2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Privacy-aware LLM-assisted Task Planning for Home Robots

Zhanjie Chen, Weihua Sheng*

Abstract— **Multi-modal large language models (LLMs) are expected to significantly enhance the intelligence of home service robots. However, reliance on cloud processing of raw visual data poses critical privacy risks. To address this problem, we propose a novel two-stage cloud-edge hybrid architecture for robots in domestic environments. This architecture employs a lightweight local LLM to perform sensitive content screening and semantic abstraction before transmitting the data to a more powerful cloud-based LLM for high-level planning and reasoning. Experiments with our end-to-end system demonstrate that it effectively protects a wide range of private data with minimal impact on task success rates. Without modifying cloud models, our approach offers a deployable performance–privacy trade-off for home robots, advancing safe and socially acceptable autonomy.**

I. Introduction

Task planning is one of the most important problems in robotics, which reflects the intelligence of robots. Robotic task planning has long been dominated by symbolic and rule-based approaches. Classical AI planners generate logically sound action sequences given a formal model of the world and goals. Planning Domain Definition Language (PDDL) [1] and Answer Set Programming (ASP) [2] are two representative technologies. These methods take a well-defined initial state, a goal state, and a set of allowed actions, and generate a sequence of actions that achieve the goal. The main advantage of this type of method lies in its ability to generate provably correct and often optimal plans within its strictly defined domain. However, they require expert knowledge to manually design the symbolic operators and cannot interpret everyday human instructions, making them ill-suited for the dynamic and varied demands of a domestic service environment where users expect to interact with the robot through natural language conversation.

Traditional methods force humans to conform to the rigid languages of machines (such as PDDL or state transitions), while the emergence of LLMs enables machines to understand human natural language. This paradigm shift, which endows robots with an agent-like capability for commonsense reasoning, is entirely absent in all classical approaches. The robotic systems can leverage the rich world knowledge and reasoning capabilities embedded in LLMs to translate high-level natural language instructions into executable sequences of robotic actions.

Such core reasoning capabilities heavily rely on large-scale foundation models hosted in the cloud (such as GPT-4 [3], Gemini [4] and DeepSeek [5]). For normal operation, the robot must transmit the multimodal data streams from its environment—including high-resolution images, video feeds, and user voice commands—to these external cloud APIs for processing [6], [7]. However, in home application scenarios, the data required to complete these tasks is inherently sensitive. Observation data (such as letters, medicine and financial documents) may be transmitted and stored on third-party servers. To protect privacy, the most intuitive approach is to perform sensitive information removal before the visual data is transmitted to the cloud. Currently, the most commonly used techniques for visual anonymization are blurring and pixelation [8], which distort image regions to conceal identifiable features. However, such anonymization techniques often result in the loss of task-relevant information and lack adaptability to context. For instance, blurring a medicine bottle to protect privacy may prevent the robot from completing tasks that rely on identifying or localizing that object. An emerging and more promising solution is context-aware or semantic-aware privacy protection, which applies protection measures based on the content and context of the data [9]. However, existing implementations of context-aware privacy typically rely on predefined rules or classifiers trained for a limited set of contexts. They lack the flexibility and common-sense reasoning required to handle diverse privacy scenarios in home environments. Therefore they are incapable of making nuanced judgments such as "This is a credit card, it is private" versus "This is a playing card, it is not private," especially when encountering new objects.

In this paper, we propose a privacy-aware LLM-assisted task planning system for home robots. The purpose of this system is to reduce the risk of privacy leakage in home-robot task planning using cloud LLMs, without compromising performance.

The main contributions of this paper are as follows:

- **A Novel "Local-Filter, Cloud-Decision" Hybrid Architecture**: We proposed and validated a novel two-stage robotic architecture. Using a lightweight local model for privacy filtering and semantic extraction, this architecture significantly reduces privacy risks in home environments without compromising the powerful planning capabilities of cloud-based LLMs.
- **An Effective Local Semantic Privacy Filter**: We designed and implemented an effective local privacy filter. Through the judgment of privacy attributes, the generation of descriptions, and the analysis of task relevance, the filter provides essential information for cloud-based planning while preserving user privacy. Experiments demonstrate that the filter excels at handling

Zhanjie Chen and Weihua Sheng (corresponding author) are with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK, 74078, USA (e-mails: zhanjie.chen@okstate.edu; weihua.sheng@okstate.edu).

visually ambiguous private objects (e.g., distinguishing a bank card from a playing card), with performance that surpasses general-purpose baselines and is comparable to state-of-the-art cloud models. Furthermore, the end-to-end system enabled by this filter minimizes the risk of privacy leakage while maintaining an excellent task success rate.

