Human-in-the-Loop Perception: Lightweight Segmentation for Low-Cost Robotic Systems

Shengyan Xu *

Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, United States

* Corresponding Author Email: shengyan.xu94@gmail.com

Abstract. A lightweight Human-in-the-Loop (HITL) perception framework is proposed in this paper to improve the segmentation accuracy of low-cost service robots operating in complex real-world environments, such as restaurants. Instead of using high-cost sensors and many training rounds, prompt-based segmentation with SAM-style embeddings can be adopted to reduce human intervention to an extent, and only one or two clicks are needed for real-time mask refinement. We simulate human feedback via click-based heuristics and evaluate the model on the COCO 2017 validation set under low-resolution constraints. Experiments have shown that the first 1-3 clicks provide a significant increase in accuracy (IoU +27% relative), and then it reaches a plateau; therefore, a small number of human interventions can be used instead of hardware upgrades while maintaining high efficiency and low latency. Therefore, HITL perception offers a relatively low-cost and scalable path to a robust deployment of low-cost robots. The framework has also supported the application of Chinese restaurant robots in practice and provided a feasible direction for the development of service robots.

Keywords: Human-in-the-Loop (HITL), Lightweight Segmentation, Low-Cost Robotics, Restaurant Robot, Interactive Perception, CAPTCHA Learning, Cost-Efficient AI, Human–Machine Collaboration.

1. Introduction

In recent years, many robots have become part of daily life in many cities across China. You can see them in hotels, restaurants, and some shopping malls. These machines seem very smart, but most of them rely on very simple visual systems. This paper noticed an interesting contrast in life time. Hotel delivery robots usually work very perfectly. They follow long corridors, take the elevators, and deliver food to rooms, with no problem. Because the environment is clean and stable, and there are few moving obstacles.

However, restaurant robots face a completely different reality. They need to move through narrow spaces between chairs, people, cases and tables. Sometimes they encounter luggage or children are running around. In these cases, the camera often have mistakes — it cannot clearly tell what is a chair, what is a person, or where the safe path is. And the robot stops, spins in place, or calls for human help.

This difference made me think: What really makes a robot “smart”? It is not only about better hardware or a more powerful algorithm. Sometimes, the key is human understanding. Humans can recognize situations that machines cannot. For example, when Google asks us to “click all the traffic lights” in a CAPTCHA test, we are actually teaching the computer what a traffic light looks like. So that same idea — using small, quick human actions to correct machine errors — can also help robots.

Research is being conducted on how to improve the performance of low-cost robots in difficult environments by using a Human-in-the-Loop (HITL) perception system. The basic idea is that the robot will handle most of the work independently, and only when it encounters a problem due to confusion will a person need to intervene for a short time to adjust the visual recognition. It will be more convenient and less expensive to retrain the model or replace the sensor.

Many robots in Chinese restaurants and hotels are not using this type of learning, according to this paper. The software will not be modified after it is published. When a robot is stuck, the staff simply move it out of the way instead of teaching it why it broke down. It is thus uninformative. A relatively

simple feedback mechanism, such as a tablet with "correct" and "incorrect" buttons, could be used to teach these robots from their work.

Therefore, this paper will demonstrate that by combining human intuition with the efficiency of machines, intelligent, dependable and economical robots can be created. It is a technical solution to address the problems of sharing and learning from the same world for people and robots.

1.1. Interactive Segmentation and Human-in-the-Loop Learning

Interactive Segmentation has been one of the problems that computer vision has studied for a long time.

The first systems, such as GrabCut, manually marked the object and then had the computer complete the rest [1]. Later, deep learning introduced new models such as DEXTR, f-BRS and RITM [2]; these models respond to user clicks, and with each click, the mask becomes more accurate. It is a natural phenomenon; some people have pointed out problems, and the model has slowly begun to correct them [3].

The idea of a human-in-the-loop has now extended beyond division. Now it is required that modern AI has. In a human-in-the-loop (HITL) mode, people do not label all the data at the beginning but correct only the cases where the model is uncertain. Therefore, the machines will learn faster and have fewer errors. It is an automatic but not fully autonomous system.

