Robot–Elevator Interaction: From Physical Panel Recognition to Activation

Siheng Ye, Zejing Tong, Suiyi Zhang, and David Chieng

Next Generation Internet of Everything Laboratory (NGIOE)
Department of Electrical and Electronic Engineering,
University of Nottingham Ningbo China
{biysy32,ssyzt6,ssysz9,David.Chieng}@nottingham.edu.cn

Abstract. Existing robot-accessible elevator interfaces require manufacturer-specific retrofits which are costly and difficult to scale. Direct physical interaction with existing elevator panels offers a promising infrastructure-independent alternative, but requires robots to reliably perceive, interpret, and actuate controls despite variations in panel layouts, lighting conditions, reflections, and labeling conventions. We present an edge-based prototype system with an end-to-end perception-to-action pipeline that integrates task-specific button detection, specialized target recognition, vision-language semantic understanding, confidence-aware rejection, semantic-state maintenance, asynchronous processing, and physical button actuation. We evaluate perception performance, static-image target selection, runtime behaviour, and physical actuation through two qualitative button-activation demonstrations. Across 475 target-selection requests derived from 112 static RGB images spanning six buildings, the system selected the intended button in 404 cases, achieving an accuracy of 85.1%. On a separate benchmark of 379 human-annotated Open/Close/Other button crops, the HOG+SVM classifier achieved a macro-F1 score of 0.9963. During a 207-s runtime evaluation, the system filtered 99.8% of redundant candidate-level semantic requests, with no observed label switching, flicker, crashes, or camera interruptions. In both operator-initiated physical button-activation demonstrations, the robotic arm activated the selected button, withdrew, and returned to its standby state. These results provide practical insights and design considerations for the development of robust, edge-based, offline, and real-time robot–elevator interaction systems.

Keywords: Service robots · Elevator interaction · Semantic perception · Stable semantic targets · Edge robotics

1 Introduction

There is increasing demand for service robots to operate autonomously across multiple floors in hospitals, offices, hotels, and other large indoor environments. However, most existing elevator systems do not expose standardized interfaces that can be readily accessed by robots. Providing such connectivity typically

requires manufacturer-dependent retrofitting or additional interface hardware, which may be costly and difficult to scale. These limitations motivate infrastructure-independent approaches in which robots perceive, interpret, and physically operate existing elevator panels in a manner similar to human users.

Such interaction requires robots to reliably perceive, interpret, and physically operate elevator panels under non-standardized panel layouts, reflective surfaces, variable lighting conditions, and diverse labeling conventions, including skipped floor numbers, bilingual text, symbols, and building-specific abbreviations. Perception must both localize controls and identify their functions [5, 7] but location alone cannot distinguish floors, door controls, alarms, and other functions. Viewpoint and image-quality changes further require unreliable semantic predictions to be rejected [7, 11].

On edge hardware, the detector, specialized classifier, and vision-language model differ in computational cost and update rate, which can cause intermittent predictions, semantic instability, and latency. These issues can be mitigated by maintaining validated button–label associations, rejecting uncertain updates, and revising semantic states only when new evidence or target reassociation justifies a change.

In this paper, we present an edge-based prototype system with an end-to-end perception-to-action pipeline that integrates task-specific button detection, specialized target recognition, vision-language semantic understanding, confidence-aware rejection, semantic-state maintenance, asynchronous processing, and physical button actuation. The main contributions of this paper are as follows.

1. **Edge-based system integration:** a modular pipeline that integrates elevator button localization, complementary specialized/general semantics generation, and operator-requested actuation on existing elevator panels.
2. **Mixed-rate semantic target preservation:** We use HOG+SVM for specialized door-control recognition and Qwen3-VL for broader semantics, with confidence-aware rejection and validated-state reuse to preserve stable semantic targets during asynchronous perception updates.
3. **System evaluation:** We conducted a staged evaluation covering module-level detection and semantic recognition, static-image target selection across six buildings, runtime behavior, and two qualitative physical button-activation demonstrations.