II. Related work

Recently, researchers have leveraged LLMs to endow robots with greater linguistic and reasoning capabilities. By tapping into pre-trained LLM knowledge, the robotic system leverages the rich world knowledge and reasoning capabilities to transform high-level natural language instructions into executable sequences of robotic actions. SayCan [10] uses a cloud-based LLM (PaLM) to suggest feasible next actions given a command, combined with a value function to ensure physical executability. Similarly, Huang et al. [11] demonstrate that an LLM can serve as a zero-shot planner, directly mapping natural language requests to sequences of actions without additional training. Some works use LLMs to produce intermediate representations for robots. For example, code-as-policies [12] converts LLM output into Python scripts calling robot APIs, and ProgPrompt [13] generates step-by-step plans in a structured language. With the advancement of LLM technology, vision language models (VLMs) have been gradually applied in the field of robotics. Driess et al. [14] introduce PaLM-E, which fuses visual input with language models to reason about both what the robot sees and what it should do. RT-2 [15] performs joint fine-tuning on a large-scale pre-trained VLM using both web-scale data and robotic data, enabling knowledge transfer to comprehend novel concepts absent from its robotic training set and to carry out elementary reasoning. VoxPoser [16] leverages LLM-generated code to call VLMs and geometric perception, synthesizing composable 3D value maps in the voxel space. These maps are then used by a model-based planner to optimize trajectories, allowing zero-shot execution of diverse manipulation tasks. These LLM-driven approaches achieve advanced generality, handling diverse user requests that traditional planners cannot. However, such approaches typically rely on large-scale online models and cloud computing, uploading perceptual data or queries to third-party services (e.g., the OpenAI API). Such architectures may transmit sensitive audio and visual information to the cloud when invoking cloud-based LLMs, posing serious privacy and security risks. However, current LLM-based robotics research rarely addresses this issue, with the primary focus still placed on capability and performance.

Given the inherent privacy risks of current LLM-based robotics, we survey and evaluate relevant work in the domain of privacy protection. Visual anonymization techniques protect privacy by modifying or removing identifiable elements from sensor data. Intuitive methods like blurring or pixelation of faces are common in practice but can severely degrade image utility. More advanced approaches use generative models to inpaint or replace sensitive regions. DeepPrivacy2 [17] builds a GAN-based system that automatically obscures people in camera footage while preserving scene context. Cryptography is another feasible technique for privacy protection, such as homomorphic encryption, which allows computations on encrypted data. Luedeman et al. [18] demonstrate this concept for drone navigation, using homomorphic encryption to compute collision avoidance on encrypted flight paths so that sensitive location data remain secret. However, fully homomorphic encryption has high computational complexity and is currently difficult to apply to real-time visual or language processing tasks. In addition, federated learning keeps training data localized to preserve privacy. Instead of uploading raw data, robots collaboratively train shared models by only exchanging model parameters. Marino et al. [19] implement a ROS 2–based federated learning pipeline for robotics, enabling multiple robots to collaboratively train a global model without sharing raw data. Federated learning effectively protects privacy during the model training phase and has been widely applied in multi-robot or Internet of Things (IoT) scenarios. However, these methods are rarely integrated with visual-language understanding frameworks in practical applications. Existing anonymization or encryption methods usually process data before it is understood, and are not yet built into the reasoning process of robots that understand natural language. Recent studies have highlighted the need for such integration. Taras et al. [20] advocate "inherently privacy-preserving" perception in future autonomous systems. In the context of robot task execution, robots should semantically recognize sensitive information and integrate this ability into task planning and communication.

III. Methodology

A. System Overview

Fig. 1 shows the proposed system framework which consists of four modules. Among them, the *Cloud decision and interaction* module receives user instructions and the privacy-filtered image data and outputs a task plan. The robot executes each step of the global plan and acquires images. The *Image processing* module is designed for object boundary detection, segmentation, and masking of privacy-sensitive items. The *Privacy handling* module is a privacy filter for home environments, which is designed based on a fine-tuned local LLM. For each item in the image, this module outputs a detection result, including its privacy attribute, object description, and task relevance. The *Local information integration* module is the bridge between the local LLM and the cloud LLM which is responsible for invoking the *Image processing* module to apply privacy mask based on the attribute, and check if the description uploading is needed for each privacy-sensitive item. It then uploads the filtered image and necessary description information to the cloud LLM.