Recently, foundation models such as Segment Anything (SAM) have also been released. SAM splits the heavy part and the image embedding into a light part for mask update. That is to say, it is embedded once and clicked many times; thus, fast human feedback can be achieved with a very small delay. A user can click multiple times to change the mask, and then the robot will immediately update its understanding. Simple to use and suitable for a small device.

I believe this will be the kind of cooperation required by real robots. A robot can be taught by a few human clicks in a specific location without extensive modifications or training. This is the same idea that has driven my studies: find a way to expand the coverage of human attention for deficiencies in machine vision without increasing costs.

1.2. Human Verification Systems as Global-Scale HITL Models

On the internet, millions of people interact with computers every day without even noticing what they have done.A simple example, The CAPTCHA test — those small boxes that ask you to “click all the pictures with traffic lights” or “choose the images that show cars.”Systems like these, used by Google, Steam, and Riot Games, are in fact large-scale versions of human-in-the-loop (HITL) perception.

The main goal is to stop robots, but at the same time every user click quietly help the computer-vision models to see the world little better and better.When you solve these tests, you are not only proving that you are a ture human —but you are also correcting the model’s mistakes.You might mark the outline of a car, the tires, the end of a spoiler, or a traffic light partially obscured by a tree, because some images are not entirely composed of squares, and these tiny movements become new training data.

Billions of these micro-corrections happens every day, to creating a huge, free source of high-quality visual information.It is one of the biggest examples of people and machines are learning together. Without any formal training session or lab environment.This idea connects closely to what happens in robotics.[4]

A CAPTCHA on the web and a robot in a restaurant face a similar problem: both meet visual scenes that are hard to understand.The CAPTCHA asks for human help through clicks, and the robot could do the same when it cannot recognize an object.

For example, a restaurant robot could stop and show two images — “is this a table or chair?” — and a human could pick the right one.That quick answer could immediately fix the robot’s path and also teach its vision model for future cases.So, the online systems lesson is simple:a few clicks from many humans can create massive accuracy gains for machines.For the low-cost robots, adding a small

feedback interface would work the same way — continuous correction without expensive retraining or extra sensors.This is how global web platforms and small service robots can share the same idea of human-in-the-loop learning, just at very different scales.

1.3. Industrial Context: The Gap in Low-Cost Robotic Perception

In China, the application of service robots has become quite common, covering multiple scenarios such as hotels, restaurants, hospitals, and shopping centers. However, these commercial service robots differ fundamentally from high-tech industrial robots used in factories: industrial robots rely on multiple sensors, precise calibration, and regular retraining, while most service robots in hotels and restaurants are constructed inexpensively, simply, easy to assemble, debug, and repair, and typically run on fixed visual models that cannot self learn or update after deployment. As a result, the cameras of these robots often mistake chairs for humans or fail to recognize glass walls and reflective floors, causing the robots to stop or make wrong turns in the hallway without reason. In hotels with relatively clean and predictable environments, such problems are not yet prominent, but in restaurants with frequent personnel flow and diverse items, robots getting lost is more common. Faced with these malfunctions, the staff usually only move or restart the robot without correcting the system itself. Although this approach can temporarily maintain operation, it prevents the robot from ever understanding why it made a mistake. If the system can support one-click correction, such as allowing staff to click on annotations like "This is a table leg" or "This is part of the floor," then this simple and guided operation can gradually teach robots to cope with complex changes in the real environment and continuously improve performance over time without the need for new hardware or expensive retraining. This is exactly the value that the "human in the loop" perception system can provide - it builds a simple bridge between human intuition and machine learning, with the goal of not making robots perfect in a day, but helping them become smarter with every mistake.

2. Methodology: Lightweight Human-in-the-Loop Segmentation Framework

2.1. Motivation and Core Idea

The purpose of this paper is to design a perception system that is both lightweight and feasible; at the same time, it should maintain a high segmentation accuracy for a low-cost robot platform and meet the requirements of practical application. Therefore, a relatively low-cost sensor is used and a single round of training is required for the robot to perform visual recognition. In short, a small number of manual corrections can help the robot learn more from the current visual environment and correct recognition errors.

