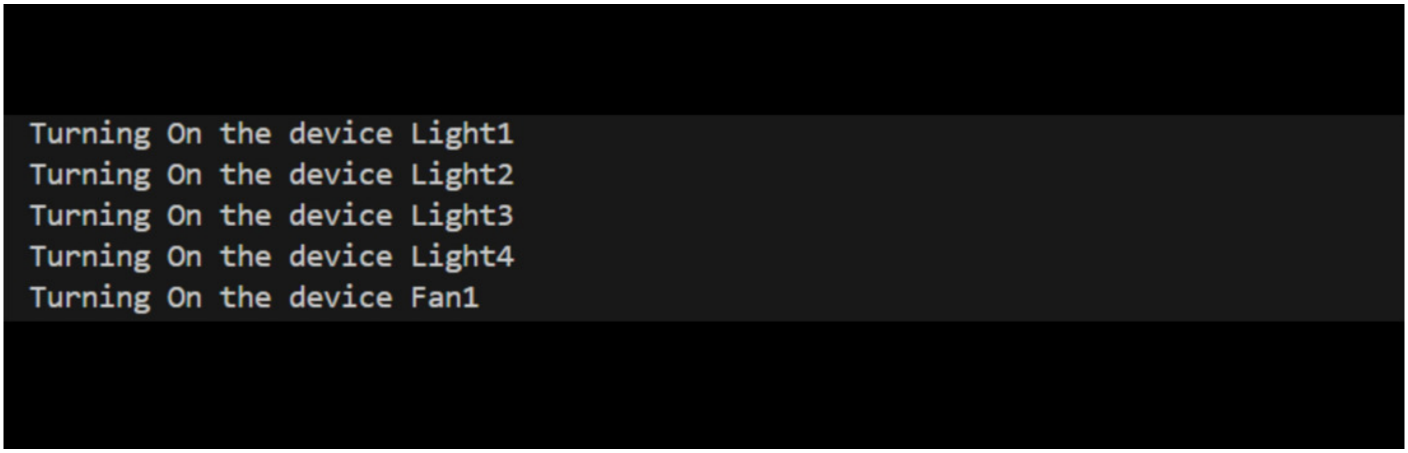
Intent-Based Agentic Command Synthesis



~~~
Recording... Speak now!
Recording complete!
```json
{
"Light1": "On",
"Light2": "On",
"Light3": "On",
"Light4": "On",
"Fan1": "On"
}
```
{'Light1': 'On', 'Light2': 'On', 'Light3': 'On', 'Light4': 'On', 'Fan1': 'On'}
~~~

Figure 10

Execution Module



```
Turning On the device Light1
Turning On the device Light2
Turning On the device Light3
Turning On the device Light4
Turning On the device Fan1
```