B. Image Processing Module

Each item in an image is evaluated by the *Privacy handling* module. To avoid the interference between objects, we

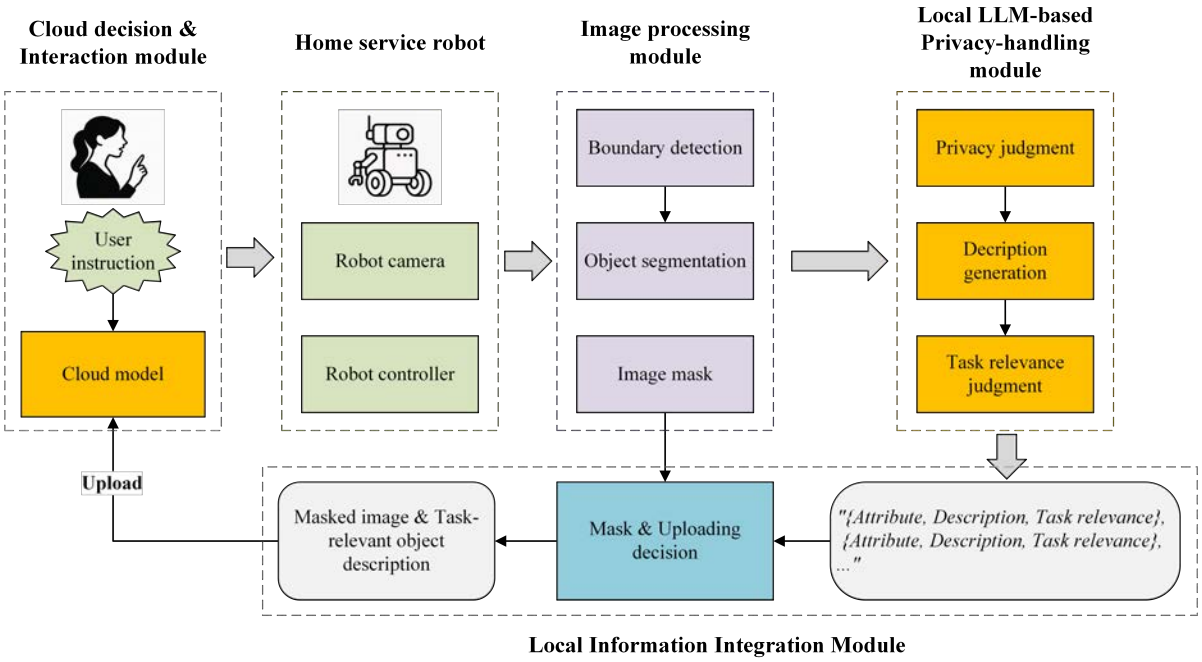

Fig. 1. Overview of the system framework

perform object segmentation as a pre-processing step. LLMs (particularly local LLMs) currently suffer from insufficient real-time performance and accuracy in boundary detection and instance segmentation. Therefore, a dedicated *Image processing* module is designed to handle this pre-processing step. Fig. 2 shows the design and workflow of this module. First, YOLOv8 [21] is adopted to perform real-time boundary detection on the input image. Although the YOLO model can also provide classification outputs, this functionality is not used because the context-aware recognition required for privacy detection is handled by the proposed *Privacy handling* module, which is far more capable for this specific task than YOLO's general-purpose classifier. Next, the image is segmented according to the coordinates of each bounding box, ensuring that each segmented image contains only one object, which is then fed into the *Privacy handling* module. Meanwhile, this module integrates a mask function: based on the output of the *Privacy handling* module, the system invokes the function to mask objects judged as privacy-sensitive items in the original image.

C. Privacy Handling Module

The *Privacy handling* module functions as a local privacy filter: it takes as input the segmented single-object images from the *image process* module and outputs, for each object, its privacy attribute, a refined object description, and task relevance. The output is defined according to the following rules:

- **Privacy attribute**: Using the general knowledge of a multimodal LLM, the module determines whether the item involves privacy in a home scenario (e.g., a bank card is *yes*; a bottled water is *no*). Output format: *yes/no*.