2 Related Work

2.1 Elevator Button Detection and Recognition

Elevator-button recognition is challenging because physical panels vary in appearance and are affected by reflections, viewpoint changes, and occlusion. Early work therefore combined local visual evidence with panel-level structure. Kim et al. [6] used a pre-learned panel model, homography-based matching, and button alignment to improve recognition under specular reflections and partial occlusion. This work showed the value of geometric context when local appearance alone is unreliable, although it required prior knowledge of the target panel.

Later studies adopted learning-based methods for broader panel variation. Dong et al. [4] combined contour-derived button candidates with convolutional neural network(CNN) classification and panel-order correction, and explicitly evaluated recognition on four previously untrained elevators. Zhu et al. [11] proposed OCR-RCNN to jointly detect buttons and recognize character sequences on a substantially larger dataset containing more than 21,000 button labels and hundreds of label classes. Together, these studies show a progression from panel-specific geometric reasoning toward learned, cross-elevator, and larger-vocabulary recognition.

2.2 Robotic Button Pressing and Integrated Elevator Operation

Elevator perception has also been connected directly to physical robot operation. Klingbeil et al. [7] combined button detection, grid reasoning, OCR, and HMM-based consistency with three-dimensional targeting and robotic pressing. Their mobile robot was evaluated on three previously unseen elevators, locating the requested target in all 14 trials and successfully pressing it in 13. This work established an early example of connecting unfamiliar-panel perception directly to physical action.

Li et al. [9] later proposed the AEBRO framework, which combines numeral recognition, spatial and numerical consistency constraints, RGB-D-based target localization, and Kinova manipulation. Four of five reported button-operation trials were successful. More recently, Lee et al. [8] integrated YOLO-based button recognition, position estimation, myCobot manipulation, and mobile-robot navigation into a complete elevator-use sequence. They compared YOLOv4 and YOLOv4-Tiny and selected the latter for its substantially higher frame rate at similar reported recognition performance. Across eight physical trials, all hall-button presses and seven of eight floor-button presses succeeded.

Prior work has demonstrated important advances in elevator-panel recognition and robotic button actuation. Accordingly, our work further contributes in integrating complementary specialized and general-purpose semantic recognition, maintaining validated semantic target states under mixed-rate edge inference, and enabling physical actuation of operator-specified elevator buttons. Specifically, HOG+SVM provides specialized Open/Close/Other recognition, while Qwen3-VL supports broader semantic interpretation. The modular architecture further enables independent evaluation of each component and its contribution to the end-to-end perception-to-action pipeline.

3 System and Methods

3.1 System Hardware

Figure 1 shows the experimental platform, which integrates visual sensing, onboard computation, and a robotic manipulator on a three-tier utility cart. An Orbbec Gemini 335 RGB-D camera provides panel-level color observations and

depth measurements for three-dimensional button localization. A wrist-mounted TYS-CLB30W-V1 RGB camera provides a close-range image for one-time target-position refinement before pressing. A Waveshare RoArm-M3 manipulator carries a Button Actuation End-effector (BAE) to press the selected button.

An NVIDIA Jetson AGX Orin 64GB Developer Kit serves as the onboard computing unit. A laptop provides remote desktop monitoring through NoMachine and terminal access through SSH. For panel-interaction experiments, the cart is manually positioned and kept stationary during arm operation; autonomous navigation is outside the scope of this work.



Fig. 1. Experimental robot platform and major hardware components.

3.2 End-to-End Perception-to-Action Pipeline

Figure 2 summarizes the perception-to-action pipeline. Localized button candidates and semantic observations form a stable semantic target. After the user selects a localized, labeled candidate, a single PRESS request passes that target identity and fresh geometry to the physical interaction module (Section 3.3).

