ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robust Unknown Object Detection and Tracking for Vision-Language-Action Models on Edge Devices*

Subin Joo, and Deokgi Jeung

***Abstract*—This study proposes a Stepwise Vision-Language-Action (VLA) framework for the robust detection and tracking of unknown objects in edge device environments (NVIDIA Jetson AGX Orin). Conventional end-to-end VLA models face challenges such as massive memory requirements and a "black-box" nature that complicates debugging. To address these issues, we adopt a modular architecture, specifically integrating Depth-Guided Gaussian Sampling with MobileSAM in the vision module. This approach achieves over 99% detection success for unlearned objects. Furthermore, we demonstrate real-time 6-DOF pose tracking at over 30 FPS through ORB feature matching and ROI-based localization following the initialization phase.**

I. Introduction

With the increasing deployment of upper-body humanoids in diverse service and industrial sectors, research on VLA foundation models is accelerating. However, state-of-the-art LLM and VLA models often require dozens of gigabytes of memory, predominantly necessitating high-performance workstations or server-grade infrastructure [1, 2]. While efforts to implement these models on small-scale platforms and edge devices exist, end-to-end architectures suffer from long training times and an inherent inability to debug internal errors due to their black-box nature.

This study aims to overcome these limitations by deconstructing the end-to-end model into a lightweight Stepwise VLA model that allows for modular optimization and debugging. Specifically, to address the recurring problem of "unknown objects"—objects not included in training data or rendered unrecognizable by lighting and camera angles—we developed a vision algorithm capable of perceiving any item as an "object" and precisely tracking its motion. The proposed system targets an integrated processing speed of over 30 FPS on the NVIDIA Jetson AGX Orin.

II. Related Work

General-purpose end-to-end VLA models proposed in recent humanoid robotics research demand significant computational resources and tend to exhibit increased error in unlearned motions or specific hardware environments [2, 3]. Particularly in surgical and precision manufacturing robotics, where accuracy and real-time performance are critical, on-device execution and privacy preservation are essential.

*Research supported by the Research Program of the Korea Institute of Machinery and Materials and the Ministry of Trade, Industry and Energy (MOTIE) (No. RS-2024-00469885).

Subin Joo and Deokgi Jeung are with the Department of Robot Application, Korea Institute of Machinery and Materials, Daegu, South Korea (corresponding author to provide phone: +82-53-670-9020; fax: +82-53-670-9001; e-mail: sbjoo@kimm.re.kr).

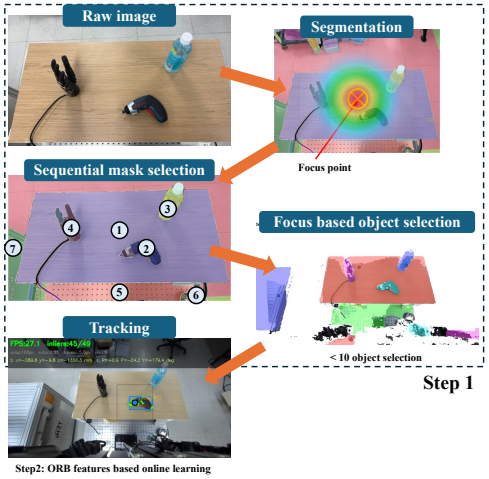

Figure 1. Robust unknown object detection framework

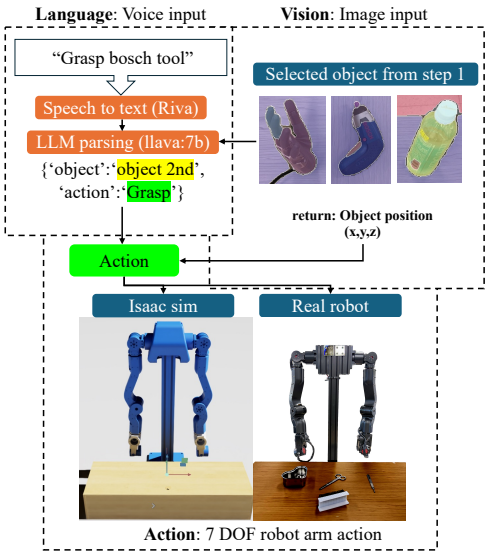

Figure 2. stepwise VLA framework

As an alternative, a prior Stepwise VLA framework proposed modularly integrating Faster Whisper (STT), Llama 3.1:8B (intent recognition), and YOLO-World (object detection). This modular approach demonstrated that granular tuning and debugging at each stage could ensure high reliability and efficiency in embedded environments.

