Last-Meter Delivery: Solving the Unattended Delivery Challenge from Streets to Doorsteps

by

Wen-Xin Xiao

B.S., National Taipei University of Technology (2018)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2024

Authored by____

Wen-Xin Xiao
Program in Media Arts and Sciences
August 16, 2024

Certified by____

Kent Larson
Professor of the Practice
Massachusetts Institute of Technology
Thesis Supervisor

Accepted by____

Joseph Paradiso
Academic Head
Program in Media Arts and Sciences

Last-Meter Delivery: Solving the Unattended Delivery Challenge from Streets to Doorsteps

by

Wen-Xin Xiao

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 16, 2024, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

The rise of e-commerce has led to a surge in package deliveries, resulting in the proliferation of unattended delivery methods to address the "last-meter" problem – the challenge of delivering packages from the roadside or sidewalk to the customer's front door. This thesis proposes a methodology for implementing Large Language Model (LLM), and Vision Language Model (VLM) to enable delivery robots to identify the final delivery target and navigate the complex terrain from the curb to the front door. The proposed solution aims to enhance the autonomy and safety of last-mile delivery systems, addressing the "last-meter" challenge and improving the customer experience.

This thesis presents a comprehensive overview of the last-meter delivery concept, aiming to bridge the gap between the roadside/sidewalk and the customer's front door. It begins by introducing the significance of last-meter delivery in the growing e-commerce industry and the challenges posed by unattended deliveries. The thesis then reviews the existing literature on autonomous and unmanned delivery systems, multimodal delivery approaches, and the application of large language models and vision language models in robotics. This research identifies the advancements and gaps in the field that the proposed methodology aims to address.

The thesis primarily focuses on leveraging Large Language Models, the Segment Anything Model, and the open-source Florence-2 vision foundation model to enable the transmission of customers' delivery instructions to the final delivery target in the context of last-meter delivery. It outlines the methodology for data preparation, object detection and labeling, as well as the integration of Large Language Models to handle customer instructions and coordinate delivery target. It also describes the experimental design and methodologies employed to validate the effectiveness of the proposed system. This includes the use of a last-meter dataset and the evaluation of last-meter scene and target coordinate identification.

The thesis concludes by summarizing the key findings and contributions, discussing the broader implications of the proposed methodology, and suggesting directions for future work, such as enhancing system robustness and scalability.

KEYWORDS: Last-Mile Delivery, last-meter Delivery, Large Language Models (LLM), Vision Language Models (VLM), Robotics, Segment Anything Model (SAM), Open-Vocabulary Object Detection (OVD).

Thesis Supervisor: Kent Larson
Title: Professor of the Practice, Massachusetts Institute of Technology

Last-Meter Delivery: Solving the Unattended Delivery Challenge from Streets to Doorsteps

by

Wen-Xin Xiao

This thesis has been reviewed and approved by the following committee members:

Thesis Supervisor________________________________

Kent Larson
Professor of the Practice
Massachusetts Institute of Technology

Thesis Reader________________________________

Luis Alberto Alonso Pastor
Research Scientist
Massachusetts Institute of Technology

Thesis Reader________________________________

Sekimoto Yoshihide
Professor of Center for Spatial Information Science
University of Tokyo

Thesis Reader________________________________

Kevin Esvelt
Associate Professor
Massachusetts Institute of Technology

Contents

Abstract **3**

1 Introduction **13**

2 Related Works **17**
2.1 Autonomous and Unmanned Delivery Systems 18
2.2 Multimodal Delivery Systems 20
2.3 Large Language Models and Vision Language Models 21
2.4 Human-Robot Interaction and Safety 22
2.5 Sustainability 22

3 Methodology **23**
3.1 Data Preparation 24
3.2 Object Detection and Labeling 25
3.2.1 Segment Anything Model 2 (SAM2) 26
3.2.2 Florence-2 Open-Vocabulary Object Detection 27
3.3 Processing Customer Instructions by LLM 27
3.3.1 Chain-of-Tought (CoT) Prompting 28

4 Experiment Results and Discussion **30**
4.1 Last-Meter Data Collection 31
4.2 Last-Meter Scene Analysis 32
4.2.1 Evaluation Method 32
4.2.2 Results 33
4.3 Target Coordinate Identification by Chain-of-Thought Prompting 35
4.3.1 Evaluation Method 35
4.3.2 Results 36
4.4 Discussion 38

5 Conclusion **41**
5.1 Contribution 41
5.2 Limitation 42
5.3 Potential 43
5.4 Future Work 44

Acknowledgments 46

A Object Detection Output 47

B Code of LLM Chain-of-Though 49

Bibliography 52

List of Figures

1-1 Definition of Last-Meter 14
1-2 Comparison of Delivery Methods by Number of Packages and Miles Traveled 16

3-1 Conceptual Model of the Last-Mile and Last-Meter Delivery Ecosystem . . 24
3-2 The SAM 2 architecture, Image Source: Ravi et al. [32] 26
3-3 Florence-2 data engine, Image Source: Xiao et al. [51] 27
3-4 Standard Input-Output Prompt v.s. CoT Prompt, Image Source: Wei et al. [49] 29

4-1 Experiment Overview 30
4-2 Last-Meter Dataset 31
4-3 Coordinate Output from GPT4-o 37

List of Tables

4.1 Comparison of IoU between Florence-2 and SAM2 + Florence-2 for various objects in different locations. 33

4.2 Ground Truth Coordinates and Coverage Rate Comparison 36

A.1 Comparison of IoU between Florence-2 and SAM2 + Florence-2 at the dorm of the last-meter dataset 47

A.2 Comparison of IoU between Florence-2 and SAM2 + Florence-2 at the house#1 of the last-meter dataset 48

A.3 Comparison of IoU between Florence-2 and SAM2 + Florence-2 at the house#2 of the last-meter dataset 48

Chapter 1

Introduction

The rapid growth in e-commerce and online shopping has sparked a substantial and widespread increase in parcel deliveries. For instance, in Sweden, e-commerce sales grew by an average of 18% per year between 2005 and 2019, with a further 40% increase in 2020 during the COVID-19 pandemic [43]. This expansion has led to heightened challenges for the last stage of the delivery process, known as last-mile logistics [4]. According to Bosona [7], last-mile delivery refers to the final transportation segment of a supply chain. Last-mile delivery is often the most inefficient and costly, accounting for up to 50% of the total delivery expense [35]. Changes in consumer shopping behaviors and rising urban population have made last-mile delivery particularly problematic in urban areas, leading to issues with congestion, delivery times, and sustainability [21]. This challenge is especially pronounced in highly urbanized regions with a substantial share of e-commerce, such as Europe and the United States. Further, innovative transport modes and systems are rapidly changing the conditions for last-mile deliveries, particularly with regard to unattended delivery [24].

At the same time, customers demand faster, more predictable, and more flexible deliveries [40] [43]. This results in last-mile delivery being an expensive and inefficient component of the supply chain [31]. Concurrently, technology is undergoing significant advancements in digitalization and autonomous vehicles, with substantial ongoing research [12]. Autonomous vehicles are increasingly becoming commercially available and expanding into new domains.

Automated guided vehicles have been commonplace in warehouses for decades [45] and have more recently emerged on sidewalks in the form of autonomous delivery robots. The externalities, coupled with the rise in e-commerce and urbanization mentioned above, have created pressure to find innovative solutions to enhance the efficiency of last-mile deliveries. One approach to addressing this problem could be the utilization of new advanced technologies to develop automated delivery systems, potentially leading to a more efficient and sustainable last-mile delivery system [39].

To address these challenges, the concept of "last-meter" delivery, which aims to bridge the gap between the roadside/sidewalk and the customer's front door, has gained significant attention. This approach seeks to optimize the efficiency and sustainability of the final delivery by exploring innovative solutions that can overcome the obstacles posed by unattended deliveries. Researchers have explored various last-mile delivery concepts, including the use of unmanned aerial vehicles (UAVs) [5] [10] [34] and autonomous delivery robots (ADRs) [1] [2] [3], which have the potential to reduce delivery time, increase efficiency, and cut costs. However, most unattended delivery robots still require human involvement to retrieve packages. How to fully automate the last-meter delivery process in a way that ensures the safe and secure delivery of packages to customers' doorsteps remains a critical challenge that requires further research and technological advancements [6] [8].