Figure 11

Completed Workflow

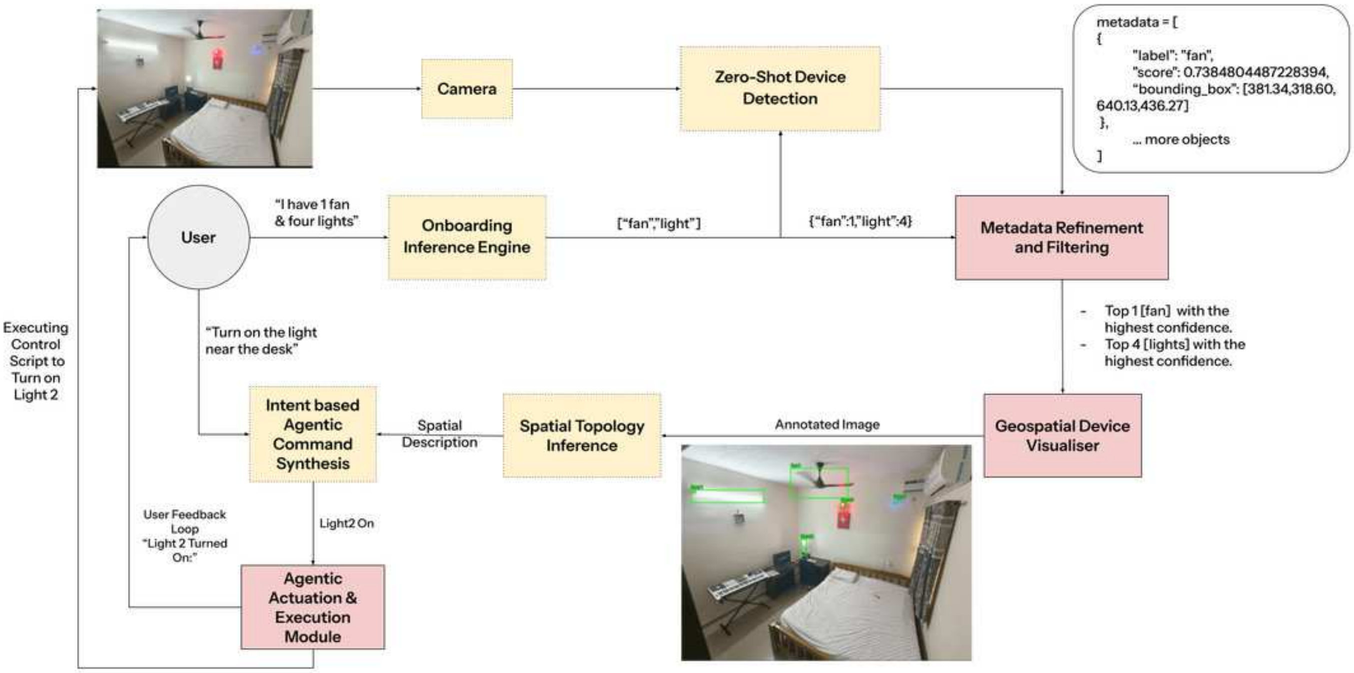

Camera
Zero-Shot Device Detection
metadata = [
{
"label": "fan",
"score": 0.7384804487228394,
"bounding_box": [381.34,318.60,
640.13,436.27]
},
... more objects
]
User
"I have 1 fan & four lights"
Onboarding Inference Engine
["fan","light"]
{"fan":1,"light":4}
Metadata Refinement and Filtering
- Top 1 [fan] with the highest confidence.
- Top 4 [lights] with the highest confidence.
Executing Control Script to Turn on Light 2
"Turn on the light near the desk"
Intent based Agentic Command Synthesis
Spatial Description
Spatial Topology Inference
Annotated Image
Geospatial Device Visualiser
User Feedback Loop "Light 2 Turned On:"
Light2 On
Agentic Actuation & Execution Module