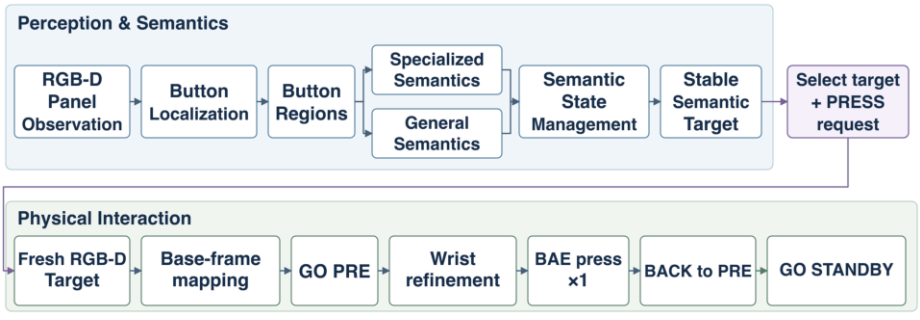


Fig. 2. End-to-end perception-to-action pipeline.

Panel Perception and Semantic Understanding Button candidates are cropped for two complementary recognition branches. Specialized door-control recognition handles a narrow vocabulary, while the general branch provides broader semantics. Both pass observations to semantic-state management for validation and target generation.

Stable Semantic Target Generation Mixed-rate observations are associated with button candidates and validated before maintained state is updated. Uncertain predictions do not overwrite a validated label, which remains available until new evidence or target reassociation triggers re-evaluation. The resulting stable semantic target pairs the current button instance with its validated operational label.

3.3 Physical Interaction Execution

Selection and admission. RECOGNIZE shows localized, labeled button boxes in the Gemini RGB view; the user selects one, which is highlighted. Neither step moves the arm. A single PRESS request locks the target, obtains fresh RGB-D geometry, transforms it to the arm-base frame, and records the stationary joint pose. Before motion, the system checks readiness, inverse-kinematics feasibility, joint limits, step size, and proxy clearances. Execution may begin from any suitable measured pose; GO STANDBY is not required first. These admission checks assess motion feasibility, while semantic labeling is completed by the perception pipeline.

Approach and press. Within the same PRESS request, internal GO PRE moves the BAE tip to a nominal 70 mm non-contact standoff. After the arm stops, one wrist-camera image refines the position. The target identity and geometry are then frozen for one position-controlled press and BACK withdrawal to PRE; no second confirmation or target switch occurs.

Return and limits. After BACK and a new readiness check, the controller issues internal GO STANDBY; Standby is accepted only on controller completion. STOP requests a motion stop and invalidates the active flow so late callbacks cannot resume it. Execution faults stop the sequence without automatic retry or return.

4 Implementation and Experimental Setup

4.1 Perception and Semantic Implementation

The deployed button-localization block uses a single-class YOLO11n model trained to detect elevator buttons. Input images are processed at 640×640 pixels with a confidence threshold of 0.40, followed by non-maximum suppression with an intersection-over-union (IoU) threshold of 0.70. Detected boxes are expanded by 15% before semantic recognition. Aligned depth and camera intrinsics are then used to estimate the corresponding 3-D button position in the camera frame.

The specialized semantic branch combines Histogram-of-Oriented-Gradients (HOG) features with a linear Support Vector Machine (SVM) classifier [3, 2]; this configuration is referred to as HOG+SVM throughout the paper. Crops are resized to 160×160 pixels, converted to grayscale, and processed using contrast-limited adaptive histogram equalization before feature extraction. The classifier uses $C = 0.1$, balanced class weights, and probability estimates. The broader semantic branch uses Qwen3-VL-2B-Instruct as a vision-language model for floor, hall-direction, and general-button semantics [1].

Specialized predictions are mapped to `UNKNOWN` unless the maximum class probability is at least 0.60 and the margin between the two highest probabilities is at least 0.05. A door-control label requires agreement in at least two of the three most recent valid observations, whereas the broader semantic branch requires two consecutive identical valid outputs.

State revalidation is triggered when the associated-box IoU falls below 0.35 or the crop difference-hash Hamming distance exceeds 18 differing bits. Repeated semantic requests for unchanged instances are filtered before inference to avoid redundant processing, association loss invalidates the maintained state, and asynchronous semantic processing retains the latest validated state across processing gaps. Figure 3 illustrates the resulting processing sequence on a representative cabin panel, from RGB-D observation and button localization to complementary semantic recognition and validated button-level output.