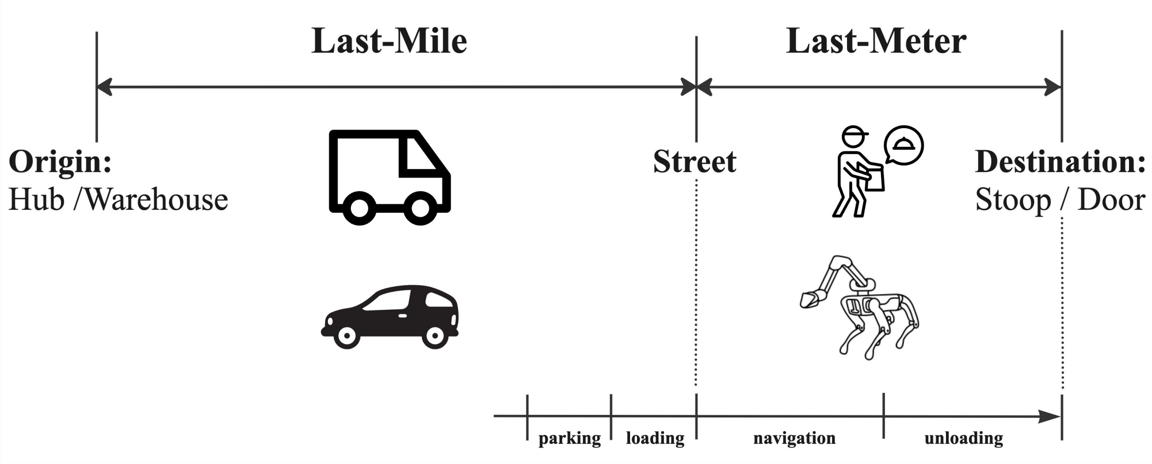


Figure 1-1: Definition of Last-Meter

In this context, the concept of a multimodal autonomous last-mile delivery system has emerged as a promising approach. Land vehicles transport parcels to the curb, and a smaller robot then delivers them to the customer's front door, enabling a seamless transition between different modes of transportation and reducing the overall delivery time. The last-mile modes should be equipped with basic sensors commonly found in autonomous vehicles [16], such as cameras, LiDAR, radars, and GPS, to navigate and detect obstacles. Furthermore, integrating advanced communication technologies can enhance the coordination between these vehicles and the surrounding infrastructure, ultimately leading to a more streamlined and efficient delivery process [38]. The last-meter modes will depart from the last-mile modes to safely approach and deposit the packages at the customer's doorstep. This requires the implementation of communication and charging systems within the last-mile and last-meter delivery modes. The last-meter delivery robots must be equipped with their own sensors, such as LiDAR and radars, to detect unexpected obstacles or environmental changes as they move, as well as advanced computer vision and object detection capabilities to precisely locate the customer's door and safely drop the package. Once the last-meter delivery robot reaches the final destination, it should be able to autonomously complete the package delivery without any human intervention, send a confirmation signal and an image indicating the package's arrival back to the last-mile modes or the central system, marking the delivery as complete, and then return to the last-mile mode. As many e-commerce platforms now offer customers the option to decide where to have their packages dropped off, this enhances customer convenience and flexibility but also increases the delivery complexity and challenge. This data can be fed into the last-meter robots to better locate and identify each customer's preferred delivery location. This paper examines the potential of last-meter delivery in addressing the unattended delivery problem, drawing from the latest research and industry developments. The aim of this thesis is to leverage Large Language Models, the Segment Anything Model, and the open-source Florence-2 vision foundation model to enable the transmission of customers' delivery instructions to the final delivery target in the context of last-meter delivery.

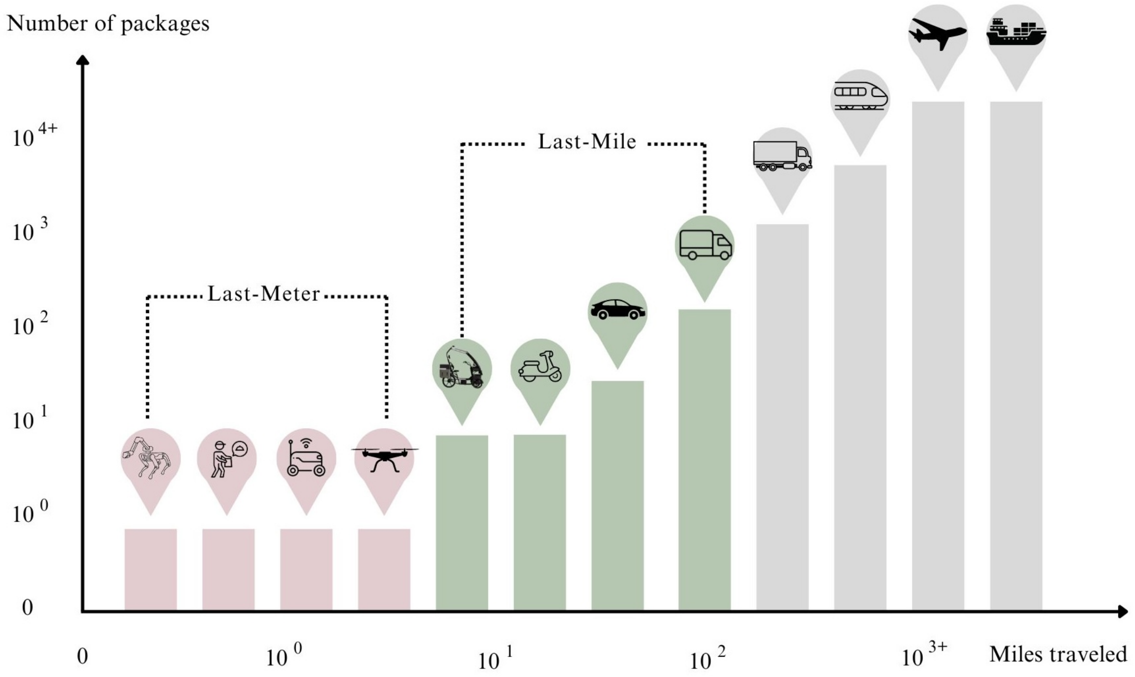


Figure 1-2: Comparison of Delivery Methods by Number of Packages and Miles Traveled

Chapter 2

Related Works

This section presents a review of the existing scholarly literature on the challenges associated with last-mile delivery, as well as the development of autonomous delivery systems and multimodal delivery approaches. It explores the potential benefits of employing multimodal solutions that leverage Large Language Models and Vision Language Models to enhance the efficiency and reliability of last-mile delivery systems.

Additionally, this section delves into the critical importance of human-robot interaction and safety considerations in urban settings. It highlights the growing emphasis on sustainability and environmental factors in the context of last-mile delivery, discussing how innovative solutions can contribute to reducing greenhouse gas emissions and mitigating the environmental impact of the growing e-commerce industry.

2.1 Autonomous and Unmanned Delivery Systems

The unattended delivery problem has garnered substantial attention in recent years, with researchers and industry professionals exploring various solutions to enhance the efficiency and reliability of last-mile delivery systems. The most common approaches include the use of parcel lockers [29] [52], smart mailboxes [44] [41], and secure delivery boxes [23] [28], which allow for the safe storage of packages until customers can retrieve them at their convenience [46]. These solutions, however, do not fully address the “last-meter” problem, as they often require customers to travel to a designated location to collect their packages.

To tackle the unattended delivery problem more comprehensively, autonomous delivery robots (ADRs) and unmanned aerial vehicles (UAVs) have emerged as promising solutions. ADRs are designed to navigate urban environments and deliver packages directly to customers’ doorsteps. Researchers use simulators to navigate on sidewalks, campuses, and other urban scenarios, avoiding real-world damage from algorithm failures or programming errors. This accelerates development and reduces costs [42]. Recent advancements in robotics, artificial intelligence, and computer vision have significantly improved the capabilities of ADRs, enabling them to operate with minimal human intervention. Companies such as Starship Technologies and Nuro have successfully deployed ADRs in various urban settings, demonstrating the potential for these robots to enhance last-mile delivery efficiency [18]. However, these approaches primarily focus on wheeled robots that can only navigate roads and sidewalks, and are not capable of traveling up to the customer’s doorstep. This requires customers to retrieve their packages from the curb or roadside areas, which can be inconvenient and may not fully address the unattended delivery problem.

In parallel, unmanned aerial vehicles (UAVs), or drones, represent another innovative approach to addressing the unattended delivery problem. Drones offer the advantage of bypassing ground traffic, potentially reducing delivery times and costs [20]. Companies like Amazon and Google have invested heavily in developing drone delivery systems, with pilot programs demonstrating their feasibility in specific use cases. However, regulatory challenges, safety concerns, and technical limitations, such as battery life and payload capacity, continue to hinder the widespread adoption of UAVs for last-mile delivery.

2.2 Multimodal Delivery Systems

The integration of various delivery methods, including ADRs, UAVs, and traditional delivery vehicles, into a cohesive multimodal delivery system has the potential to address the limitations of individual approaches. Multimodal delivery systems leverage the strengths of each delivery method to optimize efficiency and reduce costs [37]. For example, ADRs can handle deliveries in densely populated urban areas, while UAVs can service remote or difficult-to-access locations. The development of intelligent logistics platforms that can coordinate these different delivery modes is crucial for the success of multimodal delivery systems.

Prior studies have explored the concept of an unmanned vehicle-robot system, where a self-driving vehicle acts as a mothership transporting multiple autonomous robots [25]. In this approach, the self-driving vehicle travels from a central depot to various dispersed stations, and the onboard robots are then deployed to handle the individual pickup and delivery tasks. Furthermore, some researchers have investigated the integration of autonomous vehicles with drone-based solutions to fully automate the last-mile delivery process [17].