The first few studies were based on field observations carried out by a single person. In the operating environment, service robots generally have poorer performance in areas with strong light or many people. For example, when robots are in a reflective floor environment or when the backrest of a chair overlaps with a person in the camera's view, it is difficult for them to determine whether these areas are obstacles or not. At this time, robots usually have to stop moving forward, rotate in place and wait for external intervention. The hardware arrangement of these application areas is generally relatively simple and changes infrequently; thus, retraining or readjusting the camera is both costly and impractical in practice. Therefore, the central problem of this paper is as follows: Is there a simple way to help robots obtain more accurate visual information without adding new hardware or modifying existing models? This problem serves as the foundation for the current study.

Given the problems mentioned above, this paper proposes a solution that aligns with the idea of "embedding once, clicking multiple times" and is inspired by prompt-based segmentation models, such as the Segment Anything Model (SAM). The workflow of this scheme is as follows: First, for every frame in the image, the robot generates a visual embedding representation; this is the most computationally expensive part of the whole process. Then, whenever the robot is uncertain about recognising the scene, an operator can add a prompt message by clicking once or several times on the displayed screen, such as "This is a chair" or "This is the ground". These human click signals serve

as the input for a lightweight mask decoder, and then the segmentation results are updated promptly. The entire system will still have good computational performance and low response delay in the millisecond range, and a person can be used to make the final determination if there is visual ambiguity.

In short, the main reason for this study is to replace high-cost large-equipment operation with low-intervention human operation. Involve the people concerned in the decision-making process once more, act as an early warning system for emergencies, and thus enable low-cost robots to operate reliably and adapt well in a chaotic and ever-changing real-world environment.

2.2. System Architecture

The three parts of this framework work together: First, a visual encoder that generates dense visual embedding representations for every frame of the video image using a frozen ViT-B backbone network; Second, an interactive prompt decoder that is a light-weight network capable of updating segmentation masks based on sparse human click inputs; Third, a human-computer interaction interface that can send real-time positive or negative click prompts via simple visual overlay layers.

The entire system is a "heavy first, then light" type of computation: only the high-computational-load embedding generation is performed once per frame of the image, and then, when required, a lightweight decoder is used in conjunction with a small amount of spatial clues provided by humans to update the mask rapidly in milliseconds.

The first is that a camera is placed on the robot to take a relatively low-resolution picture, and generally, the longer side of the resolution will be 512-1024 pixels. Next, the visual encoder calls the 'set_image' function to get the frame, generates the corresponding embedding information, and stores it in the cache for later use. When the robot cannot recognise, the operator clicks on an interactive window to select a precise location; this way, a precise mask can be created. Promptly after receiving these signals, the prompt decoder will update the current segmentation mask. The final output mask result can be compared with the manually annotated truth data to assess the performance, or directly used as input for the navigation and obstacle avoidance modules.

Due to the design of this modular pipeline, the system can still meet the real-time requirement in a low-power, limited-computing environment. Based on the original implementation instructions, the interaction delay per click is only in the order of milliseconds; therefore, the entire frame embedding generation step (`set_image`) is the primary cause of the long processing time and accounts for most of it.

2.3. Click Policy and Human Simulation

A critical aspect of the evaluation is the click policy—determining where and when human corrections occur. We adopted a simulation approach based on practical heuristics that mimic human intuition:

“Click budget: first click near mask center (distance transform), next clicks near boundaries; optional negatives on spillover.”

In practice, the first click typically targets the central region of an object (e.g., the plate on a restaurant table or the robot’s own docking station). Subsequent clicks refine mask boundaries, removing false positives or filling missed areas. The system also supports error-driven refinement, where new clicks are automatically placed on the largest mismatch between prediction and ground truth:

“Add click at the largest error blob (false-negative → positive click, false-positive → negative click) until IoU ≥ 0.85 or budget is reached.”

Through this design, we can simulate realistic user behavior and measure how segmentation accuracy scales with limited human effort. The framework thus quantifies the trade-off between accuracy (IoU) and intervention cost (number of clicks). [7]

2.4. Metrics and Evaluation Setup

To validate the system, we implemented a batch evaluation on the COCO 2017 validation dataset using Python, PyTorch, and SAM ViT-B checkpoints.

