CMRM: A Cross-modal Reasoning Model to Enable Zero-shot Imitation Learning for Robotic RFID Inventory in Unstructured Environments

†Yongshuai Wu, §Jian Zhang, †Shaoen Wu, ‡Shiwen Mao, and *Ying Wang
†Department of Information Technology, Kennesaw State University, Marietta, GA 30060, USA
§Department of Electrical and Computer Engineering, Kennesaw State University, Marietta, GA 30060, USA
‡Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA
*Department of Robotics and Mechatronics Engineering, Kennesaw State University, Marietta, GA 30060, USA
Email: ywu26@students.kennesaw.edu, jianzhang@ieee.org, swu10@kennesaw.edu, smao@ieee.org, ywang34@kennesaw.edu

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437833

***Abstract*—The fast development in Deep Learning (DL) has made it a promising technique for various autonomous robotic systems. Recently, researchers have explored deploying DL models, such as Reinforcement Learning and Imitation Learning, to enable robots for Radio-frequency Identification (RFID) based inventory tasks. However, the existing methods are either focused on a single field or need tremendous data and time to train. To address these problems, this paper presents a Cross-Modal Reasoning Model (CMRM), which is designed to extract high-dimension information from multiple sensors and learn to reason from spatial and historical features for latent cross-modal relations. Furthermore, CMRM aligns the learned tasking policy to high-level features to offer zero-shot generalization to unseen environments. We conduct extensive experiments in several virtual environments as well as in indoor settings with robots for RFID inventory. The experimental results demonstrate that the proposed CMRM can significantly improve learning efficiency by around 20 times. It also demonstrates a robust zero-shot generalization for deploying a learned policy in unseen environments to perform RFID inventory tasks successfully.**

***Index Terms*—Imitating Learning, RFID inventory, Long-horizon tasks, Cross-modal reasoning, multiple sensing spaces**

I. Introduction

The Radio-frequency Identification (RFID) technology provides a low-cost and easy-to-deploy inventory solution that has been widely deployed in retail stores, factories, and warehouses [1], [2]. This paper presents a learning-based model that enables robots to perform RFID-based automated inventory in unstructured environments. From previous studies [3], [4], this requires a robot with continuous control capabilities for long-horizon action planning, which remains to be a significant challenge for autonomous embodied agents [5], [6]. To efficiently and effectively scan all RFID tags in an unstructured environment, the robot must perceive the surrounding spatial space and align it with the RFID sensing space. Furthermore, optimized action planning should rely on current and historical observations to build complete tasking state information since the robot's sensors can only partially observe the environmental spaces.

Recent developments in Deep Reinforcement Learning (DRL) and Imitation Learning (IL) have shown great promise in enabling robots to learn policies for more accomplishing complex tasks [7]–[9]. However, the current DRL, IL, and their combined methods require tremendous training data to succeed [4], [10], [11], resulting in unsustainable training costs for our RFID inventory tasks in an unstructured real-world environment. This data-hungry problem is imposed by their inherent low learning efficiency. First, training policy from observed low-level features causes the lack of the ability to learn critical features that affect the action. Second, the latent relationship between features and actions is inefficiently and insufficiently explored by current learning models. Other learning-based methods [12], [13] tend to bypass the complexity of long-horizon tasks by predicting a sub-goal and then rely on the robot's built-in ability to achieve the sub-goal. Usually, the sub-goal is selected from its existing known states, such as the position in an image from a static top-view camera. These methods usually work well in small-scaled scenarios, such as a fixed robotic arm for pick-and-place tasks, which only require a few steps with all states known or being easy to predict. However, the robot must explore large and unknown spaces in the RFID inventory task, making these methods unsuitable. Additionally, training the model to learn a robotic policy directly in a physical environment is costly, sometimes even infeasible and dangerous. To bridge this gap, we can train a policy in a virtual environment and then fine-tune it for a real robotic application [14]. However, the "reality gap" challenges this methodology, a phenomenon where the virtual-learned policies cannot be directly applied to real robotic applications [15].

This research addresses the above challenges with a proposed Cross-Modal Reasoning Model (CMRM), which could efficiently align the information between RFID sensing and spatial spaces and learn the latent cross-space relations from current and historical observations. The main contribution of our work is summarized as follows:

This work is supported in part by the NSF under Grants ECCS-1923163, ECCS-1923717, ECCS-2245607, and ECCS-2245608.