Figure 12

Boxplot of TLX scores mean and standard deviation

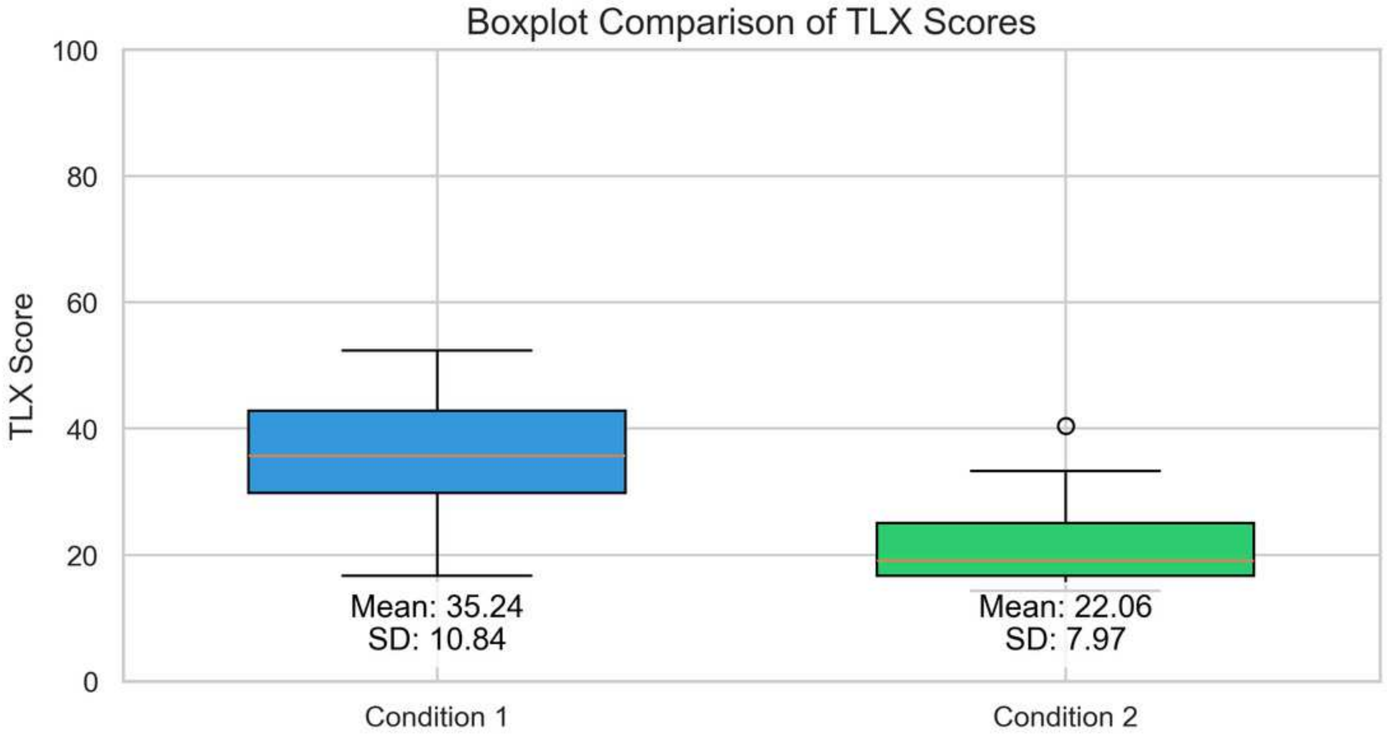


Figure 13

TLX scores Density distribution

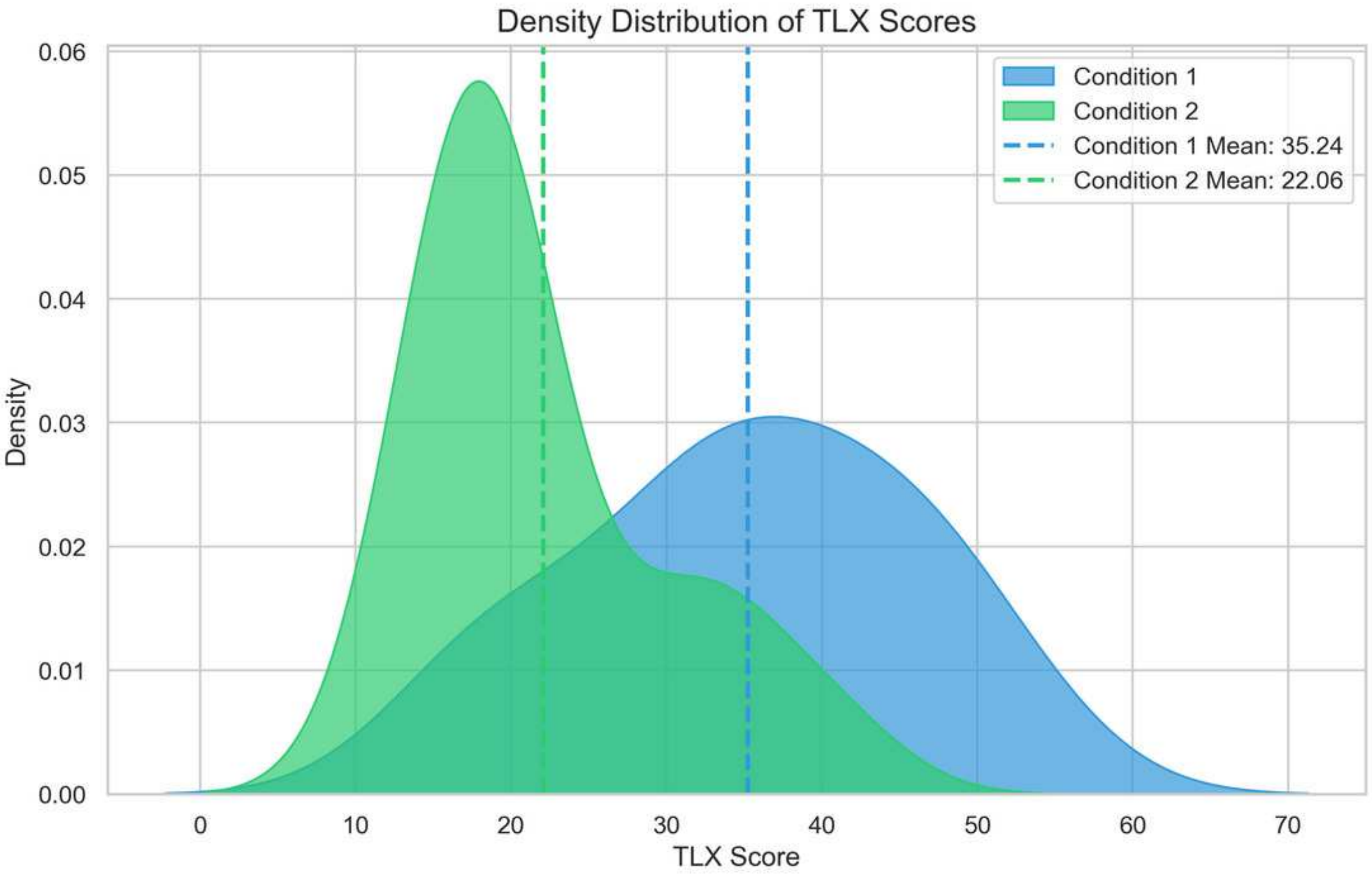
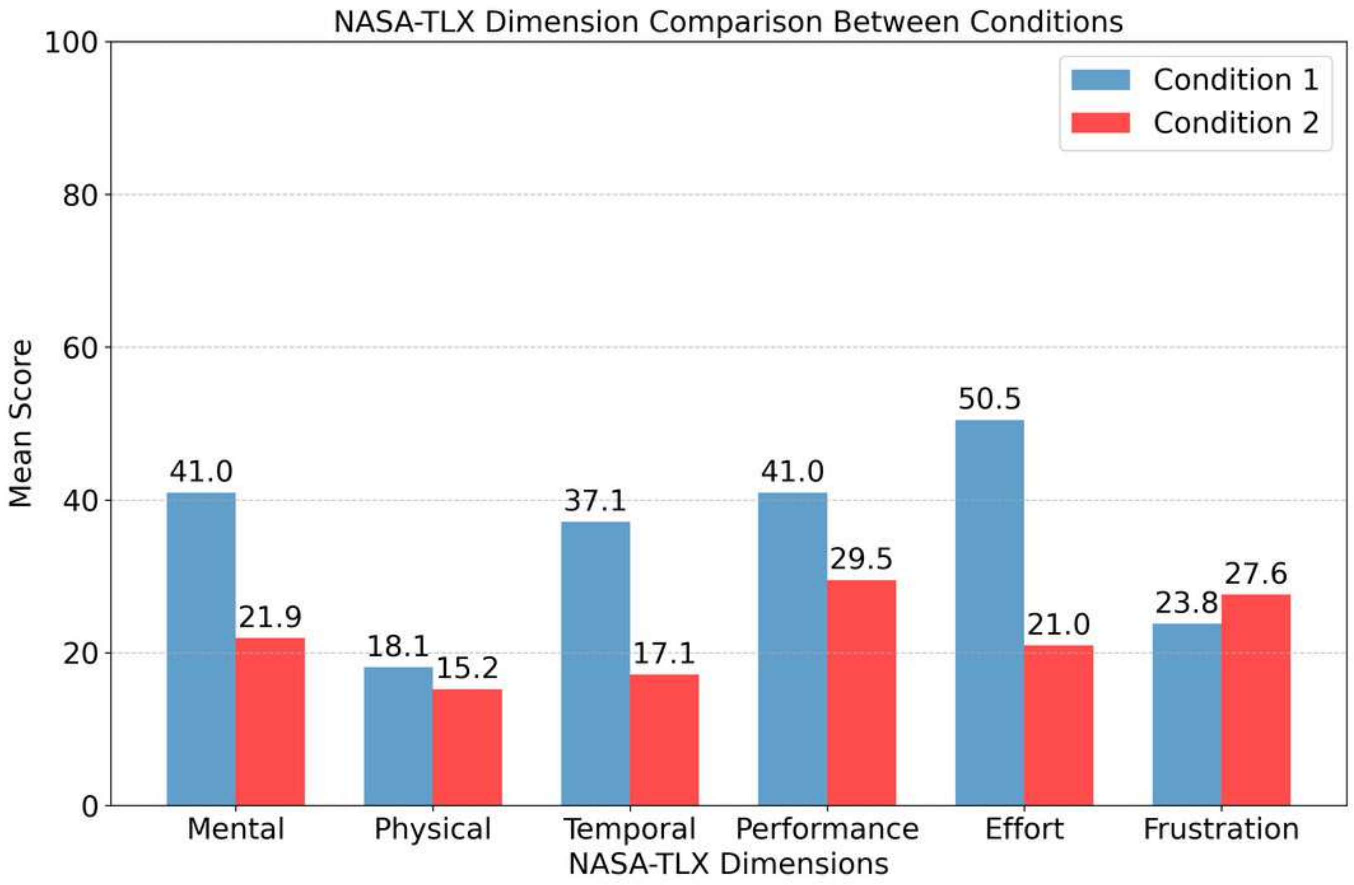


Figure 14

NASA TLX Dimensions


NASA-TLX Dimension Comparison Between Conditions
Condition 1
Condition 2
Mean Score
Mental
Physical
Temporal
Performance
Effort
Frustration
NASA-TLX Dimensions