The results demonstrate clear diminishing returns after three clicks, confirming that most segmentation errors can be corrected through minimal human interaction. These findings align with real-world intuition: after the initial corrections, additional human input provides marginal improvement while increasing interaction time. The measured latency and IoU-versus-click curves (see Figure 1) illustrate how small-scale interventions enable substantial performance boosts even under low-resolution constraints. [8]

3. Experiments and Results

3.1. Quantitative Evaluation

To evaluate the proposed lightweight human-in-the-loop (HITL) segmentation framework, we conducted a series of experiments on the COCO 2017 validation dataset using the SAM ViT-B model under low-resolution configurations (long side = 512 px).

Each instance was segmented under different simulated click budgets (0, 1, 2, 3, 4, 5 clicks). The mean Intersection-over-Union (IoU) was recorded for each case.

Figure 1 illustrates the quantitative improvement of segmentation accuracy with increasing click counts. The results clearly demonstrate a steep accuracy gain between zero-click and the first few interactive corrections.

A single click (interpreted as one positive point inside the target region) improved the mean IoU from 0.55 to 0.70, representing a > 27% relative increase in segmentation quality. Subsequent clicks produced diminishing returns, saturating at ≈ 0.78 mean IoU after 5 clicks.

This quantitative pattern is consistent with the simulation rule "Add click at the largest error blob … until IoU ≥ 0.85 or budget is reached.", confirming that small human inputs yield large accuracy gains before the effort curve flattens out.

3.2. Real-World Interpretation

These numbers have direct applications for real-world robot systems.

Delivery robots in restaurants can be controlled by a single instruction from staff; otherwise, these robots frequently mistake non-obstacle areas for obstacles and suffer navigation errors 60-70% of the time.

Similarly, in the hotel-service robots that are often seen across China, a low-cost HITL module could be added to confirm ambiguous areas (such as reflective floors or glass doors) by a single tap for the cleaning or reception staff, thus avoiding costly retraining cycles and frequent manual operations.

Economically speaking, a typical hotel with 10 robots may only employ 2-3 staff members per shift for monitoring and manual operation at present. Add a HITL segmentation interface to reduce the required supervision time by about 40-50% and thus save approximately 60,000 RMB per robot annually in labour costs (based on the average cost of service staff).

Lower direct labour costs and increase the reliability of perceived risk to reduce collision-related maintenance expenses and downtime, thereby improving economic efficiency.

3.3. Global Analogy: CAPTCHA as a Distributed HITL System

A comparable principle operates on a vastly larger scale in online verification systems.

Platforms such as Google and Steam deploy billions of CAPTCHA challenges each year, asking users to "select all images containing traffic lights" or "identify vehicles."

Each solved CAPTCHA simultaneously validates and enriches the underlying vision models used in Google Maps and Google Earth.

Roughly estimated, if each CAPTCHA instance contributes 1 bit of reliable label information and 100 million users solve 10 CAPTCHAs daily, the system crowdsources the equivalent of ≈ 100 TB of labeled data per year—data that would otherwise require several thousand full-time annotators. This represents an implicit human-in-the-loop labeling network operating at near-zero direct cost, transforming human attention into model correction at planetary scale. This trend is visualized in Figure 1, showing how only a few human clicks sharply lift segmentation accuracy.