- A zero-shot model that learns the tasking policy conditioned on abstracted semantic features, which allows it to be deployed in unseen environments.
- The proposed model provides cross-modal reasoning to effectively learn the latent relations from multiple varied sensing spaces in current and historical observations.
- The proposed model outperforms previous works by improving the learning efficiency around 20 times.
- We conduct extensive experiments to validate that CMRM can provide robust zero-shot generalization in unseen tasking environments. Furthermore, the virtual-learned policy can be directly deployed in a real-world robot, proving that CMRM can effectively bridge the "reality gap."

The remainder of the paper is organized as follows. We present the problem statement in Section II and the proposed solution in Section III. We evaluate the proposed CMRM model in Section IV and conclude this paper in Section V.

II. Problem Statement

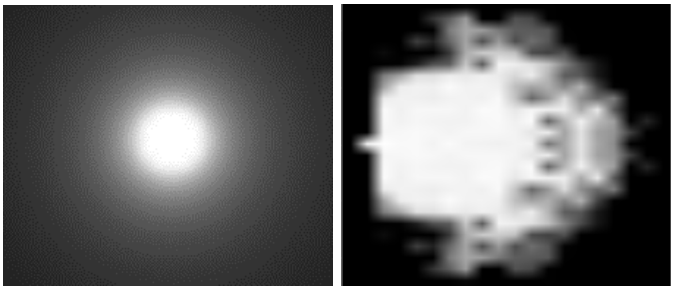
This work will enable a mobile robot to perform autonomous RFID inventory. Our goal is to train a policy $\pi(a_t|\{s_i\}_{i=t-T}^{t};\{v_i\}_{i=t-T}^{t})$ conditioned on a sequence of environmental observations $\{s_i\}_{i=1}^{t}$ and Radio Frequency (RF) observations $\{v_i\}_{i=1}^{t}$, where a_t is the predicted action, t is the current step, and $T \leq t$ is the sequence length.

The *environmental observation* $s_t \in \mathcal{S}$ at step t consists of the outputs from the robot's primary sensors, such as the camera and Lidar, to represent the instantly observed surrounding environment. This paper only considers the robot equipped with a 2D Lidar, which scans the surroundings in 360° at a fixed increment angle $\alpha = 360°/n$. The Lidar provides observations $s_t = [l_1, l_2, \ldots, l_n]$, where l_j is the distance to the object detected on the angle αj.

The *RF observation* $v_t \in \mathcal{V}$ provides an approximated observing area of the robot's RFID reader at step t. We deployed the RFID model proposed in [16] to define our RF observation as $v_t \sim P(\cdot|d)$, which is the probability for a tag to be observed at a distance d to the RFID reader's antenna. Fig. 1 shows an example of v_t for an isotropic antenna and a directional RFID antenna: the observed probability of tags decreases when the distance d is increased.

III. Proposed Approach

We propose a Cross-Modal Reasoning Model (CMRM) to enable robots for RFID inventory by learning tasking policy π from a few demonstrations. The learned policy shall provide a robust generalization for the target unstructured environments, such as an apparel store. To this end, the proposed model extracts high-level and abstracted features from the observations and then efficiently explores the latent relations among those features from multiple sensing spaces and steps. Finally, the model learns the tasking policy π conditioned on the high-level features. The architecture of the proposed method is presented in Fig. 2, which comprises three modules: Data Pre-processing, Tasking Reasoner, and Action Decoder.



(a) An isotropic antenna. (b) A directional antenna.

Fig. 1: Typical RF observations v_t obtained by two reader antennas, the brighter pixel denotes a probability of being observed.