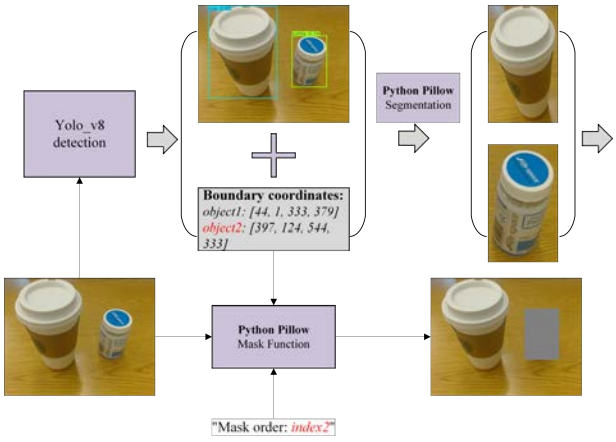

Fig. 2. Pipeline of the image processing module

- **Object description**: Identify the target and generate concise and objective target description statements.
- **Task relevance**: By combining the understanding of the task planning from the cloud model with the object in the image, determine whether the item is relevant to the current task. Output format: *yes/no*.

The attribute judgment is the most critical output of the *Privacy handling* module, as it directly determines whether the image has any risk of privacy exposure. However, due to parameter size constraints (typically smaller than 8B), LLMs, especially local LLMs deployed on the robot, cannot guarantee the expected accuracy across all objects. To improve its performance, we fine-tune the local LLM on a custom dataset of commonly encountered household items.

The fine-tuning focuses particularly on low-discriminability object pairs (e.g., bank cards vs. playing cards), which are the most error-prone cases. The specific composition of our training dataset is detailed in the *Experimental Setup* section.

In this work, "sensitive content" is defined as any visual object that may reveal personally identifiable, financial, medical, legal, educational, or account-related information in a home environment. To improve reproducibility, we operationalize this definition with three screening criteria: (1) *identity linkage* (can the object be linked to a specific person), (2) *harm potential* (could exposure cause financial, legal, or safety harm), and (3) *task necessity* (is disclosure required to complete the current user instruction). The module marks an object as privacy-sensitive when criteria (1) or (2) are satisfied; criterion (3) is then used by the local integration module to determine whether a textual description should be uploaded to the cloud.

To further optimize performance, we design a classification-style prompt for the local LLM. The prompt describes the role, task, and output format of the model to improve stability and accuracy while ensuring that the results are machine-friendly. The key components of the prompts as follows:

- **Role** : bind the model to a household service robot to constrain reasoning within the target scenario.
- **Visual focus**: "identify clearly visible main objects", which is aligned with the object segmentation strategy in the pre-processing stage to avoid judgment interference.
- **Attribute judgment**: the model must determine whether the object involves privacy in a household scenario, with the output strictly defined as yes or no.
- **Description generation**: produce a clear, non-sensitive textual description that supports cloud-based task planning.
- **Task relevance**: combine the object understanding with the task planning from the cloud model to decide relevance, with output yes or no.
- **Output template**: the output must strictly follow the format: [*attribute, object description, task relevance*]. e.g. [*yes, credit card, no*].

D. Local Information Integration Module

The *local information integration* module serves as a critical bridge connecting the local LLM with the cloud LLM. It receives the per-object analysis from the *Privacy handling* module and generates the final, privacy-filtered data for upload. This process is governed by a strategy designed to effectively balance user privacy protection with the performance of cloud-based task planning.

Specifically, the module executes the following operations based on each object's privacy judgment and task relevance judgment:

- **Private but Not Task-Relevant Objects (Attribute: yes, Task relevance: no)**:The module applies a mask to the corresponding image region to protect user privacy. Moreover, the textual description of the object will not be uploaded to the cloud.
- **Private and Task-Relevant Objects (Attribute: yes, Task relevance: yes)**: The module masks the private region, while the object's textual description is retained and uploaded.
- **Non-Private Objects (Attribute: no)**: The image region remains in its original state with no mask applied. The object's textual description is also not uploaded.

E. Cloud Decision and Interaction Module

In this module, the cloud LLM is utilized to translate a user's natural language command (U_{task}) into a structured, executable task plan for the robot. It is responsible for global task decomposition and dynamic decision-making. In each decision cycle, the module processes the user's high-level command and the uploaded filtered image, then generates a structured data object containing a detailed action plan and the next directive for the robot. The specific inputs and outputs for this module are defined as follows:

- **Input:**
  1) **User's Overall Task (U_{task}):** A natural language string describing the final goal of the task.
  2) **Processed Environmental Perception ($E_{\text{perception}}$):** Data which includes a selectively masked image (I_{masked}) and a set of textual descriptions for task-relevant private objects (D_{privacy}).
- **Output:** A standardized, structured data object ($A_{\text{structured}}$) containing the following key fields:
  1) **Overall Plan:** An ordered list of sub-task steps and their corresponding status (e.g., Pending, Completed). Meanwhile, the plan is transmitted to the *Privacy handling* module to determine task relevance.
  2) **Next Action Directive:** A standardized, high-level semantic command to be sent to the robot.
  3) **Reasoning:** The logical explanation for generating the next action.