Zhu et al. [54] have proposed and investigated the feasibility of a multi-agent modular robotic delivery system. This innovative system can latch onto a single package, functioning as a legged robotic system where the package itself serves as the body. This research aims to address the inefficiencies often encountered with the use of traditional legged and humanoid robots, which tend to occupy large amounts of space and limit scalability and task parallelization, thereby reducing the overall package carrying capacity of delivery vehicles. By utilizing a modular design where the package acts as the body, this system can potentially achieve greater flexibility, improved space utilization, and enhanced task parallelization capabilities, ultimately enhancing the efficiency and scalability of last-mile delivery operations.

2.3 Large Language Models and Vision Language Models

Recent advancements in computer vision and machine learning have paved the way for more sophisticated localization and navigation systems for delivery robots. Vision-based systems enable robots to accurately identify delivery targets and navigate complex urban environments. By leveraging large language models (LLMs) and vision language models(VLMs), researchers have developed algorithms that allow robots to interpret visual data, recognize landmarks, and plan efficient delivery routes [27]. These technologies are essential for overcoming the “last-meter” challenge, as they enable robots to autonomously navigate the final few meters to the customer’s doorstep.

Researchers have also explored the use of Large Language Models (LLMs) and other multi-agent AI systems to enable robot to adapt and navigate complex environments. Gu et al. [13] use depth-camera and semantic segmentation to help robots identify objects, understand the spatial context and identify target locations within indoor environments. Furthermore, studies [15] [26] [48] [53] have demonstrated the potential of LLMs in facilitating natural language interaction between humans, robots, and enviroments. This has enabled advancements in interactive object search tasks involving mobile manipulation [14] as well as scalable approaches to robot task planning [30]. However, most existing research in this space has focused on laboratory or controlled environments, an important next step is to validate the performance and robustness of these AI-powered systems in real-world settings.

Moreover, Wei et al. [49] demonstrated that sufficiently large large language models can perform similar tasks to the examples provided, without requiring any adjustments to the model weights. The concept of Chain-of-Thought prompting, which involves providing the model with a series of examples, can be conveniently implemented to leverage these capabilities.

2.4 Human-Robot Interaction and Safety

Ensuring the safe and secure operation of delivery robots in urban environments is a critical aspect of last-meter delivery research. Human-robot interaction (HRI) studies focus on understanding how robots can interact safely and effectively with pedestrians [36], cyclists, and other road users [50] [19]. Researchers are developing advanced sensors, collision avoidance systems, and communication protocols to enhance the safety and reliability of ADRs. Additionally, regulatory frameworks and industry standards are being established to govern the deployment and operation of delivery robots.

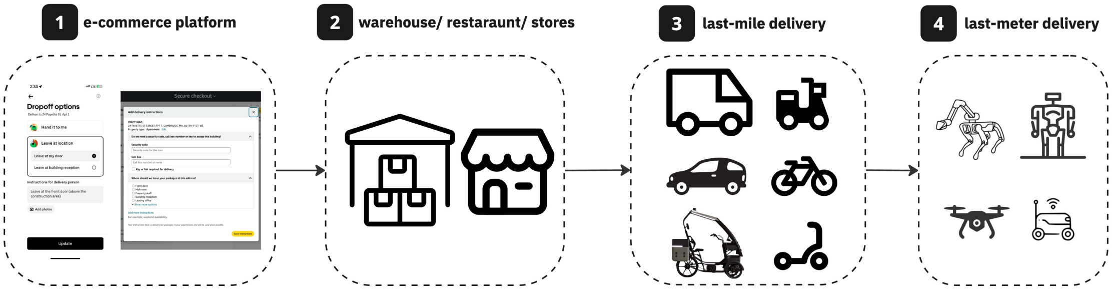
2.5 Sustainability

The field of last-mile delivery is progressing rapidly, with a growing emphasis on sustainability and environmental considerations. Recent research has indicated emerging trends and future directions in this area, highlighting the importance of adopting more eco-friendly delivery solutions [21]. Compared to traditional delivery methods relying on fossil-fuel-powered vehicles, autonomous delivery robots and unmanned aerial vehicles have the potential to significantly reduce energy consumption and greenhouse gas emissions [11]. By decreasing the number of traditional delivery vehicles on the road, these autonomous systems can contribute to decreased traffic congestion and improved air quality in urban areas, making last-mile delivery more sustainable. Additionally, the integration of various delivery modes, such as ADRs and UAVs, into a cohesive multimodal delivery system can further optimize efficiency and reduce environmental impact.

Chapter 3

Methodology

The methodology aims to leverage the Segment Anything Model [32] and Florence-2 [51] for scene analysis in outdoor settings, and employ large language models to interpret customer instructions from e-commerce platforms. This information is then used to coordinate the delivery target, and for the further utilization of quadrupedal robots to navigate the diverse terrain encountered during the last-meter delivery scenario, ultimately delivering the packages. The methodology involves several core components: assembling a comprehensive dataset for the last-mile delivery scenario, utilizing SAM2 [32] and Florence-2 [51] to segment images and analyze outdoor scenes, and using large language models and vision-language models for processing customer instructions to pinpoint delivery destinations, and finally coordinating the target delivery location for navigation of quadrupedal robots to the final drop-off locations. The conceptual model of the last-mile and last-meter delivery ecosystem depicted in Figure 3-1 encompasses an initial state characterized by package information, such as product details, dimensions, warehouse location, delivery address, and customer-provided delivery instructions. This information is then communicated to relevant logistics entities, including warehouses, restaurants, or merchandisers. These entities subsequently allocate the most suitable last-mile delivery mode, such as vans, motorcycles, vehicles, bikes, Persuasive Electric Vehicles, and scooters, as well as the appropriate last-meter delivery mode, including quadrupedal, humanoid, drone, or delivery robot, based on factors like capacity, travel distance, terrain capability, and autonomy level.



2:33
Dropoff options
Deliver to 24 Fayette St. Apt 1
Hand it to me
Leave at location
Leave at my door ●
Leave at building reception ○
Instructions for delivery person
Leave at the front door (above the construction area)
Add photos
Update

Secure checkout
Add delivery instructions
VINCY XIAO
24 FAYETTE ST STREET APT 1, CAMBRIDGE, MA, 02139-1157, US
Property type: Apartment Edit
Do we need a security code, call box number or key to access this building?
Security code
Security code for the door
Call box
Call box number or name
□ Key or fob required for delivery
Where should we leave your packages at this address?
○ Front door
○ Mailroom
○ Property staff
○ Building reception
○ Leasing office
Show more options
Add more instructions
For example, weekend availability
Your instructions help us deliver your packages to your expectations and will be used when possible.
Save instructions

Figure 3-1: Conceptual Model of the Last-Mile and Last-Meter Delivery Ecosystem

3.1 Data Preparation

The initial stage of the study focused on procuring outdoor data through the use of consumer-grade devices. Given the absence of a standardized dataset tailored to the last-meter delivery domain, I employed an Apple iPad Pro equipped with LiDAR sensors. This LiDAR technology enabled the acquisition of high-resolution, precise three-dimensional point cloud data, which was essential for constructing detailed and reliable maps of the delivery environment. Furthermore, the research utilized Polycam, a free drone photogrammetry application, to efficiently generate high-quality three-dimensional models from photographs captured with any iPhone or iPad. This enabled the rapid creation of detailed scans of the delivery environments using LiDAR technology, as well as the capture of comprehensive 360-degree imagery.

This data collection will predominantly focus on residential areas, as the study is centered on the last-meter delivery domain. Additionally, a publicly available last-meter outdoor dataset will be shared with the research community to encourage further exploration and advancement in this field.

- **Hardware**: Apple iPad Pro (6th generation) with iPadOS 16.0
- **Application**: Polycam version 3.5.7

3.2 Object Detection and Labeling

Accurate object detection and recognition is a critical component in addressing the last-meter delivery challenge. This is because real-world residential environments present significant variability and complexity that autonomous delivery robots must navigate. In contrast to standardized delivery points like parcel lockers, residential areas feature diverse doorstep characteristics, such as stairs, plants, and other obstacles. By employing robust object recognition capabilities, ADRs can accurately identify these environmental features and adapt their actions accordingly. This ensures precise and reliable package placement, even in the absence of customer presence.

Following the collection of last-meter delivery data, the subsequent step is to leverage the Segment Anything Model 2 to segment and classify various elements within the environment, such as buildings, trees, sidewalks, and parked vehicles. Additionally, this research employs the Florence-2, an open-vocabulary object detection model, to label the segmented objects and provide their spatial coordinates, enabling a comprehensive scene analysis of the delivery environment.

The Florence-2 model exhibits the flexibility of vision-language models employed in open-vocabulary object detection. These models can recognize a diverse range of objects without necessitating extensive retraining for specific items yet maintain the ability to be fine-tuned using customized datasets, unlike conventional approaches like YOLO [33] [47], which would require training on a custom dataset tailored to the last-meter delivery domain. The training process for such models involves manually segmenting and labeling numerous objects, which is a time-intensive endeavor and restricts the models to only the trained objects, thus reducing their versatility compared to the open-vocabulary capabilities of the Florence-2 model.