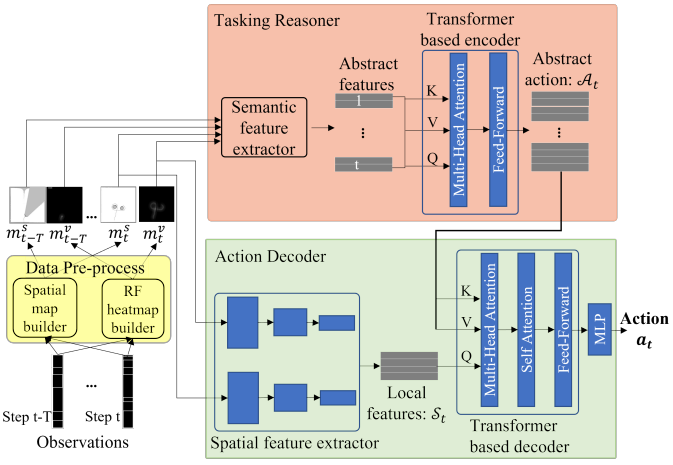
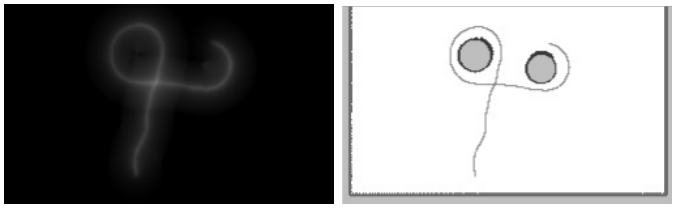


Fig. 2: High-level architecture of the proposed CMRM, only the filled blue components are trainable. It comprises three sub-modules: Data Pre-processing marked as yellow, the Tasking Reasoner in orange, and the Action Decoder in light green.

A. Data Pre-processing

The Data Pre-processing module consists of two sub-modules, *Spatial map builder* f_s and *RF heatmap builder* f_v. The spatial map builder is based on occupancy grids [17], a traditional spatial environment representation widely used in robotics. From the current observation v_t and the previous map m_{t-1}^s, it will recursively create and update an occupancy grid-based map $m_t^s = f_s(s_t|m_{t-1}^s)$, $m_t^s \in \mathbb{R}^{H\times W\times d}$, where H, W, d are the height, width, and number of channels. In this project, we deploy it as a grey map with $d = 1$. The m_t^s provides a simple and computationally efficient spatial model to facilitate the navigation of robots [18]. Besides spatial information, f_s will also record the robot tasking trajectory in m_t^s with a different grid value to distinguish it from occupancy grids. This way, m_t^s could serve as a memory to provide the downstream modules with a completed robotic spatial state. Fig. 3b shows an example of m_t^s. While the robot performs an RFID inventory, the *RF heatmap bulider* f_v will present an RF map m_t^v, as shown in Fig. 3a. Based on the RF observation v_t, the robot's position p_t and the previous RF map m_{t-1}^v, f_v will predict the current map, $m_t^v = f_v(v_t, p_t|m_{t-1}^v)$, $m_t^v \in \mathbb{R}^{H\times W\times d}$. It will provide a complete RF signal distribution in the environment considering



(a) (b)

Fig. 3: A typical output of the Data Pre-processing module while a robot performs an RFID inventory task: (a) An example of m_t^v: brighter grids/pixels indicate a denser RF signal, and the tags in those grids are more likely to be scanned. (b) An example of m_t^s: black grids/pixels indicate the detected objects, grey pixels represent the unknown spaces (areas inside the black circles), and the light grey curves present the robot's navigated trajectory.

the robot's historical navigational and RF information.

B. Tasking Reasoner

The robotic RFID inventory is a typical long-horizon task in cross-sensing spaces. The robot should reason in spatial and RF spaces and examine historical observations to maintain complete knowledge of tasking. Additionally, to avoid overfitting the local information, it should understand and learn the tasking policy from high-level and semantic features. We design our Tasking Reasoner to meet all those requirements. The input is a sequence of maps $\mathcal{M}\{(m_{t-T}^s, m_{t-T}^v), \cdots, (m_t^s, m_t^v)\}$ that is converted from observations by the proposed Data Pre-process, where T is the length of the sequence. As shown in the orange block in Fig. 2, the Tasking Reasoner module comprises a semantic feature extractor and a Transformer-based encoder.