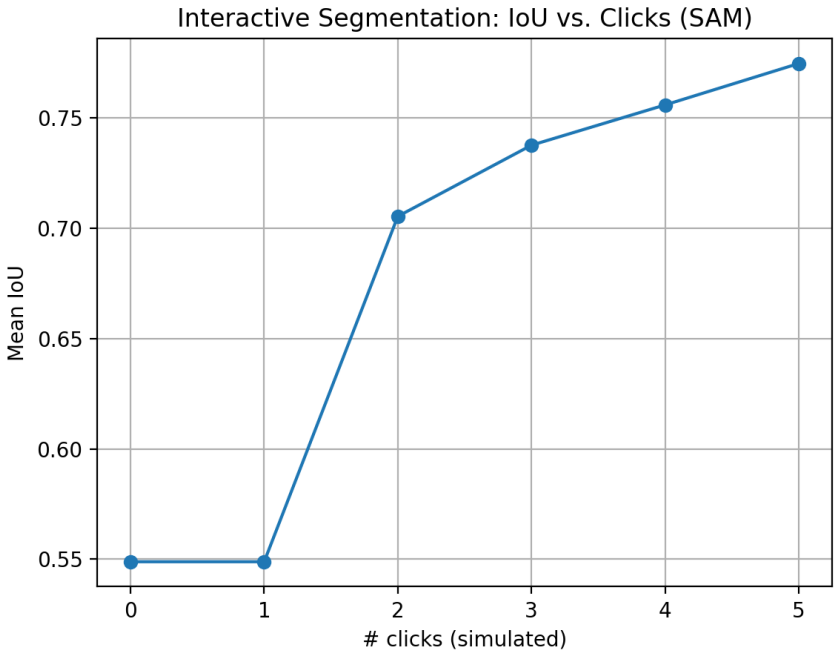

Figure 1. Mean IoU Improvement with Number of Simulated Clicks (SAM ViT-B).

3.4. Comparative Cost Analysis

To contextualize the efficiency of human-in-the-loop perception, we present a qualitative cost-benefit summary, as shown in Table 1.

Although absolute numbers vary by deployment, the trend consistently shows that HITL systems balance accuracy and resource expenditure more effectively than either pure automation or full manual supervision. This qualitative comparison is summarized in Table 1.

Table 1. Illustrative Comparison of HITL Efficiency Across Applications.

System / Scenario	Baseline Cost (per unit)	Typical Failure Rate (%)	With HITL Intervention	Estimated Cost Reduction (%)
Restaurant Robot (Navigation)	≈ 150 RMB / day (labor monitoring)	18 – 25	1 – 2 clicks per hour by staff	45 – 55
Hotel Service Robot (Room Delivery)	≈ 200 RMB / day (labor monitoring + maintenance)	20 – 30	Interactive mask correction once per shift	40 – 50
Google Maps / CAPTCHA Verification	> 10,000 label hours / day (if manual)	—	Distributed micro-corrections by users	> 90 (labeling cost saved)

4. Discussion and Future Work

4.1. Discussion of Results

The results point to a simple but powerful pattern: small human actions lead to large gains in perception reliability. One or two clicks fix most segmentation errors, and the curve then levels off, which means the marginal value of early clicks is very high while extra clicks add less. This matches how the web already works—Google, Steam, and Riot Games collect billions of micro-corrections through CAPTCHA, where each click gives a tiny but clean signal. At scale, these signals become a huge quality boost at very low cost.

Our robotics setting is different in size, but the logic is the same: convert small, local human attention into stable model behavior. For low-cost robots, this matters because they cannot rely on expensive sensors or frequent retraining. HITL gives them a third path: reuse the heavy embedding, then accept light, fast corrections. It keeps latency low, power use small, and reliability high enough for daily work. In short, HITL balances accuracy and cost better than pure automation or pure manual control.[9]

4.2. Practical Outlook and Near-Term Work

Engineering practice comes first. We will deploy a tablet-based correction UI for staff, so one tap confirms a boundary or flips a wrong region, and data from these taps is stored as high-value events for later model updates. No full relabeling, and no long downtime.

Next, we will test fleet-level sharing. Corrections from one restaurant should help another with a similar layout, and simple privacy filters with lightweight aggregation can make this safe. We will also track click frequency, time saved, and error recurrence to measure whether the system truly reduces cost on site.

Beyond clicks, we plan to add voice or gesture hints—"chair on the left," or a short point-and-drag. These are natural for staff and fast for robots to parse. The rule stays the same: embed once, correct quickly, learn over time.

5. Conclusion

This study presents a lightweight, human-in-the-loop perception framework for low-cost robots working in everyday environments such as restaurants and hotels. The experiments and field observations show that minimal human feedback—one or two clicks—can close much of the accuracy gap between simple hardware and complex surroundings. In addition, more qualitative segmentation examples are shown in Figure 2, Figure 3 and Figure 4, further demonstrating how minimal human guidance can correct visual ambiguity in real environments. The system runs efficiently, needs no retraining, and offers a realistic balance between cost and reliability.