This comprehensive scene understanding, achieved through the integration of SAM2 and Florence-2, will contribute to the development of robust navigation strategies for the quadrupedal robots as they navigate the diverse and dynamic last-meter delivery setting.

3.2.1 Segment Anything Model 2 (SAM2)

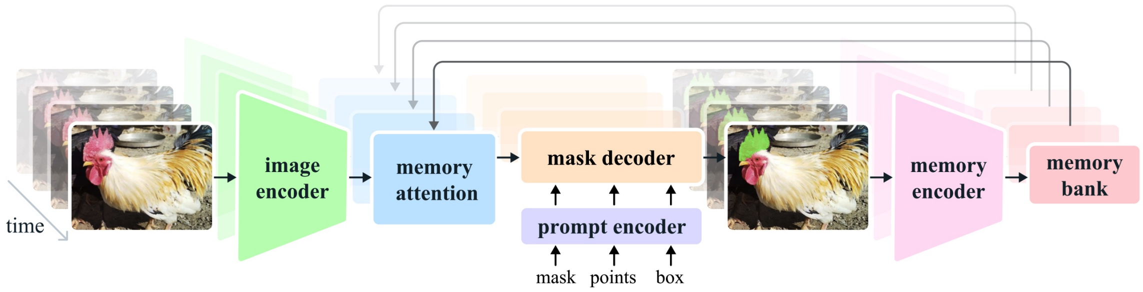


Figure 3-2: The SAM 2 architecture, Image Source: Ravi et al. [32]

The Segment Anything Model 2 is a foundation model designed to address the challenge of promptable visual segmentation in images and videos. It is capable of segmenting and classifying various elements within the environment, such as buildings, trees, sidewalks, and parked vehicles. The model employs a simple transformer architecture with streaming memory, enabling real-time video processing. According to the research, SAM 2, trained on a comprehensive dataset, demonstrates strong performance across a wide range of tasks. Specifically, in video segmentation, the model exhibits improved accuracy, requiring 3 times fewer interactions than previous approaches. Furthermore, in image segmentation, SAM 2 is found to be more accurate and 6 times faster than the original Segment Anything Model [22].

3.2.2 Florence-2 Open-Vocabulary Object Detection

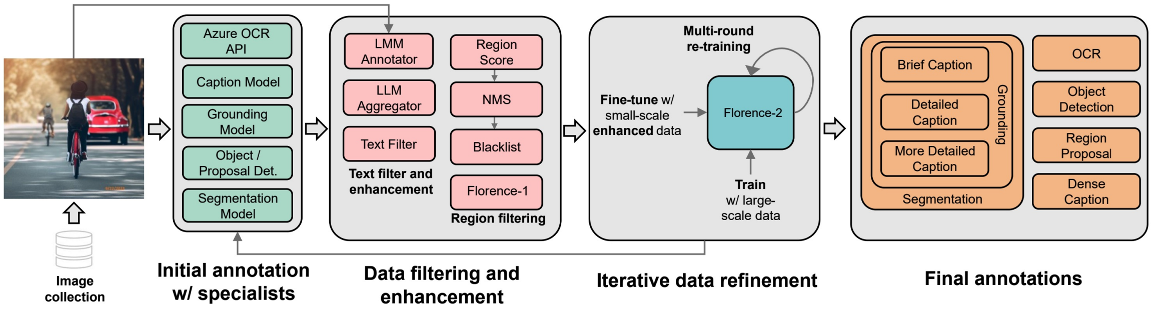


Figure 3-3: Florence-2 data engine, Image Source: Xiao et al. [51]

Florence-2 is a novel vision foundation model that uses a unified, prompt-based approach to handle various computer vision and vision-language tasks. It can label objects within images, provide their coordinates, and describe the spatial relationships between objects to understand their relative positions and interactions. Florence-2 is designed to take text prompts as task instructions and generate appropriate results, whether for image captioning, object detection, grounding, or segmentation.

3.3 Processing Customer Instructions by LLM

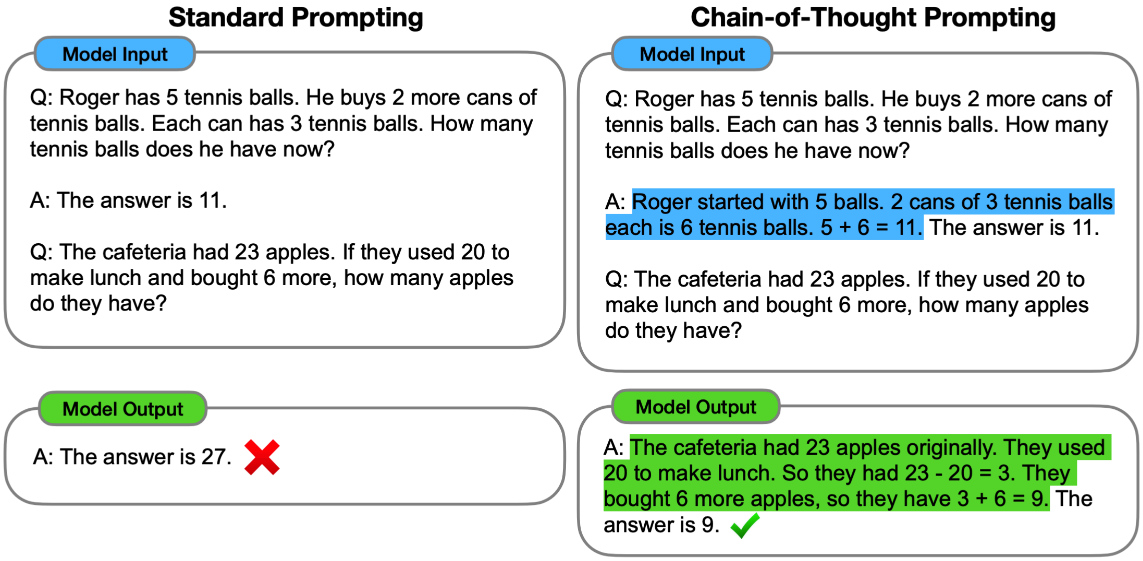
The system will leverage large language models to analyze the natural language delivery instructions provided by customers. This will enable the extraction of key information, such as the intended delivery location or any specific objects that need to be considered during the delivery process. To facilitate this, the system will leverage the labeled objects and their spatial coordinates obtained through scene analysis. Furthermore, the chain-of-thought prompting technique will be employed to guide the language model in determining the target delivery coordinates by aligning the customer's instructions with the understanding of the delivery environment.

3.3.1 Chain-of-Tought (CoT) Prompting

The chain-of-thought (CoT) prompting technique enhances the capability of LLMs to engage in complex reasoning. This approach involves providing the model with a sequence of intermediate reasoning steps as exemplars, allowing the model to autonomously develop its own reasoning abilities.

The rationale for employing this technique is that standard LLMs or VLMs alone could not reliably determine precise delivery coordinates based solely on customer instructions. By utilizing chain-of-thought prompting in conjunction with examples of detected objects and their coordinates, the language model is guided to deduce the final optimal target coordinate, which aligns more closely with the customer's specific delivery instructions. Example code demonstrating this technique is provided in Appendix B.

Incorporating the CoT prompting technique within the language model prompting framework can augment the system's ability to interpret the objects referenced in customer delivery instructions and infer the corresponding delivery target. To illustrate the difference, a side-by-side comparison between a standard input-output prompt and a chain-of-thought prompt is presented in Figure 3-4, highlighting the example-based approach of the chain-of-thought prompting method.



Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Figure 3-4: Standard Input-Output Prompt v.s. CoT Prompt, Image Source: Wei et al. [49]

Chapter 4

Experiment Results and Discussion

This thesis presents an experimental framework to validate the proposed methodology for addressing the last-meter delivery challenge. The experiments are designed to demonstrate the viability of the system, which leverages Segment Anything Model, open vocabulary object detection, large language models, and vision-language models to process customer delivery instructions. The key aspects evaluated include the construction of detailed scene graphs, the retrieval of objects based on textual queries, and the identification of target delivery coordinates through chain-of-thought prompting.