The semantic feature extractor is a frozen pre-trained image encoder from CLIP [19], which is designed to align vision and language representations and trained with millions of image-caption pairs. The trained model can efficiently extract features that are more sensitive to natural language. Furthermore, those extracted features are resistant to the local noises and discrepancies in the low-level features, which help the model transfer to new conditions. It will extract the semantic-level features from input maps, i.e., $m_t^s \rightarrow e_t^s \colon \mathbb{R}^{1\times 512}$ and $m_t^v \rightarrow e_t^v \colon \mathbb{R}^{1\times 512}$, respectively. We concatenate these two features as $e_t = [e_t^s; e_t^v]$. The input sequence of maps $\mathcal{M}$ will be transformed to $\{e_{t-T}, \cdots, e_t\}$. We further concatenate them to obtain a feature map $E_t = [e_{t-T}; \cdots; e_t]$. To retain the sensing spaces and temporal information, we introduce a two-layer positioning embedding:

$$\mathcal{E}_t = E_t + E_{sp} + E_{st}, \tag{1}$$

where E_{sp} is the embedding of feature space id to indicate the features from e_t^s or e_t^v, and E_{st} is the embedding of step id that denotes the temporal information t. Thus, the final $\mathcal{E}_t$ provides an integrity feature map with all semantic information for tasking.

With the integrity feature map $\mathcal{E}_t$, we then design a Transformer-based encoder to obtain a contextualized abstract action $\mathcal{A}_t$. It could efficiently learn relationships in the two maps m_t^s and m_t^v at the same step t, and among the maps at all steps $t-T, \cdots, t$. This enables the learned $\mathcal{A}_t$ to retain action-related features from both spatial and RF space conditioned on the semantic level features. Here $\mathcal{A}_t \in \mathbb{R}^{d_\mathcal{A}}$ provides an abstracted action embedding that represents a high-level semantic act instruction, where $d_\mathcal{A}$ is the dimension. We deployed the Transformer's attention mechanism [20] to learn such an abstract action representation:

$$\text{Atten}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, and V are three matrices that represent queries, keys, and values in the attention mechanism, and d_k is the dimension of K. We deploy self-attention layers to capture all the above relationships and obtain $\mathcal{A}_t$ as

$$\mathcal{A}_t = \text{Atten}(\mathcal{E}_t, \mathcal{E}_t, \mathcal{E}_t). \tag{3}$$

Thus, our Tasking Reasoner explores all the available historical and current observations to predict an abstract action $\mathcal{A}_t$ conditioned on high-level semantic features. This module will significantly improve the generalization of our method, because the semantic features are highly abstracted to work as an efficient and effective means to overcome the detail discrepancies in task environments.

C. Action Decoder

The abstracted action $\mathcal{A}_t$ can provide sufficient high-level tasking instructions, but it still cannot guide a robot to perform RFID inventory. Because there is still a lack of detailed information that is only available in the raw maps, such as the accurate spatial structural measurements of the surrounding environment. The proposed Action Decoder will anchor $\mathcal{A}_t$ by reasoning with the current maps m_t^s and m_t^v to predict the vivid action $a_t \in \mathbb{R}^2$ that enables the robot to efficiently perform the RFID inventory task in an unstructured environment.

The proposed Action Decoder consists of a spatial feature extractor and a Transformer-based decoder. We deploy two identical Fully-Convolutional-Networks (FCNs) as our spatial feature extractor. It receives the current maps, m_t^s and m_t^v, to extract features τ_t^s and τ_t^v, respectively. Then we concatenate them to form local features $\mathcal{T}_t = [\tau_t^s; \tau_t^v]$. Prior works [12], [21] have proven that FCNs can efficiently retain spatial information in given input images. Therefore, the extracted $\mathcal{T}_t$ provides sufficient spatial information of the current step t. Then, we deploy a Transformer-based decoder to instantiate $\mathcal{A}_t$ by reasoning the local spatial information $\mathcal{T}_t$. It comprises cross-attention layers and a fully connected (MLP) layer to predict an action a_t, given by

$$a_t = \phi(\text{Atten}(\mathcal{T}_t, \mathcal{A}_t, \mathcal{A}_t)), \tag{4}$$

where $\phi(\cdot)$ represents the final MLP layer that maps the output of cross-attention layers to the desired dimension of action a_t.