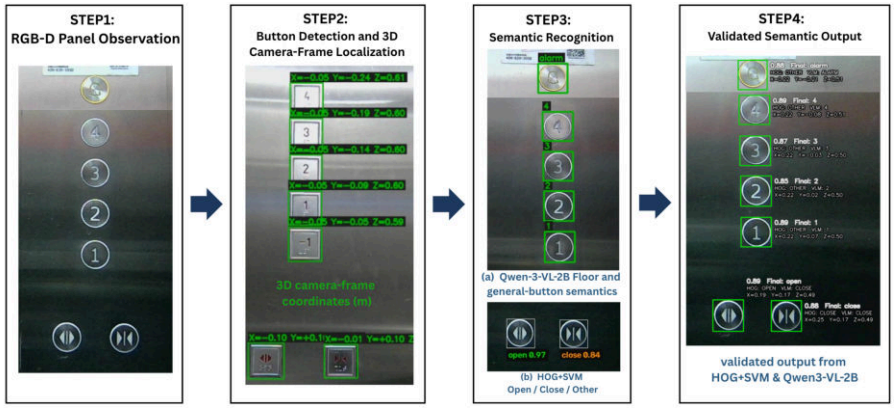


Fig. 3. Perception and semantic processing pipeline.

4.2 Physical Interaction Implementation

Coordinate calibration. Gemini button points and panel normals are transformed from the camera optical frame to the arm-base frame using the fixed camera-to-base calibration. Wrist-camera intrinsics and distortion coefficients are used

to correct lens distortion. Park–Martin hand–eye calibration [10] estimates the camera's fixed pose relative to the robot tool frame. The BAE contact point is represented by a fixed positional offset of [111.789, 21.475, 1.302] mm in the tool frame.

Press geometry. Gemini depth provides a button point and a panel normal fitted from nearby depth samples. PRE is set at a nominal 70 mm outward standoff. After the arm stops at PRE, one distortion-corrected wrist-camera ray is mapped through the camera-to-tool calibration and the current tool pose into the arm-base frame. It is then intersected with the frozen panel plane to update the target position once. The tangent-plane correction is limited to 35 mm, and its upward component is disabled. Commanded overtravel is the extra tip displacement beyond the estimated button plane along the inward normal. Both demonstrations used a fixed 4 mm commanded overtravel under position control, with a single wrist-camera refinement at PRE.

Calibration checks. Table 1 summarizes calibration-consistency residuals for the camera-to-base and camera-to-tool transforms. Wrist-camera intrinsic fitting yields a 0.324-pixel root-mean-square(RMS) reprojection error.

Table 1. Calibration-consistency residuals for the camera-to-base and camera-to-tool transforms.

Calibration	Residual (unit)	Median	Max
Gemini camera-to-base	Translation (mm)	6.920	8.609
22 fitting / 8 held-out poses	Rotation (deg)	1.274	2.081
Wrist camera-to-tool frame	Translation (mm)	7.181	10.512
20 calibration / 5 diagnostic poses	Rotation (deg)	2.060	3.764

The reported residuals quantify geometric transform consistency; physical BAE activation is demonstrated separately in Figure 4.

4.3 Experimental Protocols

Perception and Recognition Evaluation Table 2 summarizes the module-level datasets and roles. The task-specific YOLO11n detector, trained for elevator buttons, was compared with the text-conditioned open-vocabulary YOLO-World and Grounding DINO under a shared button-detection protocol. Geometry is a non-learning geometric baseline; HOG+SVM specializes in Open/Close/Other recognition; Qwen3-VL-2B-Instruct (abbreviated as Qwen3-VL-2B in tables) is the vision-language model used for broader semantic interpretation.

Detection was evaluated using precision, recall, F1, and AP_{50}, with predicted and ground-truth boxes matched at intersection-over-union (IoU) ≥ 0.50. The shared 379-crop Open/Close/Other benchmark reports per-class F1, Macro F1,

Table 2. Module-level datasets and evaluation roles.