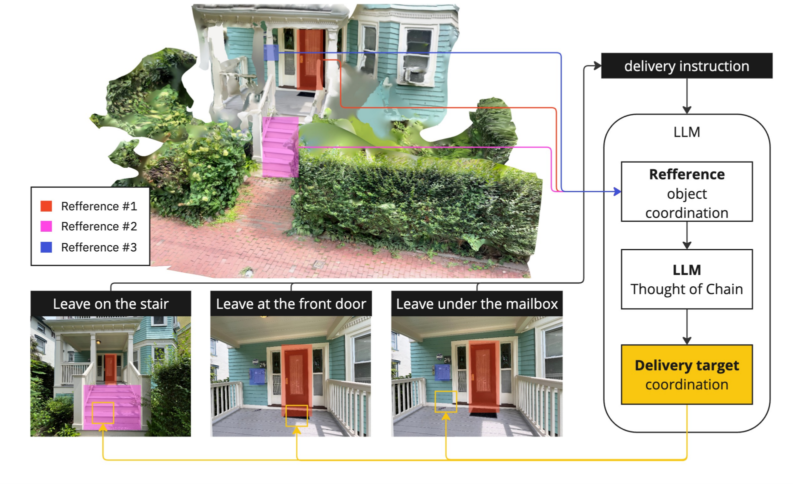


Figure 4-1: Experiment Overview

4.1 Last-Meter Data Collection

The initial experiment focuses on constructing detailed scene graphs utilizing the collected LiDAR data. This involves capturing high-resolution three-dimensional point cloud data by equipping an Apple iPad Pro with LiDAR sensors and leveraging the Polycam application to efficiently generate high-quality three-dimensional models from the LiDAR-captured photographs. The last-meter dataset comprises four main categories of data: RGB, NeRF, geometry, and instance segmentation, collected from three residential places in Cambridge, Massachusetts.

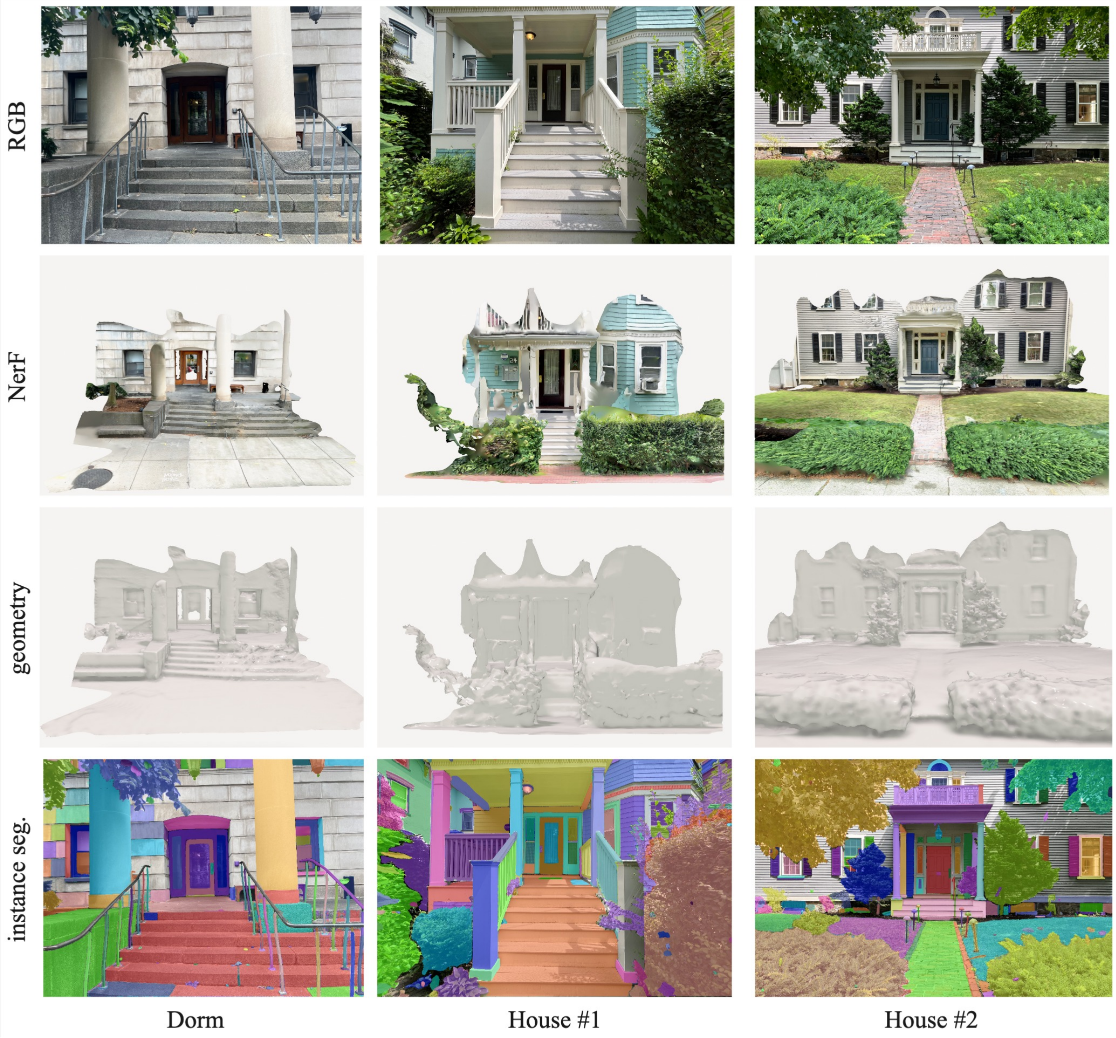


Figure 4-2: Last-Meter Dataset

4.2 Last-Meter Scene Analysis

This experiment evaluates the performance of the Segment Anything Model 2 and the Florence-2 Model in analyzing the collected data. It compares the number of detected objects and the average Intersection over Union (IoU) between the standalone use of the Florence-2 Model for object labeling and the combined use of SAM2 and Florence-2. Specifically, the Segment Anything Model 2 is utilized to segment various elements within the scene, such as buildings, trees, sidewalks, and parked vehicles. Additionally, the Florence-2 model is employed to identify and label specific objects in the scene, as well as their spatial relationships.

4.2.1 Evaluation Method

To evaluate if the additional information provided by SAM2 leads to a better understanding of the scene, we calculated the average Intersection over Union (IoU) between the standalone use of the Florence-2 Model for object labeling and the combined use of SAM2 and Florence-2. Additionally, the comparison of the number of objects detected between the standalone use of the Florence-2 Model and the combined use of SAM2 and Florence-2 was also evaluated.

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

- Intersection: This refers to the area of overlap between the predicted bounding box and the ground truth bounding box which is manually labeled. It is the common area that both boxes share.

- Union: This is the total area covered by both the predicted and ground truth bounding boxes combined. It includes the overlapping area plus the areas covered by the boxes but not in overlap.

- IoU Value: The IoU value ranges from 0 to 1, where 0 means no overlap, and 1 meands perfect overlap.

4.2.2 Results

The integrated application of the SAM2 and the Florence-2 Model identified a similar quantity of objects to the independent use of the Florence-2 Model, with 27 and 26 objects detected, respectively.

	Objects	**Florence-2**	**SAM2 + Florence-2**
Dorm	door	0.84	1
	houseplant	0.77	0
	lamp #1	0.83	0.95
	lamp #2	0.80	0.98
	stairs	0.76	0.84
	window #1	0.95	1
	window #2	0.90	1
	window #3	0.95	1
	pillar #1	0	0.98
	pillar #2	0	0.90
House #1	door	0.83	0.86
	lamp	0.93	0
	porch	0.70	0
	stairs	0.73	0.92
	air conditioner	0.98	0
	window #1	0.96	1
	window #2	0.72	1
	window #3	0.84	0.84
	window #4	0.90	1
	window #5	0.86	0.93
	window #6	0.82	0.84
House #2	door	0	0.98
	lamp	0.84	1
	porch	0.89	0.97
	stairs	0	0.97
	pillar #1	0	0
	pillar #2	0	0.95
	window #1	1	0.91
	window #2	0.76	0.81
	window #3	1	0.92
	window #4	1	1
	window #5	1	1
Average IoU		**0.705**	**0.798**

Table 4.1: Comparison of IoU between Florence-2 and SAM2 + Florence-2 for various objects in different locations.

However, the average IoU increased from 0.705 for the standalone Florence-2 Model to 0.798 for the combined SAM2 and Florence-2 approach. This approximately 10% increase in IoU is statistically significant, suggesting that the integrated application of the two models resulted in a demonstrable improvement in the ability to segment and accurately label objects within the observed scenes. Furthermore, the results indicate that the standalone approach using the Florence-2 Model identified an air conditioner, which is less pertinent for last-meter delivery operations. In contrast, the integrated method leveraging both SAM2 and Florence-2 did not detect the air conditioner but successfully identified other objects more relevant for last-meter delivery, such as doors and stairs. A similar pattern was observed in house #2 of the dataset, where the combined approach effectively identified the door and stairs, while the standalone Florence-2 Model approach did not.

4.3 Target Coordinate Identification by Chain-of-Thought Prompting

This experiment evaluates the system's ability to employ LLMs in conjunction with the CoT prompting technique. The aim is to leverage the coordinates of detected objects, obtained from the previous experiment, to identify the target delivery location. By incorporating the CoT approach and providing examples of the detected objects and their corresponding coordinates, the LLM is guided to deduce the optimal target coordinate, which aligns more closely with the customer's specific delivery instructions. Appendix B provides an example of the CoT approach employed in this experiment.