D. Policy Training

We train the proposed CMRM by Behavioral Cloning (BC) [22], a prevalent IL framework. Consider a given set

of expert demonstrations $\mathcal{D} = \{\chi_0, \chi_1, \cdots, \chi_n\}$. Here each demonstrated episode $\chi_i = \{(s_0, v_0, a_0^e), (s_1, v_1, a_1^e), \cdots\}$ is a sequence of multiple observation-action pairs (s_t, v_t, a_t^e), with s_t, v_t, and a_t^e denoting the environmental observation, RF observation, and expert action, respectively, at step t of the demonstration. We will uniformly sample the observation-action pairs from dataset $\mathcal{D}$ to form training sequences $\mathcal{Q} = \{\mathbf{Q}_0, \mathbf{Q}_1, \cdots, \mathbf{Q}_m\}$, with each sequence $\mathbf{Q}_i = \{(s_{t-T}, v_{t-T}, a_{t-T}^e), \cdots, (s_t, v_t, a_t^e)\}$ with a length of T. During the training, $\mathbf{Q}_i$ will be fed into CMRM to gain a predicted action a_t. Then, a simple Mean Squared Error (MSE) loss function will be deployed:

$$\mathcal{L} = \sum_{a_t \sim \pi} (a_t - a_t^e)^2. \tag{5}$$

By minimizing the error between a_t and the expert action a_t^e, we can train the Tasking Reasoner and Action Decoder of the proposed CMRM in an end-to-end manner.

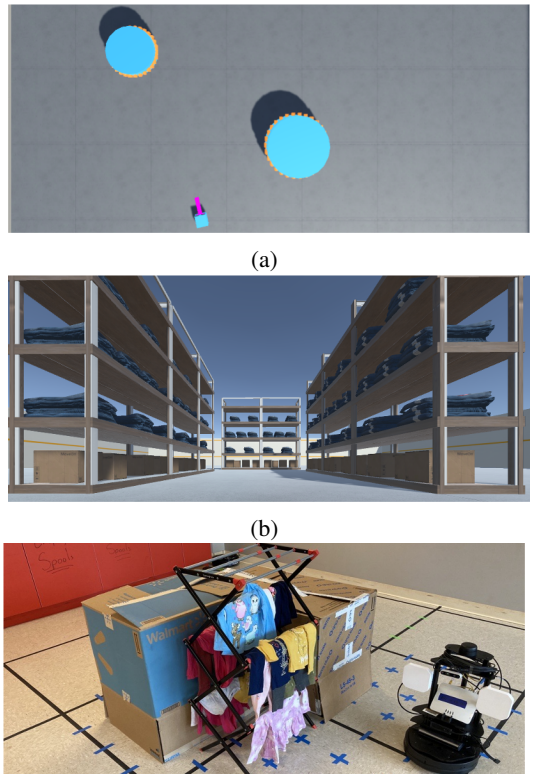
IV. Experimental Study

A. Experiment Setup

In this project, the robot aims to scan all the RFID tags in a given environment using fewer steps. An inventory task will be accomplished if the robot scans all RFID tags or when it reaches the maximum number of steps $M_{ax} = 5,000$. We assess the performance of CMRM by the learning efficiency for tasking policy and generalization to new tasking environments. The quality of a learned policy will be evaluated by the percentage of scanned RFID tags, the steps to complete the task, and the number of collisions during the inventory process. In our experimental setting, the robot could collide with other objects (such as racks), which simulates a robot equipped with bumper sensors to "softly" touch the environment. An optimized policy should perform a successful inventory task with less or without any collision. To thoroughly evaluate the proposed CMRM, our experiments will be conducted in the following three environments:

a) Conceptual Apparel Store: Inspired by previous works [3], [4], we implement a conceptual apparel store to collect the demonstration data and conduct experiments. This virtual environment is developed on the Unity3D platform, a widely used game engine. As shown in Fig. 4a, it is highly abstract and plain: blue cylinders, which hold orange cuboids, representing the apparel racks holding RFID-attached items; and the blue cube with a purple bar is the simulated mobile robot. The robot is equipped with a 2D Lidar and a simulated RFID reader, and their outputs match the requirements in Section II. Its purple bar indicates the heading direction. This conceptual store is a $25 \times 25\text{m}^2$ enclosed room with a random number, size, and position of cylinder-shaped racks.

b) Photo-realistic Virtual Environment: We also developed a photo-realistic virtual apparel store based on the Unity3D game engine. As shown in Fig. 4b, it is a photo-realistic virtual environment that can provide accurate physical properties (such as collision, gravity, inertia, etc.). The size of this store is also $25 \times 25\text{m}^2$. Standard four-layer racks with stored jeans and boxes will be randomly placed in the room, and the virtual Jeans and boxes all have RFID tags attached. This environment is a high-fidelity digital replica for a practice small apparel store to provide close to real tasking conditions and complexities. The same virtual robot as the one in the conceptual apparel store will be deployed here for the RFID inventory task.