To guide the LLM in performing the planning task, our prompt is designed with four core components:

- **Role Assignment:** the LLM is assigned the role of a "dynamic planner for a home service robot."
- **Multi-modal Fusion Instruction:** the prompt instructs the model that it must combine information from the masked image (I_{masked}) and the corresponding privacy text descriptions (D_{privacy}) to achieve a comprehensive understanding of the environment.
- **Task Decomposition Structure:** the model should first generate a step-by-step "Overall Plan" before providing the "Reasoning" for the next action, thereby guiding its planning process.
- **Output Format Enforcement:** the prompt strictly enforces a predefined, standardized output format. This ensures that the generated content can be reliably parsed and executed by downstream systems.

Fig. 3. Overview of experimental environment

IV. Experiments

A. Experimental Setup

All experiments were conducted in a controlled laboratory environment designed to simulate a typical home. As shown in Fig. 3, this mock apartment is comprised of several core living areas, including a living room, bedroom, kitchen, and bathroom. This setup provides an ideal platform for the repeatable testing of our system in realistic domestic scenarios.

For our experiments, we developed a robot platform named ASCCBot. As illustrated in Fig. 4, ASCCBot is composed of two main components: a head and a body. The head is a tabletop unit that functions as a detachable, standalone conversational robot. The body integrates a Pioneer-3DX mobile base, a Hokuyo 2D LiDAR, and an Intel RealSense D415 RGB-D camera as sensing devices, along with an NVIDIA Jetson Orin NX board serving as the processing unit.

The proposed system is implemented within the Robot Operating System (ROS) framework, which manages communication between all modules. Our proposed *Privacy handling* model is the result of fine-tuning the local model gemma-3n-E4B (8B). The fine-tuning process utilized a custom dataset collected from real-world home scenarios. The dataset, detailed in Tab. I, comprises 15 categories of common objects with low discriminability - specifically 10 private objects and 5 non-private objects. We collected 15 images for each category, resulting in a total of 225 images for training.

B. Performance of Privacy Detection

1) Experimental design: This experiment is designed to evaluate the core privacy detection capability of our local LLM-based privacy handling module, which involves two key tasks: privacy attribute judgment and object description generation. We collected three pairs of low-discriminability objects in a realistic environment, as depicted in Fig. 5. These pairs belong to the categories of cards, bottles, and documents. Each pair consists of a private item (e.g., a driver's license) and a visually similar, non-private item (e.g., a playing card), designed to test fine-grained semantic differentiation. All images were captured at a resolution of 640x480.

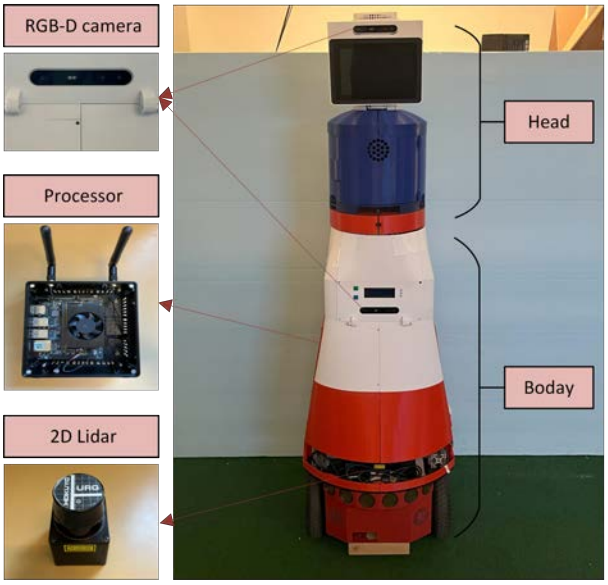

Fig. 4. The ASCCBot mobile robot as the experimental platform

TABLE I
Object Categories for Fine-tuning on Low-Discriminability Items

Categories	Private Objects	Non-Private Objects
Card	bank card driver's license student ID card	playing card
Picture	family photo ID photo	cartoon character photos
Documents	business contract bank statement insurance card prescription slip	magazine book
Bottle	medication bottle	vitamins

We evaluated the performance of our proposed method on the low-discriminability image pairs by comparing four models (3 local, 1 cloud). To establish a performance benchmark, we also included ChatGPT-5 as a state-of-the-art cloud-based reference.