4.3.1 Evaluation Method

The performance of this experiment was evaluated by calculating the coverage rate of the predicted bounding boxes, defined as the ratio of the area of overlap between the predicted bounding box and the manually labeled ground truth area to the total area of the predicted bounding box.

$$\text{Coverage Rate} = \frac{\text{Area of Overlap}}{\text{Area of Predicted Bounding Box}}$$

- Overlap: This refers to the area of intersection between the predicted bounding box and the manually labeled ground truth bounding box. It is the common area that both boxes share.

- Predicted Bounding Box Area: This is the total area of the bounding box predicted by the model.

- Coverage Rate Value: The coverage rate ranges from 0 to 1, where 0 indicates no overlap, and 1 indicates that the predicted bounding box fully overlaps the ground truth area.

4.3.2 Results

The results demonstrated that both models performed well in scenarios where the relevant object was clearly detectable, achieving high coverage rates approaching 1. However, when instructions involved objects not present in the observed environment, both models failed to accurately identify the correct delivery coordinates, resulting in a coverage rate of 0. This highlights a key limitation: the models' dependency on the presence of detectable objects to accurately determine the target delivery location.

The coverage rates exhibited slight variations between the models, with the GPT-4-o benchmark generally achieving higher accuracy compared to the LLaMA 3.1 model. For instance, GPT-4-o attained a coverage rate of 0.97 for the instruction "Leave at the front door" at House #1, compared to 0.8 for LLaMA 3.1. These findings suggest that while both models can effectively interpret straightforward delivery instructions, the precision of the coordinate identification improves when using the more advanced LLM model.

Overall, these results underscore the effectiveness of CoT prompting technique in guiding the LLMs to align customer instructions with the detected objects, thereby deducing precise delivery coordinates. However, the findings also highlight to the need for further enhancements, such as implementing default fallback prompts or leveraging additional cues like low-cost RFID tags, to address cases where the specified objects are not detected within the observed scene.

Location	**Delivery Instruction**	**Ground Truth**	**GPT 4-o**	**Llama 3.1**
Dorm	Leave on the stair	(172, 612, 1278, 958)	1	1
	Leave at the front door	(489, 275, 753, 565)	1	1
	Leave under the mailbox	-	0	0
House #1	Leave on the stair	(179, 358, 976, 958)	1	0.92
	Leave at the front door	(583, 312, 673, 518)	0.97	0.8
	Leave under the mailbox	(465, 252, 110, 100)	0	0
House #2	Leave on the stair	(484, 485, 832, 568)	0.82	0.81
	Leave at the front door	(761, 1053, 277, 1136)	1	0.96
	Leave under the mailbox	-	0	0

Table 4.2: Ground Truth Coordinates and Coverage Rate Comparison

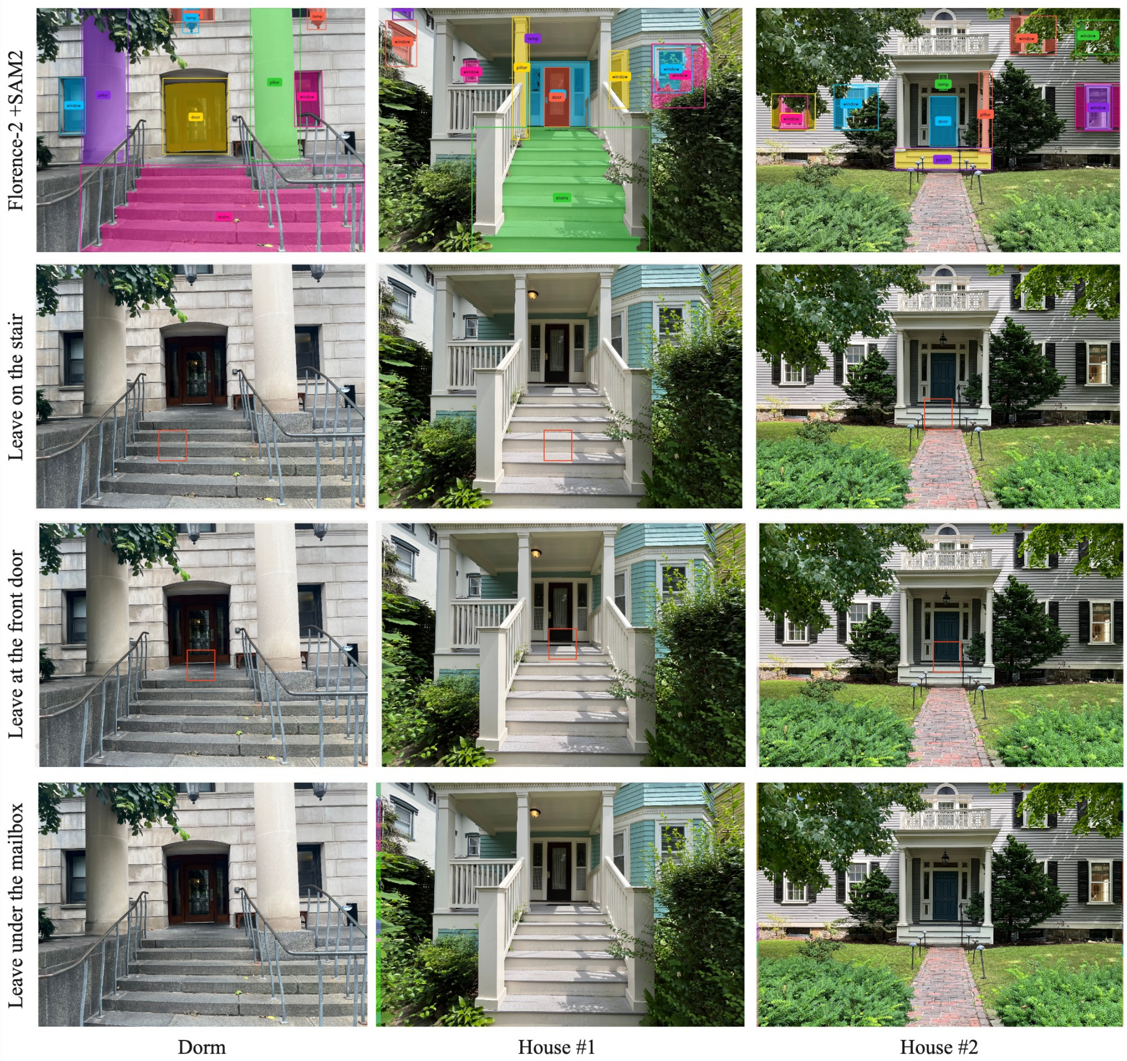


Figure 4-3: Coordinate Output from GPT4-o

4.4 Discussion

The experiments conducted demonstrate that the integrated application of the Segment Anything Model 2 and the Florence-2 Model exhibits better performance in object detection and labeling compared to the standalone use of the Florence-2 Model. Additionally, the GPT-4-o benchmark displayed a higher Intersection over Union value relative to the LLaMA 3.1 benchmark. However, when the object mentioned in the customer's instruction is not detected within the scene, neither the GPT-4-o nor the LLaMA 3.1 language models can accurately identify the appropriate delivery coordinates, as the referenced object is not present. This issue could potentially be addressed by implementing a series of default prompts to handle such scenarios where the requested object is not found in the observed environment.

The conducted experiments show that using the Segment Anything Model 2 and the Florence-2 Model together performs better in object detection and labeling compared to using just the Florence-2 Model alone. The increase in accuracy, indicated by the higher average Intersection over Union, supports the claim that combining advanced models leads to more reliable detection results, which is important for real-world applications like last-meter delivery scenarios.

The standalone application of Florence-2 identified an air conditioner in house #1 of the dataset, which is less relevant for the task of last-meter delivery. In contrast, the integrated approach utilizing both SAM2 and Florence-2 did not detect the air conditioner, but successfully identified other objects more pertinent to last-meter delivery, such as doors and stairs. A similar pattern was observed in house #2 of the dataset, where the combined method effectively recognized the door and stairs, while the standalone Florence-2 Model approach did not. This suggests that the integrated application of the two models leads to a more targeted and relevant object detection and labeling, which is crucial for improving the performance of last-meter delivery systems. Furthermore, the potential benefit of fine-tuning the Florence-2 Model to better recognize specific objects, such as mailboxes, could further enhance the system's ability to identify critical elements for last-meter delivery in future research iterations.

Additionally, the GPT-4-o benchmark displayed a higher Intersection over Union value relative to the LLaMA 3.1 benchmark. However, when the object mentioned in the customer's instruction is not detected within the scene, neither the GPT-4-o nor the LLaMA 3.1 language models can accurately identify the appropriate delivery coordinates, as the referenced object is not present. This challenge could potentially be addressed through the implementation of a decision-making framework that dynamically adapts to the available environmental cues. By employing a decision tree approach, the system can leverage the information gathered from the scene analysis to determine the most appropriate course of action. Furthermore, a set of default prompts can be developed to handle scenarios where the specific object referenced in the customer's instructions is not detected within the observed environment. In such cases, the system could send a follow-up message to the customer, informing them that the requested object was not identified and, therefore, the delivery instructions could not be fully executed. This adaptive approach would enhance the system's ability to navigate complex real-world situations and provide a more reliable and responsive last-meter delivery service. Moreover, exploring the integration of low-cost wireless RFID tags that customers can easily install at their doorsteps could assist in more accurately locating the target delivery destination during the last-meter stage of the delivery process.