Protocol	Data	Role
YOLO11n development	278/82 images; 1,187/469 boxes	Train / validation
Detector comparison	Shared detector benchmark	Three-model comparison
HOG+SVM	481 crops	Specialized training
Open/Close/Other	379 crops	Semantic comparison
Qwen3-VL-2B	494/483 crops	Broader-semantic diagnostic

Balanced Accuracy, and the `UNKNOWN` rate, the fraction of samples that are rejected rather than assigned a class. The separate broader-semantic Qwen3-VL-2B-Instruct diagnostic reports overall accuracy and, for human-standard crops, floor-label and hall-direction accuracies.

System-Level Target Selection and Runtime Evaluation The retrospective evaluation used 112 static RGB images with 544 human-reviewed ground-truth annotations, yielding 475 image–target requests after unknown annotations were excluded. Same-label boxes within an image were retained as acceptable alternatives, and panel mode was supplied from annotation metadata. A request was counted as correct when the system returned a unique requested-label target overlapping an acceptable ground-truth box at IoU ≥ 0.50.

Live runtime tests were conducted on the deployed NVIDIA Jetson AGX Orin under Ubuntu 22.04 and Robot Operating System 2 Humble using 1920×1080 camera input from the Orbbec Gemini 335. The protocol comprised three viewer trials, three semantic-state stability trials, six voting trials, and a 207-s integrated run, measuring frames per second, image-to-display latency, redundant semantic-request filtering, locked-state coverage, and stability events. The 95th percentile is the delay at or below which 95% of measurements fall. The redundant semantic-request filtering rate is the fraction of candidate-level semantic requests skipped as unchanged repeats before inference; locked-state coverage is the fraction of sampled runtime observations carrying a retained validated label; and stability events are observed label-switch or display-flicker incidents. Voting used `window`/`majority`=1/1 for the OFF condition and 3/2 for the ON condition under otherwise unchanged settings.

Physical Button Activation Evaluation Two demonstrations were conducted with the cart stationary: one with a hall down-call button and the other with a cabin floor-2 button. Both used the 4 mm commanded-overtravel setting in Section 4.2. The execution sequence comprised approach, pressing, withdrawal, and return to Standby. Video review assessed off-to-on illumination during BAE contact and documented the subsequent withdrawal. These qualitative cases evaluate physical actuation after the operator selects a system-localized, semantically identified button target; they do not measure semantic-recognition accuracy or repeated-trial reliability. Illumination was not used as a controller success or retry gate.

5 Results

5.1 Panel Perception Performance

Button Detection Under the common button-detection protocol, YOLO11n achieves the strongest precision, recall, F1, and AP_{50}. Grounding DINO retains high recall but lower precision, while YOLO-World performs worse overall.

Table 3. Button-detection model comparison.

Detector	Precision	Recall	F1	AP_{50}
YOLO11n	**0.9889**	**0.9763**	**0.9826**	**0.9876**
YOLO-World	0.5366	0.7226	0.6159	0.5720
Grounding DINO	0.5531	0.9690	0.7042	0.6104

Door-Control Recognition On 379 human-annotated Open/Close/Other crops, HOG+SVM achieves the highest Macro F1 and balanced accuracy with the lowest rejection rate.

Table 4. Open/Close/Other recognition benchmark on 379 human-annotated crops.

Method	Open F1	Close F1	Other F1	Macro F1	Bal. Acc.	Unknown
Geometry	0.201	0.123	0.169	0.164	0.144	65.4%
Qwen3-VL-2B	0.345	0.090	0.915	0.450	0.411	23.0%
HOG+SVM	**1.000**	**0.991**	**0.998**	**0.996**	**0.993**	**0.5%**

General Semantic Recognition In a separate broader-semantic diagnostic, Qwen3-VL-2B-Instruct achieves 366/494 (74.09%) accuracy on actionable human-standard crops and 353/483 (73.08%) on detector-matched crops. On human-standard crops, floor-label and hall-direction accuracies are 93.08% and 92.22%, respectively.