(a) (b) (c)

Fig. 4: The experimental environments: (a) A conceptual virtual apparel store; (b) A photo-realistic virtual apparel store; (c) A mobile robot performs RFID inventory in our Lab. All tags are placed in a low position to be scanned by the RFID reader antennas.

c) A Real Environment: We also validate CMRM by deploying a LoCoBot mobile robot in our lab, a room with an area of $5 \times 5 = 25m^2$. As shown in Fig. 4c, several paper boxes and a clothing rack are randomly placed in the room. RFID tags are attached to each piece of clothing and each paper box, with a total of 51 tags in our room. The LoCoBot mobile robot equipped with a Zebra FX7500 RFID reader with two AN720 antennas is deployed for the inventory task. The Lidar sensor on the robot's top is the only active primary sensor to perceive the environment.

The training demonstrations $\mathcal{D}$ are only collected in the *conceptual apparel store* by manually operating the robot for the RFID inventory. We will evaluate the generalization of the proposed CMRM by whether the learned policy from this plain and simple simulated room can be transferred to the two other vivid environments.

B. Experiment Results and Analysis

a) Learning Efficiency: We first verify that CMRM can efficiently learn the RFID inventory from a few demonstra-

tions. In this experiment, we train a task policy π by the collected demonstrations $\mathcal{D}$ and evaluate π in the conceptual apparel store. During the demonstration and evaluation, racks' positions, sizes, and RFID tags' amount are randomly generated in each episode. To reduce the burden of manual demonstration collection, we limit the racks' positions to a $10 \times 10m^2$ area adjacent to the robot's start position. The entire demonstrated set $\mathcal{D}$ consists of 103 episodes of successful RFID inventory, a total of 247K robot operation steps. We create three more demonstration sets, $\mathcal{D}_{40}$, $\mathcal{D}_{60}$, and $\mathcal{D}_{80}$, by sampling the original $\mathcal{D}$ with 40, 60, and 80 episodes, respectively. To assess the learning efficiency of our CMRM, we independently train four policies, π_{40}, π_{60}, π_{80}, and π with $\mathcal{D}_{40}$, $\mathcal{D}_{60}$, $\mathcal{D}_{80}$, and $\mathcal{D}$. We set the sequence length $T = 64$ to train all the policies. Then, we deploy these policies in the conceptual apparel store and assess them regarding the percentage of scanned tags (i.e., the number of all scanned tags to the ground-truth number of tags in all episodes, including the failed ones), average tasking steps, and average number of collisions. During this evaluation, the environment is randomized in the same setting as we collect the demonstrations regarding the positions, size of racks, and RFID tags amount. Thus, experimental scenarios are different from the demonstrations but with similar distributions. Each policy is used for 30 episodes of inventory tasks, and the experimental results are presented in Table I.

TABLE I: Learning Efficiency

Policy	Scanned percentage	Average steps	Average collisions
$\mathcal{D}$	100%	2407.8	0.8
π_{40}	76.1%	4088.2	1.9
π_{60}	82.2%	4029	0.3
π_{80}	85.20%	3485.4	1.44
π	93.6%	3288.6	1.21

Table I shows that the proposed CMRM can learn an effective policy to conduct the task from 103 episodes of demonstration that only consume about $247K$ robot operation steps. Compared with the prior work [4], which requests around $250 \times 20,000 = 5,000,000$ steps to reach a similar task performance, the proposed CMRM can significantly improve the learning efficiency by about 20 times. We train our model of 18.9 million parameters on a single NVIDIA RTX3090 24GB GPU with 250 episodes, which takes 0.4 hours with a 75% occupancy rate, that's 10.7 TeraFLOPs in theoretical.

b) Generalizing to Unseen Scenarios: Next, we test whether CMRM generalizes to new environments. In this experiment, we deploy the well-trained policy, π, for RFID inventory tasks in two unseen environments: a new conceptual virtual apparel store and a photo-realistic apparel store. In the conceptual apparel store, the racks' positions are randomized in the $25 \times 25m^2$ room, with a greater variety of experimental scenarios out of the distribution of the training set $\mathcal{D}$. The photo-realistic apparel store provides a close to the real environment with complex object shapes, virtual appearances, and spatial structure. Illustrated by Fig. 4b, the photo-realistic virtual apparel store imposes significant tasking complexities and challenges for π, which was trained by the demonstrations collected from the conceptual store. We assess π with 50 episodes in each environment, and the results are summarized in Table II.