2) Results analysis: The qualitative and quantitative results of our experiments are presented in Table II and Table III, respectively. As summarized in Table III, all local models tested achieve over 65% in accuracy of privacy judgment (APJ), which demonstrates the fundamental feasibility of building a semantic-level privacy filter using local LLMs. Specifically, the Qwen2.5-VL model performed well in terms of accuracy of object description (AOD) and the baseline gemma-3n-E4B was respectable in APJ, with each achieving a score of 83% in their respective area of strength. Despite this, each model had a disqualifying weakness. The Qwen2.5-VL model committed significant privacy judgment failures (e.g., misclassifying the insurance card as non-private), resulting in a low APJ (66.7%). Conversely, the baseline gemma-3n-E4B produced overly generic descrip-

Fig. 5. Low-Discriminability image pairs. Note: To protect personal privacy for publication, sensitive information in images (1-a) and (3-a) has been manually anonymized. The original, un-anonymized images were used in our experiments.

tions (e.g., "medical product" instead of "medicine bottle"), leading to a poor AOD of 50%. In contrast, our fine-tuning methodology successfully addresses these shortcomings. Our fine-tuned model is the only local model to achieve a perfect 100% score in both APJ and AOD, a performance on par with the state-of-the-art ChatGPT-5 reference. This result confirms that while using local LLMs for this task is a viable direction, targeted fine-tuning can effectively enhance the performance, providing a truly reliable and robust solution for home privacy protection.

C. System Performance Test in Real Applications

1) Experimental design: This experiment evaluates the end-to-end performance of our proposed system in completing realistic user tasks. We designed four distinct test cases to represent a spectrum of operational scenarios with increasing complexity, ranging from a simple non-private task to a complex multi-privacy scenario. Tab. IV details these four cases, including the specific user instructions and target objects for each, while Fig. 6 shows the corresponding visual inputs captured in a real-world home environment.

To provide a comprehensive and in-depth analysis, we compare the performance of our proposed system against three crucial baselines: Raw Data Baseline (raw images are transmitted), Naive Anonymization (Naive Anon.; all private objects are masked and no descriptions are sent), and Full Description (Ablated) (all private objects are masked, but all of their descriptions are sent).

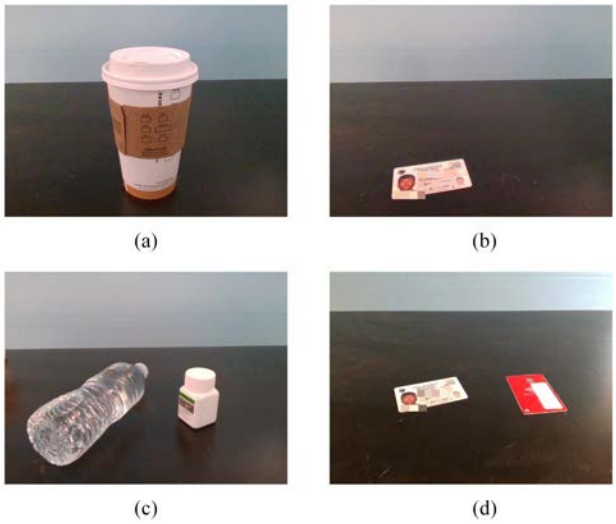

Fig. 6. Original inputs for each task. a)coffee cup; b)driver's license; c)water and medication; d)driver's license and bank card.

We designed two core metrics for this assessment: Task Success Rate (TSR), defined by the cloud model's ability to generate a valid plan based on the received data, and Privacy Upload Ratio (PUR), the percentage of private object descriptions uploaded from the scene. Together, these metrics measure the trade-off between task functionality and privacy preservation across the different strategies.

It should be noted that as the objective of this experiment is to validate the proposed privacy-protection strategy, the scope of our evaluation does not cover the physical navigation and manipulation required for task completion.

2) Results analysis: The results of the system performance evaluation are presented in two parts. Fig. 7 illustrates the visual outputs of our local *Privacy handling* module, showing how the original inputs from Fig. 6 are processed before being sent to the cloud. Table V summarizes the TSR across all four scenarios, while Table VI details the PUR for the scenarios involving sensitive items.