Chapter 5

Conclusion

This research outlines the implementation and evaluation of the Segment Anything Model 2, Florence-2, and large language models for the purpose of identifying delivery destinations based on customer instructions. The study develops a methodology to bridge the gap between last-mile and last-meter delivery operations, and assesses the viability of the proposed approach for future last-meter delivery applications in outdoor environments. The findings indicate that the utilization of SAM2, Florence-2, and LLMs has the potential to enhance last-meter delivery by providing an efficient solution for navigating complex outdoor environments and delivering packages to their intended recipients with minimal human intervention.

5.1 Contribution

This thesis makes several contributions to the field of last-meter delivery:

- **Last-Meter Dataset**: Introduced the first last-meter dataset, incorporating high-resolution three-dimensional LiDAR data, detailed geometric information, segmented objects, and labeled elements within the environment.

- **Feasibility Study**: Conducted a thorough evaluation of Segment Anything Model, Florence-2, and LLMs for outdoor last-meter delivery applications, demonstrating its potential to accomplish the last-meter delivery.
- Developed a methodology to bridge the gap between last-mile and last-meter delivery operations: Ultimately, this thesis lays the groundwork for future research and development in the emerging field of autonomous last-meter delivery, which holds potential to address the growing challenges of urban logistics and transportation.

5.2 Limitation

While the results are promising, this research also encountered several limitations:

- **Data Collection**: The quality and completeness of LiDAR data collected under various weather conditions and environments can affect the accuracy of scene representation. Additionally, obtaining permission from property owners to scan and make data open-source limited the scope of the dataset, as securing such permissions can be time-consuming and challenging.
- **Model Accuracy**: Limitations in the accuracy of 3D object detection, semantic segmentation, and spatial reasoning models can impact the overall performance of the delivery system.
- **Computational Demands**: The proposed method requires significant computational resources for real-time scene analysis and navigation, which may limit its scalability and practical deployment in resource-constrained settings.
- **Regulatory Challenges**: Legal and regulatory constraints related to the deployment of autonomous delivery robots in public spaces pose additional challenges to widespread adoption.

5.3 Potential

The potential of LLM and VLM in revolutionizing last-meter delivery is vast:

- Last-meter delivery is a critical and unsolved challenge in the broader context of urban logistics, as it involves navigating complex outdoor environments to reach the final destination.

- **Enhanced Autonomy**: By providing detailed 3D scene representations and relational information, LLM can significantly enhance the autonomy of delivery robots, reducing the need for human intervention.

- **Broader Applications**: Beyond last-meter delivery, VLM's scene representation and dynamic updating capabilities have the potential to be applied in various other domains, such as search and rescue operations, autonomous driving, and smart city infrastructure management.

- **Improved Customer Experience**: The ability to accurately map customer instructions to delivery targets can improve the reliability and efficiency of delivery services, leading to a better customer experience.

5.4 Future Work

Future research may involve incorporating temporal dynamics into the model and preloading local maps to address out-of-sight issues, as well as evaluating the system's performance in less structured and more challenging environments to further enhance its reliability and applicability. Fine-tuning Florence-2 for specific objects, such as mailboxes, could also be a viable approach for future research iterations and would likely improve recognition rates for such critical items.

Additionally, exploring the integration of VLM with emerging technologies, such as 5G communication networks and advanced sensor arrays, could potentially improve its robustness and scalability. We can consider integrating with RFID localization to enable more robust last-meter delivery and new applications. A recent study [9] shows the use of robotic systems to achieve efficient and accurate localization of RFID tags, which enables new possibilities in robot delivery services. Expanding the dataset to encompass a wider range of environments and conditions will also be crucial for enhancing the generalizability of the proposed method.

Furthermore, compared to wheeled robots, quadrupedal and humanoid robots may be better equipped to navigate the diverse terrain and obstacles encountered during the last-meter delivery scenario. Consequently, another area for future work could be the utilization of quadrupeds to complete the final phase of the delivery workflow.

Acknowledgments

I am deeply grateful to a number of people for their unwavering support and guidance throughout the process of writing this thesis. I have been truly blessed to have such a wonderful support system. I would like to express my sincere gratitude to Professor Kent Larson for his invaluable guidance and support on this topic. I am particularly grateful to Michael, who has been an incredible mentor during my time as a visiting student and has guided me into the field of last-mile transportation. Additionally, I am fortunate to have Luis, Kevin, and Professor Sekimoto Yoshihide on my thesis committee, providing their expertise and encouragement.

Markus and Jason have always been there to bounce around ideas, provide feedback, and constantly motivate me. I also want to thank my friends Jonny, Andres, Tammy, Kai, Jue, Phil, Weitung, Leti, Naroa, and Eggi for their mental support and encouragement throughout this wonderful yet challenging journey.

I am grateful to my family, who have been a constant source of encouragement and strength during this process. Finally, I would like to thank Mr. Rubber for enhancing the cooperation between Taipei Tech and MIT, which led me here.

Appendix A

Object Detection Output

Comparison of Object Detection Output		
Subject	**Dorm**	
Objects	**Florence2**	**SAM2 + Florence2**
door	0.84	1
houseplant	0.77	0
lamp #1	0.83	0.95
lamp #2	0.80	0.98
stairs	0.76	0.84
window #1	0.95	1
window #2	0.90	1
window #3	0.95	1
pillar #1	0	0.98
pillar #2	0	0.90

Table A.1: Comparison of IoU between Florence-2 and SAM2 + Florence-2 at the dorm of the last-meter dataset

Comparison of Object Detection Output		
Subject	**House #1**	
Objects	**Florence2**	**SAM2 + Florence2**
door	0.83	0.86
lamp	0.93	0
porch	0.70	0
stairs	0.73	0.92
air conditioner	0.98	0
window #1	0.96	1
window #2	0.72	1
window #3	0.84	0.84
window #4	0.90	1
window #5	0.86	0.93
window #6	0.82	0.84

Table A.2: Comparison of IoU between Florence-2 and SAM2 + Florence-2 at the house#1 of the last-meter dataset

Comparison of Object Detection Output		
Subject	**House #2**	
Objects	**Florence2**	**SAM2 + Florence2**
door	0	0.98
lamp	0.84	1
porch	0.89	0.97
stairs	0	0.97
pillar #1	0	0
pillar #2	0	0.95
window #1	1	0.91
window #2	0.76	0.81
window #3	1	0.92
window #4	1	1
window #5	1	1

Table A.3: Comparison of IoU between Florence-2 and SAM2 + Florence-2 at the house#2 of the last-meter dataset

Appendix B

Code of LLM Chain-of-Though

```
from openai import OpenAI
client = OpenAI()


system_message = {
    "role": "system",
    "content": '''You're a help assistant of package delivery. When given a text instruction like "put the package infront of the door", you can infer where is the most possible coordinates of the target area based on the reference objects and your position.
    Example Input:
    {
      "instruction": "Leave the package at the front door.",
      "reference objects": {
        Door: [3,0,1.5],
        Stairs:[1.5,0,0.75],
        Mailbox:[3,0.5,1.5]
      },
      "your position": [2,0,1.5]
    }
    Example Output:
    {
      Task: allocate the coordinates in font the door.
      Observation: the door is at [3,0,1.5], my position is at [2,0,1.5]
      Though: the door is in front of me, the area in front of the door should be somewhere between my position and the door, but closer to the door.
      Answer: {2.8,0,1.5}
    }
  },
  '''
  },

user_message = {
    "role": "user",
    "content": '''{
        "instruction": "Leave the package under the mailbox."
        "reference objects": {
        Door: [3,0,1.5],
        Mailbox:[3,0.5,1.5]
        },
        "your position": [2,0,1.5]
    }
    '''
}

completion = client.chat.completions.create(
  model="gpt-4o-mini",
  messages = [
    {"role": "system", "content": '''You're a help assistant of package delivery. When given a text instruction like "put the package infront of the door", you can infer where is the most possible coordinates of the target area based on the reference objects and your position.
    Example Input:
    {
      "instruction": "Leave the package at the front door.",
      "reference objects": {
        Door: [3,0,1.5],
        Stairs:[1.5,0,0.75],
        Mailbox:[3,0.5,1.5]
      },
      "your position": [2,0,1.5]
    }
    Example Output:
    {
      Task: allocate the coordinates in font the door.
      Observation: the door is at [3,0,1.5], my position is at [2,0,1.5]
      Though: the door is in front of me, the area in front of the door should be somewhere between my position and the door, but closer
```

```
          to the door.
      Answer: {2.8,0,1.5}
    }
  },
  '''},
    {"role": "user", "content": '''{
        "instruction": "Leave the package under the mailbox."
        "reference objects": {
        Door: [3,0,1.5],
        Mailbox:[3,0.5,1.5]
        },
        "your position": [2,0,1.5]
    }
    '''}
  ]
)

print(completion.choices[0].message)
```

Listing B.1: Chain of Thought

References

[1] Kiwibot. (2024). kiwibot official website. retrieved 16 august, from https://www.kiwibot.com/.