Offline Visual Target-Selection Performance Across 112 static RGB images containing 544 human-reviewed ground-truth boxes, the system correctly selects a unique target for 404 of 475 requests, yielding an offline visual target-selection accuracy of 85.1% (Table 5). Performance is better for cabin than hall requests, while Open and Close requests achieve the strongest subgroup results.

Among the 71 non-correct outcomes, 40 were rejected or absent (8.4% of all requests) and 15 were ambiguous (3.2%). The remaining outcomes comprised 12 semantic mismatches (2.5%) and four upstream or closed-set cases (0.8%). The available aggregate results do not distinguish incorrect actionable selections from rejected or unavailable targets.

Table 5. Offline visual target-selection results. Counts are image–target requests.

Subset	Correct / requests	Accuracy
Overall	404 / 475	85.1%
Cabin	332 / 386	86.0%
Hall	72 / 89	80.9%
Open	54 / 55	98.2%
Close	51 / 55	92.7%

Runtime and Semantic-State Stability Across three viewer trials, the system averaged 4.84 frames per second, with a mean 95th-percentile image-to-display delay of 41.5 ms. Across three stability trials, the redundant semantic-request filtering rate averages 97.7% and locked-state coverage averages 99.3%, with no observed label switches, display-flicker events, or crashes.

The 207-s integrated run records 3,540 candidate-level semantic requests and filters 3,532 unchanged repeats before inference (99.8%), with no observed label switches, display-flicker events, crashes, or camera interruptions.

Voting OFF and ON both produce no observed switches or display-flicker events and 99.7% redundant semantic-request filtering. With caching and state reuse active in both conditions, no additional measurable independent benefit of temporal voting is observed under the stable-panel protocol.

5.2 Physical Button Activation

Figure 4 shows button illumination during BAE contact and subsequent withdrawal in both cases; the buttons remain lit in the withdrawal frames. The arm then returns to Standby. These qualitative cases demonstrate activation of system-localized, semantically identified button targets selected by the operator, but do not establish a success rate or cross-elevator robustness.

6 Discussion

6.1 Complementary Roles of Specialized and General Semantic Recognition

Both Geometry and Qwen3-VL-2B-Instruct struggle with the evaluated door-control icons, motivating HOG+SVM for Open/Close/Other and Qwen3-VL-2B-Instruct for broader semantics. The result supports a system-integration contribution, but these module-level comparisons do not isolate end-to-end component effects.

Because detection, broader semantics, rejection, ambiguity, and target availability all contribute, the 85.1% result characterizes complete request-level target selection rather than isolated semantic recognition. Further improvement therefore depends on reducing non-actionable outcomes across these stages.

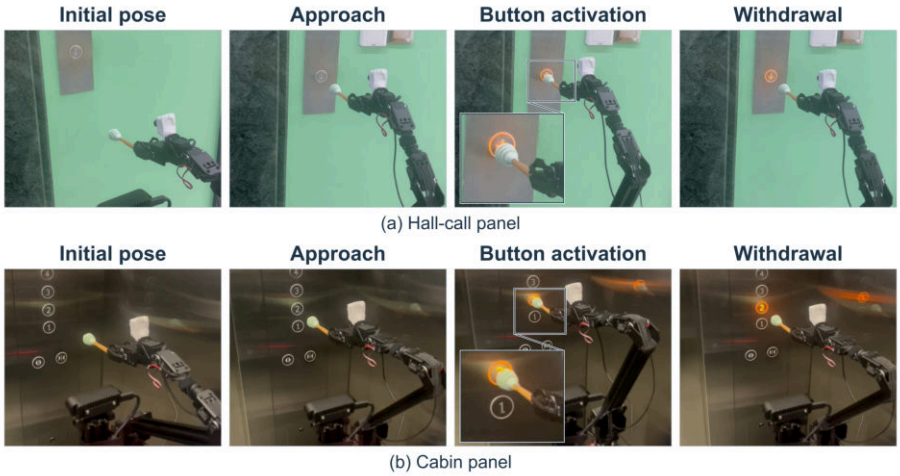


Fig. 4. Button-activation sequences on (a) hall-call and (b) cabin panels.