As shown in Table V, three of the four systems achieved a perfect 100% Overall TSR: the Raw Data Baseline, the Full Description baseline, and our proposed system. This indicates that our privacy-protection mechanism does not compromise the robot's ability to complete its tasks. In contrast, the Naive Anon. system failed any task where the target object was private (Single-Privacy and Multi-Privacy tasks), resulting in a low TSR of 50%, highlighting its functional limitations.

Table VI reveals the different privacy trade-offs among the systems that achieved high task success rates. Both the Raw Data Baseline and the Full Description baseline consistently uploaded all available private information in relevant scenarios, maintaining a 100% PUR. This indicates a significant leakage of unnecessary private data. In contrast, our proposed system demonstrates an intelligent and context-aware approach to data handling. In the Single-Privacy Task, it correctly identified the private item as task-relevant and

TABLE II

OUTPUT COMPARISON OF DIFFERENT MODELS ON HOUSEHOLD COMMON LOW-DISCRIMINABILITY IMAGE PAIRS

Categories	Images	Qwen2.5-VL	gemma-3n-E4B (Baseline)	Our Fine-tuned Model	ChatGPT-5 (Cloud Reference)
Cards	1-a	Yes, Oklahoma learner permit	Yes, ID card	Yes, driver's license	Yes, driver's license / learner permit
	1-b	No, playing card	No, playing card	No, playing card	No, playing card
Bottles	2-a	No, medicine bottle	Yes, medical product	Yes, medicine bottle	Yes, medicine bottle
	2-b	No, Omega-3 Fish Oil	Yes, Fish Oil	No, Omega-3 Fish Oil	No, Omega-3 Fish Oil
Documents	3-a	No, Insurance ID card	Yes, privacy information	Yes, Insurance card	Yes, Insurance document
	3-b	No, Digital Signal Processing and the Microcontroller	No, book	No, book-Digital Signal Processing and the Microcontroller	No, Textbook (Digital Signal Processing and the Microcontroller)

TABLE III

OVERALL PERFORMANCE COMPARISON OF DIFFERENT MODELS

Models	Parameter size	APJ(%)	AOD(%)
ChatGPT-5	-	100.0	100.0
Qwen2.5-VL-7B	8.29B	66.7(4/6)	83.3(5/6)
gemma-3n-E4B	7.85B	83.3(5/6)	50.0(3/6)
gemma-3n-E4B(Our)	7.85B	**100.0**	**100.0**

TABLE IV

EXPERIMENTAL CASES AND CURRENT TASK PROGRESS

Task categories	Input images	Instructions	Targets
Non-privacy	a	Could you get me the coffee on the table?	coffee cup
Single-privacy	b	Help me find my driver's license.	driver's license
Mixed-privacy	c	Bring me a bottle of water.	bottled water
Multi-privacy	d	Could you help me find my bank card?	bank card

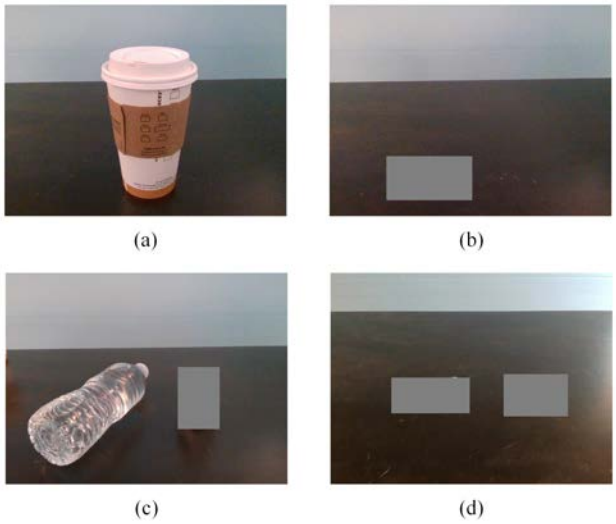

Fig. 7. Masked images of objects involving privacy

uploaded its description (100% PUR). In the Mixed-Privacy Scenario, it recognized the private item as irrelevant and withheld its description, achieving an optimal 0% PUR. Furthermore, in the more complex Multi-Privacy Scenario, it intelligently uploaded only the description of the task-relevant private target, resulting in a 50% PUR.

In conclusion, these results quantitatively demonstrate that our proposed system is the only approach capable of achieving the highest Task Success Rate while simultaneously minimizing the leakage of private information, proving its superior ability to intelligently balance functionality with privacy protection.