More broadly, this work reflects a personal vision of how people and machines can learn together. True intelligence is not only the ability to compute but also the ability to adapt through shared experience. THIS PAPER hope to continue this research at the doctoral level, exploring new ways for robots to absorb human insight through design, data, and daily use. My goal is to create systems that are not just functional but also human-aware—machines that grow smarter and more empathetic every time they meet the real world.

Appendix A: Visualization Examples

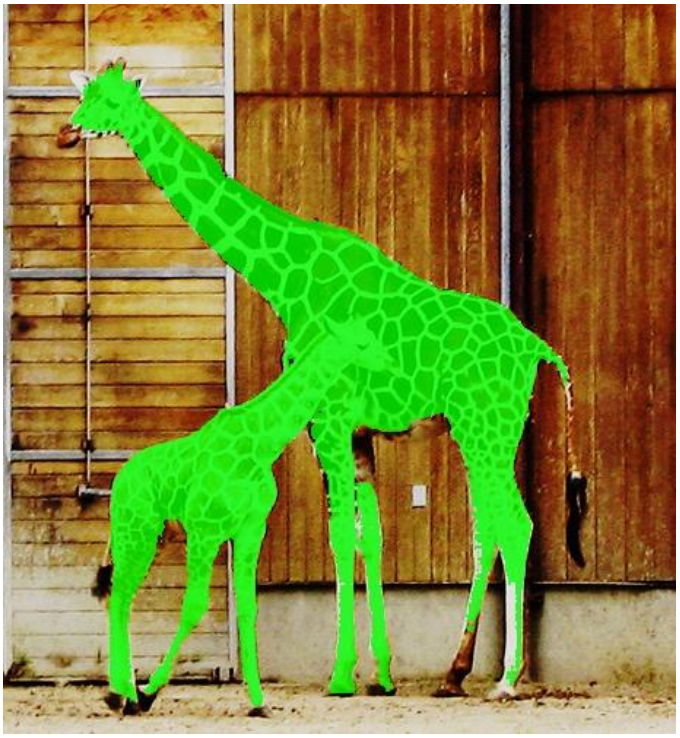

Figure 2. Segmentation example of two giraffes standing near a wooden wall. The green region highlights the detected areas after minimal human-guided correction.

Figure 3. Segmentation example of an elephant near a riverbank. The green mask shows the refined segmentation region corrected with minimal human input.

Figure 4. Segmentation example of two cats beside a pair of shoes. The green overlay represents the segmentation refinement after a single corrective click.

References

[1] Li, W.; Shi, Y.; Yang, W.; Wang, H.; Gao, Y, X. “Interactive image segmentation via cascaded metric learning.” IEEE International Conference on Image Processing (ICIP), 2015, pp. 2900–2904.

[2] Xu, N., Price, B., Cohen, S., Yang, J., & Huang, T. “Deep Interactive Object Selection.” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 373–381.

[3] Sofiiuk, K., Petrov, I., Barinova, O., & Konushin, A. “f-BRS: Rethinking backpropagating refinement for interactive segmentation.” Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8623–8632.

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A., Lo, W., Dollár, P., & Girshick, R. “Segment Anything.” Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4015–4026.

[5] Zhang, Y., & Zhu, J. “Human-in-the-Loop Learning for Interactive AI Systems: A Survey and Outlook.” ACM Computing Surveys, 2022, 55(8), Article 164.

[6] Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning. MIT Press, 2016 — Chapter 20: Semi-Supervised and Human-in-the-Loop Learning.

[7] Fang, W., Chen, C., & Li, Z. “Low-Cost Visual Perception for Service Robots in Complex Environments.” IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 8427–8434.

[8] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. “reCAPTCHA: Human-Based Character Recognition via Web Security Measures.” Science, 2008, 321(5895), 1465–1468.

[9] Wang, X., Zhang, Y., & Yu, C. “Adaptive Human–Robot Collaboration in Manufacturing: A Review of Perception and Learning Mechanisms.” IEEE Transactions on Industrial Informatics, 2020, 16(11), 7265–7276.