TABLE II: Zero-shot Generalization to New Environments

Environment	Scanned percentage	Average steps	Average collisions
Concept store	87.9%	3417	2.7
Photo-real store	83.8%	4002	3.8

In Table II, we find that CMRM provides a robust zero-shot generalization in both unseen environments by achieving the same level of performance as the simple and known environments that are shown in Table I. Qualitatively, we observe that the policy π enables the robot to precisely move toward correct target objects, such as apparel racks, clearly indicating that the learned policy conditioned on abstract features could overcome the significant discrepancies among seen and unseen environments and provide a robust task strategy.

c) Ablation Study & Comparison With Baseline Model: To evaluate the importance of several CMRM designs, we conducted 3 ablation studies and also compared them to a baseline model. The result is shown in Table III. We trained each model with 300 episodes and ran in the same conceptual apparel store. First, we trained the model π_{rt18} to evaluate the Semantic Feature Extractor design by replacing the pre-trained CLIP with a pre-trained Resnet18 with other parameters remaining the same. As we can see, the scanned percentage decreased largely, from 93.6% to 84.2%. Second, we trained the model π_{att} to evaluate the Spatial Feature Extractor by removing the FCNs in the Action Decoder and replacing the Transformer decoder layer with the Transformer encoder layer. Table III shows that its scanned percentage decreased slightly to 90.8%. Last, we implemented a *baseline model*, π_{rirl}, introduced in [3], which used simple FCNs with 128 hidden feature sizes. The result indicates this model behaves poorly on this task.

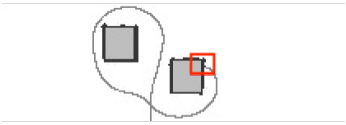
TABLE III: Ablation Studies & Compare to a Baseline Model

Policy	Scanned percentage	Average steps	Average collisions
π	93.6%	3288.6	1.21
π_{rt18}	84.2%	3993.04	1.18
π_{att}	90.8%	3305.08	1.05
π_{rirl} [3]	56.2%	4639.68	3.25

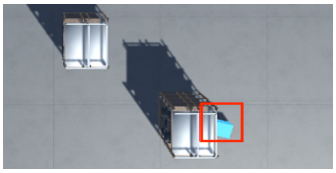
d) Real-Robot Experiment: To further evaluate the generalization of our model, we assess if the virtually trained policy can be directly deployed in a real robot. We use the same trained policy π to the LoCoBot robot in our Lab for the RFID inventory task as shown in Fig. 4c. We conducted 10 episodes by randomly placing the RFID-attached boxes and the robot in the room. Due to limited resources, we only placed 51 RFID tags in these experiments. In this small-scale experiment, the CMRM achieves a result of 96.3% scanned percentage with an average of $1,362.8$ steps and 0.75 collisions. Here, the high scan rate is due to our small group of tags, and fewer average

steps result from the smaller room. This result shows that the proposed CMRM bridges the "reality gap" thanks to its robust zero-shot generalization ability.

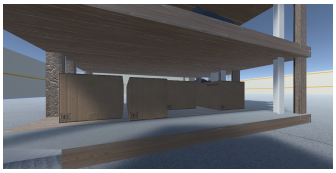
e) Failure Case Analysis: We also find that the most common failure source is "centimeter-detail" errors, and Fig. 5 illustrates an example of failed inventory in the photo-realistic store. We trained the policy to allow a soft touch with other objects. It works well in the training environment because the cylinder-shaped racks with smooth edges allow the robot to easily escape from the "touch" by making a light side movement. However, in the photo-realistic store, the rack's shape is more complex and usually has sharp corners. As the example illustrated in Fig. 5, when the robot collides with a sharp corner, the strategy of side movement will block itself to the rack.



(a) Trajectory map.



(b) Top view.



(c) Front view from the robot.

Fig. 5: A typical failed task is when the robot collides with a sharp corner of a photo-real rack: the collision point is marked by a red block in the trajectory map in (a) and the top view in (b).