However, detection-based models like YOLO-World still face recognition limits for unlearned objects and are sensitive to lighting conditions. Conversely, while the Segment Anything Model (SAM) excels in zero-shot segmentation for unknown objects, its computational cost is prohibitive for real-time control [4]. This study inherits the advantages of the

modular framework while proposing a hybrid vision module that combines the precision of MobileSAM with the speed of classical computer vision (ORB, PnP).

III. Methods: Stepwise Vision Module

The proposed vision module is divided into Phase 1 (Initialization) and Phase 2 (Real-time Tracking).

A. Phase 1: Depth-Guided Focus-Select Initialization

Define To recognize unknown objects, we focus on isolating "clusters separated from the background" rather than classifying "what" the object is.

- Depth Filtering: Utilizing ZED stereo camera point cloud data, background noise outside the workspace radius (d_{max} = 900mm) is physically filtered.
- Gaussian-Weighted Sampling based object selection: Attempting to segment all objects from an image using SAM results in dozens of redundant masks. We propose a method to select a limited number of objects using a 2D Gaussian probability distribution centered in the image, mimicking human visual fixation.
- Sequential MobileSAM Inference: Based on the Gaussian distribution, an initial pixel point 1 is selected and fed into the MobileSAM model as a point prompt to isolate object 1 and mask 1. For the remaining area (excluding mask 1), a second pixel point is selected using the same distribution. Repeating this process n times allows for the selection of n objects focused around the center.

B. Phase 2: ORB-Based 6-DOF Tracking with ROI Localization

For real-time tracking of recognized objects, a feature-based algorithm is implemented.

- ROI Localization: Computational cost is drastically reduced by cropping a localized region of interest (ROI) centered on the object's position in the previous frame.
- ORB Feature Matching: ORB features extracted within the mask are matched with the current frame's ROI, and geometric consistency is verified via RANSAC.
- 6-DOF Pose Estimation: The Perspective-n-Point (PnP) algorithm calculates the object's 3D position (x, y, z) and orientation (roll, pitch, yaw), which are then transmitted to the robot's action module

C. Stepwise VLA full concept

The full concept of the stepwise VLA equipped with the proposed vision module is shown in Figure 2. Multiple unclassified objects selected through Phase 1 are fed into a vision-capable lightweight LLM model (llava:7b) along with a text prompt. The LLM identifies the image most similar to the target "Bosch tool" among the inputs, returns its index, and the robot subsequently executes the corresponding action.

IV. Experimental Results

The performance was validated using an NVIDIA Jetson AGX Orin and a Stereolabs ZED camera. We specifically tested the operational feasibility and speed of the Vision and Language parts, excluding the Action module.

- Recognition Success Rate: Experimental results showed a success rate of over 99% for various unknown objects on a desk. This robust performance leveraged both color and depth disparity, succeeding in low-contrast situations where YOLO and SSD-based detection algorithms typically fail or lose track. While the proposed algorithm requires approximately 1.3 seconds for mask acquisition during initialization, it provides nearly 99% reliability in object classification from images
- Initialization Speed: The process of selecting 5 objects and registering their features using MobileSAM took approximately 0.5 seconds, while 10 objects required approximately 1.3 seconds. Since this occurs only once before a task begins, it does not impede overall system operation.
- Real-time Tracking Performance: Following initialization, ORB and PnP-based tracking maintained a real-time speed of over 30 FPS, reliably supporting the 100 Hz command generation cycle required for 7-DOF robotic arm control.

V. Discussion & Conclusion

The proposed Stepwise VLA framework addresses the black-box issue of end-to-end models and ensures safety in surgical or precision manufacturing robotics by enabling modular debugging. The recognition performance for unknown objects significantly enhances versatility in real-world industrial settings.

In conclusion, this algorithm provides a practical alternative for operating foundation models efficiently on edge devices like the Jetson Orin. Future work will focus on enhancing temporal consistency algorithms to overcome occlusion scenarios where tracked objects are temporarily hidden.

Acknowledgment

This research supported by the Research Program of the Korea Institute of Machinery and Materials and the Ministry of Trade, Industry and Energy (MOTIE) (No. RS-2024-00469885).

References

[1] A. Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," arXiv preprint arXiv:2212.06817, 2022.

[2] B. Zitkovich et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," Proceedings of The 7th Conference on Robot Learning (CoRL), 2023.

[3] M. J. Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model," Proceedings of The 8th Conference on Robot Learning (CoRL), 2025.

[4] C. Zhang et al., "Faster Segment Anything: Towards Lightweight SAM for Mobile Applications," arXiv preprint arXiv:2306.14289, 2023.