6.2 Stable Semantic Targets under Mixed-Rate Perception

Validated button–label associations, rather than individual predictions, support interaction readiness. Rejection and revalidation prevent uncertain observations from overwriting maintained semantics, while asynchronous reuse keeps stable semantic targets available across mixed-rate updates.

Runtime stability therefore measures availability, not correctness, and fail-closed rejection can reduce availability. Because the stable-panel ablation did not isolate an independent benefit of three-frame, two-vote voting, it remains a stabilization mechanism rather than a separately validated contribution.

6.3 Limitations and Future Work

Floor metadata and condition-specific accuracy are unavailable, so floor-wise and cross-condition generalization remain unquantified. Target selection used metadata-supplied mode and correlated requests; the Qwen3-VL diagnostic was crop-level. Voting covered stable panels only, without disturbance revalidation. The two physical demonstrations assess physical actuation of system-localized, semantically identified button targets selected by the operator, not semantic-recognition accuracy or repeated-trial reliability.

Future work should evaluate repeated end-to-end interaction cycles across multiple elevators and buildings, including dynamic live camera-to-target operation. Synchronized video and controller feedback should quantify correct-target activation, repeatability, and recovery under viewpoint, lighting, and panel changes. Short-window illumination tracking should also be investigated for online outcome verification and guarded recovery.

7 Conclusion

This work presents an edge-based system that connects stable semantic targets with guarded robotic button pressing. In offline tests, the system selects the requested target with 85.1% accuracy, while a deployed runtime test filters 99.8% of unchanged semantic requests without observed label switches or crashes. In two physical demonstrations, the robot activates selected hall-call and cabin buttons, withdraws, and returns to Standby. These staged results show the prototype's current capability, but repeated live end-to-end testing across different elevators is still needed.

References

1. Bai, S., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 886–893 (2005). https://doi.org/10.1109/CVPR.2005.177
4. Dong, Z., Zhu, D., Meng, M.Q.H.: An autonomous elevator button recognition system based on convolutional neural networks. In: 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 2533–2539 (2017). https://doi.org/10.1109/ROBIO.2017.8324801
5. Jiang, S., Yao, W., Wong, M.S., Hang, M., Hong, Z., Kim, E.J., Joo, S.H., Kuc, T.Y.: Automatic elevator button localization using a combined detecting and tracking framework for multi-story navigation. IEEE Access **8**, 1118–1134 (2020). https://doi.org/10.1109/ACCESS.2019.2958092
6. Kim, H.H., Kim, D.J., Park, K.H.: Robust elevator button recognition in the presence of partial occlusion and clutter by specular reflections. IEEE Transactions on Industrial Electronics **59**(3), 1597–1611 (2012). https://doi.org/10.1109/TIE.2011.2161249
7. Klingbeil, E., Carpenter, B., Russakovsky, O., Ng, A.Y.: Autonomous operation of novel elevators for robot navigation. In: 2010 IEEE International Conference on Robotics and Automation (ICRA). pp. 751–758 (2010). https://doi.org/10.1109/ROBOT.2010.5509466
8. Lee, J.Y., Tokuda, T., Sato, K.: Autonomous floor navigation control of mobile robot using elevator button recognition with deep learning. In: 2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE). pp. 133–138 (2023). https://doi.org/10.23919/SICE59929.2023.10354122
9. Li, S., Chen, Y., Meng, Y., Lou, Y.: Autonomous elevator button recognition and operation framework for multi-floor mobile manipulator navigation. In: 2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). pp. 383–388 (2021). https://doi.org/10.1109/AIM46487.2021.9517452
10. Park, F.C., Martin, B.J.: Robot sensor calibration: Solving AX=XB on the Euclidean group. IEEE Transactions on Robotics and Automation **10**(5), 717–721 (1994). https://doi.org/10.1109/70.326576
11. Zhu, D., Li, T., Ho, D., Zhou, T., Meng, M.Q.H.: A novel OCR-RCNN for elevator button recognition. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3626–3631 (2018). https://doi.org/10.1109/IROS.2018.8594071