V. Conclusions

In this work, we presented the CMRM, a cross-modal reasoning model that infers features and their relationships from multiple sensing spaces in historical and current observations to enable a robot to conduct RFID inventory in unstructured environments. The model extracts abstract semantic features from observations and learns a tasking policy conditioned on these high-level features. The experiments showed that an abstract feature-conditioned policy could provide a great generalization to an unseen tasking environment. It also closed the reality gap since the simulated trained policy could be deployed in real robotic applications. In our future work, we will extend the CMRM to enable robots for additional everyday tasks, such as detecting and refilling off-stock items on the sales floor.

References

[1] D. Delen, B. C. Hardgrave, and R. Sharda, "RFID for better supply-chain management through enhanced information visibility," *Wiley Prod. Oper. Manag.*, vol. 16, no. 5, pp. 613–624, Sept./Oct. 2007.

[2] J. Zhang, S. C. Periaswamy, S. Mao, and J. Patton, "Standards for passive UHF RFID," *ACM GetMobile: Mobile Computing and Communications*, vol. 23, no. 3, pp. 10–15, Sept. 2020.

[3] Z. Yu, J. Zhang, S. Mao, S. C. Periaswamy, and J. Patton, "RIRL: A recurrent imitation and reinforcement learning method for long-horizon robotic tasks," in *Proc. IEEE CCNC 2022*, Les Vegas, NV, Jan. 2022, pp. 230–235.

[4] ——, "Multi-state-space reasoning reinforcement learning for long-horizon RFID-based robotic searching and planning tasks," *Journal of Communications and Information Networks*, vol. 7, no. 3, pp. 239–251, Sept. 2022.

[5] M. Mirza, A. Jaegle, J. J. Hunt, A. Guez, S. Tunyasuvunakool, A. Muldal, T. Weber, P. Karkus, S. Racanière, L. Buesing *et al.*, "Physically embedded planning problems: New challenges for reinforcement learning," *arXiv preprint arXiv:2009.05524*, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2009.05524

[6] J. Achterhold, M. Krimmel, and J. Stueckler, "Learning temporally extended skills in continuous domains as symbolic actions for planning," *arXiv preprint arXiv:2207.05018*, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2207.05018

[7] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, Apr. 2016.

[8] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Machine Learning*, Lille, France, July 2015, pp. 1889–1897.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[10] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning *et al.*, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," in *Proc. Int. Conf. Machine Learning*, Stockholm, Sweden, July 2018, pp. 1407–1416.

[11] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, and C. Blundell, "Agent57: Outperforming the atari human benchmark," in *Proc. Int. Conf. Machine Learning*, Virtual Meeting, July 2020, pp. 507–517.

[12] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learning*, Auckland, New Zealand, Dec. 2022, pp. 894–906.

[13] P.-L. Guhur, S. Chen, R. Garcia, M. Tapaswi, I. Laptev, and C. Schmid, "Instruction-driven history-aware policies for robotic manipulations," *arXiv preprint arXiv:2209.04899*, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2209.04899

[14] A. A. Rusu, M. Večerík, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell, "Sim-to-real robot learning from pixels with progressive nets," in *Proc. Conf. Robot Learning*, Mountain View, CA, Nov. 2017, pp. 262–270.

[15] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. 2017 IEEE/RSJ Int. Conf. Intelligent Robots Systems*, Vancouver, Canada, Sept. 2017, pp. 23–30.

[16] J. Zhang, Y. Lyu, J. Patton, S. C. Periaswamy, and T. Roppel, "BFVP: A probabilistic UHF RFID tag localization algorithm using Bayesian filter and a variable power RFID model," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 10, pp. 8250–8259, Oct. 2018.

[17] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *IEEE Computer*, vol. 22, no. 6, pp. 46–57, June 1989.

[18] J. Zhang, Y. Lyu, T. Roppel, J. Patton, and C. Senthilkumar, "Mobile robot for retail inventory using rfid," in *2016 IEEE Int. Conf. Industrial Technology (ICIT)*, Bhubaneswar, India, Dec. 2016, pp. 101–106.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning*, Virtual Conference, July 2021, pp. 8748–8763.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS 2017*, Long Beach, CA, Dec. 2017.

[21] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Proc. Conf. Robot Learning*, London, UK, Nov. 2021, pp. 726–747.

[22] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. NIPS 1988*, Denver, CO, Jan. 1988, pp. 305–313.