V. SECURITY DISCUSSION AND LIMITATIONS

Our method is designed to reduce privacy leakage in the application layer by minimizing sensitive visual content transmitted to cloud LLMs. Under this threat model, even if external interception occurs during end-to-end transmission, the attacker only observes masked images and selectively retained text, which lowers direct privacy exposure compared with raw-data upload. However, this mechanism is not a replacement for transport-layer security. In practical deployment, it should be combined with secure communication and system hardening to defend against network attacks.

Our current evaluation emphasizes short-horizon task completion and privacy-preservation effectiveness in representative home scenarios. As a next step, we are extending the protocol to include long-duration operation metrics, including multi-hour and multi-day stability, memory footprint, and power consumption.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the critical privacy challenges posed by the use of cloud-based LLMs in home service robots by proposing a novel "Local-Filter, Cloud-Decision" hybrid architecture. At the core of this architecture is a lightweight, fine-tuned multi-modal LLM that runs locally to identify, mask and selectively extract semantic information from privacy-sensitive content before any visual data is uploaded. Our experimental results validated this approach on two levels. First, in component-level tests, our fine-tuned local model demonstrated accuracy comparable to SOTA cloud models for both privacy judgment and description generation on challenging, low-discriminability objects. Second, in end-to-end system tests, our proposed system was the only approach to achieve a 100% TSR while contextually minimizing the PUR (to as low as 0% in some cases).

TABLE V

TASK SUCCESS RATE (TSR) COMPARISON ACROSS DIFFERENT SYSTEMS AND SCENARIOS

Methods	Baseline Task (Non-Private)	Single-Privacy Task	Mixed-Privacy Scenario	Multi-Privacy Scenario	Overall TSR
Raw Data Baseline	Success	Success	Success	Success	**100%**
Naive Anon.	Success	**Fail**	Success	**Fail**	**50%**
Full Description	Success	Success	Success	Success	**100%**
Our	**Success**	**Success**	**Success**	**Success**	**100%**

TABLE VI

PRIVACY UPLOAD RATIO (PUR) COMPARISON (%)

Methods	Single-Privacy Task	Mixed-Privacy Scenario	Multi-Privacy Scenario
Raw Data Baseline	100%	100%	100%
Naive Anon.	0%	0%	0%
Full Description	100%	100%	100%
Our	100%	**0%**	**50%**

This indicates that our system strikes an intelligent and superior balance between robot functionality and user privacy protection. Future work will focus on two main directions. First, we will expand the fine-tuning dataset of household privacy items to cover more diverse scenarios and enhance the generalization of local LLMs. Second, we plan to explore optimization techniques, such as model quantization, to improve real-time performance on resource-constrained robotic platforms.

ACKNOWLEDGMENT

This project is supported by the National Science Foundation (NSF) Grants CPS 2212582 and TI 2329852.

REFERENCES

[1] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins Sri, Anthony Barrett, Dave Christianson, et al. Pddl—the planning domain definition language. *Technical Report, Tech. Rep.*, 1998.

[2] Gerhard Brewka, Thomas Eiter, and Mirosław Truszczyński. Answer set programming at a glance. *Communications of the ACM*, 54(12):92–103, 2011.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[5] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[6] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

[7] Yeseung Kim, Dohyun Kim, Jieun Choi, Jisang Park, Nayoung Oh, and Daehyung Park. A survey on integration of large language models with intelligent robots. *Intelligent Service Robotics*, 17(5):1091–1107, 2024.

[8] Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 15(1):101–116, 2001.

[9] Jiayu Shu, Rui Zheng, and Pan Hui. Cardea: Context-aware visual privacy protection for photo taking and sharing. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 304–315, 2018.

[10] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[11] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[12] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.

[13] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*, 2022.

[14] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[15] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[16] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[17] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1329–1338, 2023.

[18] Allan Luedeman, Nicholas Baum, Andrew Quijano, and Kemal Akkaya. Privacy-preserving drone navigation through homomorphic encryption for collision avoidance. In *2024 IEEE 49th Conference on Local Computer Networks (LCN)*, pages 1–6, 2024.

[19] Roberto Marino, Lorenzo Carnevale, Maria Fazio, and Massimo Villari. Make federated learning a standard in robotics by using ros2. In *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies*, pages 1–6, 2023.

[20] Adam K Taras, Niko Sünderhauf, Peter Corke, and Donald G Dansereau. Inherently privacy-preserving vision for trustworthy autonomous systems: Needs and solutions. *Journal of Responsible Technology*, 17:100079, 2024.

[21] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.