[2] Nuro— on a mission to better everyday life through robotics. (2024). nuro official website. retrieved 16 august, from https://www.nuro.ai/.

[3] Starship technologies. (2024). starship technologies official website. retrieved 16 august, from https://www.starship.xyz/.

[4] J. Allen, M. Piecyk, M. Piotrowska, F. McLeod, T. Cherrett, K. Ghali, T. Nguyen, T. Bektas, O. Bates, A. Friday, S. Wise, and M. Austwick. Understanding the impact of e-commerce on last-mile light goods vehicle activity in urban areas: The case of london. *Transportation Research Part D: Transport and Environment*, 61:325–338, 2018.

[5] Amazon. Amazon drones can now fly farther and deliver to more customers following faa approval, 05 2024.

[6] Mehdi Behroozi and Dinghao Ma. Last mile delivery with drones and sharing economy, 01 2023.

[7] Techane Bosona. Urban freight last mile logistics—challenges and opportunities to improve sustainability: A literature review. *Sustainability*, 12(21), 2020.

[8] Nils Boysen, Stefan Fedtke, and Stefan Schwerdfeger. Last-mile delivery concepts: a survey from an operational research perspective. *Springer Science+Business Media*, 43(1):1–58, 09 2020.

[9] Weitung Chen, Tara Boroushaki, Isaac Perper, and Fadel Adib. Reinforcement learning for rfid localization. In *2024 IEEE International Conference on RFID (RFID)*, pages 1–6, 2024.

[10] DHL. Dhl launches its first regular fully-automated and intelligent urban drone delivery service, 05 2019.

[11] Miguel Figliozzi. Carbon emissions reductions in last mile and grocery deliveries utilizing air and ground autonomous vehicles. *Elsevier BV*, 85:102443–102443, 08 2020.

[12] Rodrigo Marçal Gandia, Fabio Antonialli, Bruna Habib Cavazza, Arthur Miranda Neto, Danilo Alves de Lima, Joel Yutaka Sugano, Isabelle Nicolai, and Andre Luiz Zambalde. Autonomous vehicles: scientometric and bibliometric review*. *Transport Reviews*, 39(1):9–28, 2019. Long Term Implications of Automated Vehicles.

[13] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna M. Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning*, 09 2023.

[14] Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*, 2024.

[15] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world, 2024.

[16] Henry Alexander Ignatious, Hesham-El-Sayed, and Manzoor Khan. An overview of sensors in autonomous vehicles. *Procedia Computer Science*, 198:736–741, 2022. 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare.

[17] Navid Mohammad Imran, Sabyasachee Mishra, and Myounggyu Won. Towards fully autonomous drone-based last-mile delivery, 01 2021.

[18] Dylan Jennings and Miguel Figliozzi. Study of sidewalk autonomous delivery robots and their potential impacts on freight efficiency and travel. *SAGE Publishing*, 2673(6):317–326, 05 2019.

[19] Shyam Sundar Kannan, Ahreum Lee, and Byung-Cheol Min. External human-machine interface on delivery robots: Expression of navigation intent of the robot. 08 2021.

[20] Shyam Sundar Kannan and Byung-Cheol Min. Door delivery of packages using drones., 04 2021.

[21] Maja Kiba-Janiak, Jakub Marcinkowski, Agnieszka Jagoda, and Agnieszka Skowrońska. Sustainable last mile delivery on e-commerce market in cities from the perspective of various stakeholders. literature review. *Sustainable Cities and Society*, 71:102984, 2021.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[23] S Santosh Kumar, D Hemanth, S Dwneeth, K Dilip, and A Divyatej. Automated package delivery accepting system-smart freight box. In *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, pages 1510–1514. IEEE, 2019.

[24] Bai Li, Shaoshan Liu, Jie Tang, Jean-Luc Gaudiot, L. Zhang, and Qi Kong. Autonomous last-mile delivery vehicles in complex traffic environments. *IEEE Computer Society*, 53(11):26–35, 11 2020.

[25] Yongjian Li, Yan Chen, Gaicong Guo, Hui-Wen Wu, and Yuan Zhao. Integrated routing for a vehicle-robot pickup and delivery system with time constraints, 01 2022.

[26] Jacob P. Macdonald, Rohit Mallick, Allan B. Wollaber, Jaime D. Peña, Nathan McNeese, and Ho Chit Siu. Language, camera, autonomy! prompt-engineered robot

control for rapidly evolving deployment. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 717–721, New York, NY, USA, 2024. Association for Computing Machinery.

[27] Blake Nazario-Casey, Harris Newsteder, and O. Patrick Kreidl. Algorithmic decision making for robot navigation in unknown environments. 03 2017.

[28] Thanat Nonthaputha, Montree Kumngern, Jirapat Phookwantong, and Sompong Keawwang. Arduino based smart box for receiving parcel posts. In *2020 18th International Conference on ICT and Knowledge Engineering (ICT&KE)*, pages 1–5. IEEE, 2020.

[29] Jing Zhi Ooi and Chye Cheah Tan. Smart modular parcel locker system using internet of things (iot). In *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, pages 66–71, 2021.

[30] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *7th Annual Conference on Robot Learning*, 2023.

[31] Luigi Ranieri, Salvatore Digiesi, Bartolomeo Silvestri, and Michele Roccotelli. A review of last mile logistics innovations in an externalities cost reduction vision. *Sustainability*, 10(3), 2018.

[32] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 08 2024.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[34] David Reid. Domino's delivers world's first ever pizza by drone.

[35] Leonardo N. Rosenberg, Noemie Balouka, Yale T. Herer, Eglantina Dani, Paco Gasparin, Kerstin Dobers, David Rüdiger, Pete Pättiniemi, Peter Portheine, and Sonja van Uden. Introducing the shared micro-depot network for last-mile logistics. *Multidisciplinary Digital Publishing Institute*, 13(4):2067–2067, 02 2021.

[36] Pericle Salvini, Diego Paez-Granados, and Aude Billard. Safety concerns emerging from robots navigating in crowded pedestrian areas. *Springer Science+Business Media*, 14(2):441–462, 06 2021.

[37] Farah Samouh, Veronica Gluza, Shadi Djavadian, SeyedMehdi Meshkani, and Bilal Farooq. Multimodal autonomous last-mile delivery system design and application. 09 2020.

[38] Eyad Shaklab, Areg Karapetyan, Arjun Sharma, Murad Mebrahtu, Mustofa Basri, Mohamed Nagy, Majid Khonji, and Jorge Dias. Towards autonomous and safe last-mile deliveries with ai-augmented self-driving delivery robots, 2023.

[39] Michele D. Simoni, Erhan Kutanoglu, and Christian G. Claudel. Optimization and analysis of a robot-assisted last mile delivery system. *Transportation Research Part E: Logistics and Transportation Review*, 142:102049, 2020.

[40] Safaa Sindi and Roger Woodman. Autonomous goods vehicles for last-mile delivery: Evaluation of impact and barriers. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020.

[41] Jonathan Ross Tew and Lydia Ray. Addsmart: Address digitization and smart mailbox with rfid technology. In *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 1–6. IEEE, 2016.

[42] Zhaofeng Tian and Weisong Shi. Design and implement an enhanced simulator for autonomous delivery robot. 04 2022.

[43] Trafikanalys. E-handels effekter på transportsystemet (the effects of e-commerce on the transport system), 2022. Transport Analysis Agency, Report 2022:4, Stockholm, Sweden.

[44] Stanislava Turska and Lucia Madleňáková. Concept of smart postal mailbox. *Transportation Research Procedia*, 40:1199–1207, 2019.

[45] Günter Ullrich. *The History of Automated Guided Vehicle Systems*, pages 1–14. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[46] Ron van Duin, Bart Wiegmans, Bart van Arem, and Yorick van Amstel. From home delivery to parcel lockers: a case study in amsterdam. *Elsevier BV*, 46:37–44, 01 2020.

[47] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information, 2024.

[48] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, 2024.

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.

[50] David A. Weinberg, H. A. Dwyer, Sarah Fox, and Nikolas Martelaro. Sharing the sidewalk: Observing delivery robot interactions with pedestrians during a pilot in pittsburgh, pa. *Multidisciplinary Digital Publishing Institute*, 7(5):53–53, 05 2023.

[51] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023.

[52] Sun Ze-hong and Zhao Guang-yuan. Multi-functional parcel delivery locker system. In *2015 International Conference on Computer and Computational Sciences (ICCCS)*, pages 207–210, 2015.

[53] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models, 2023.

[54] Zhu, Taoyuanmin, Gabriel I. Fernandez, Colin Togashi, Yeting Liu, and Dennis Hong. Feasibility study of limms, a multi-agent modular robotic delivery system with various locomotion and manipulation modes. In *2022 19th International Conference on Ubiquitous Robots (UR)*, pages 30–37, 2022.