Edge-Assisted Vision-Language Perception for Resource-Constrained Unmanned Systems

Han Yuxin, Syed Murtaza Hussain Abidi, *Soo Young Shin.
Department of IT Convergence Engineering, Kumoh National Institute of Technology
ku.98953@kumoh-xin.by-works.net, murtazaabidi@kumoh.ac.kr, * wdragon@kumoh.ac.kr

Abstract

Unmanned mobile systems operating in dynamic real-world environments require not only object detection but also fine-grained understanding of object states and attributes. While large Vision-Language Models (VLMs) such as GLIP and FILIP demonstrate remarkable zero-shot generalization and pixel-level semantic alignment, their direct deployment on edge devices is hindered by stringent computational and memory constraints. This paper presents a framework that integrates fine-grained cross-modal perception with wireless cooperative inference to enable VLM-based environmental understanding in unmanned systems operating under strict latency and resource constraints. By leveraging open-vocabulary VLMs for state-aware perception and offloading computationally intensive inference tasks to nearby edge servers via latency-aware wireless coordination, the proposed framework achieves a balance between perception accuracy and system efficiency. The overall system architecture is described, key technical challenges are identified, and an experimental methodology is outlined. This work aims to advance the practical deployment of multimodal perception models on communication-constrained robotic platforms.

Keywords: Vision-Language Models, Fine-Grained Perception, Edge Intelligence,

I. Introduction

Recent advances in Vision-Language Models (VLMs) have significantly enhanced open-vocabulary object detection and fine-grained visual reasoning. Models like GLIP [1] and FILIP [2] align image regions with textual concepts at high granularity, allowing zero-shot inference without task-specific retraining ideal for unpredictable environments where unmanned aerial vehicles (UAVs) operate.

However, these models are typically large (hundreds of millions to billions of parameters) and computationally intensive, making on-device execution impractical for resource-limited platforms. Meanwhile, conventional object detectors lack the semantic depth needed for state-level understanding A co-design framework is proposed that integrates (1) fine-grained vision-language perception based on cross-modal alignment, (2) wireless cooperative inference, in which VLM computation is dynamically offloaded to edge servers according to real-time channel conditions, energy constraints, and latency requirements. This paper presents a research vision, a system architecture, and planned contributions aimed at enabling intelligent, responsive, and semantically rich perception for next-generation unmanned systems.

II. Related Work

Vision-Language Models for Robotics: Recent works explore VLMs in robotics for instruction following [3] and scene understanding [4]. However, most assume cloud connectivity or ignore real-time constraints.Edge Intelligence & Model Offloading: Prior studies [5,6] investigate DNN partitioning and offloading over wireless networks. Yet, they focus on CNNs or transformers without considering multimodal alignment or open-vocabulary semantics.

A recent study by members of the present research team introduced SafeVision[7], a vision-language reasoning framework for context-aware safety monitoring in industrial environments. The approach utilizes a frozen CLIP-ViT encoder with Low-Rank Adaptation (LoRA) to enable parameter-efficient fine-tuning and incorporates dual reasoning pathways—namely, a Scene-Aware Reasoner (SAR) for holistic scene interpretation and a Region-Focused Neural Reasoner (ReFiNER) for localized attribute verification—thereby supporting natural language–specified safety rules such as “workers must wear helmets” or “no personnel in hazard zones.” Designed for deployment on resource-constrained edge devices, SafeVision demonstrates effective zero-shot detection of safety violations

without task-specific retraining. Extending this line of work, the current study addresses the challenges of applying such vision-language reasoning capabilities to mobile unmanned systems, where dynamic mobility, limited onboard computation, and intermittent connectivity necessitate cooperative offloading strategies to achieve real-time, fine-grained environmental perception.

III. Proposed Framework

As illustrated in Fig. 1, the proposed system operates in a collaborative edge-robot setting. An UxVs equipped with a camera captures real-time visual data from the environment. Instead of running the full VLM locally, the device executes a lightweight perception frontend—which may include image preprocessing, region proposal generation, or early-stage feature extraction—to produce compact visual tokens or query embeddings. These compact representations are then transmitted over a wireless link to a nearby edge server. At the edge, a fine-tuned open-vocabulary VLM (such as GLIP) performs fine-grained inference: given both the visual input and natural-language prompts, the model leverages its cross-modal alignment capability to detect objects and classify their semantic states at pixel or region level.

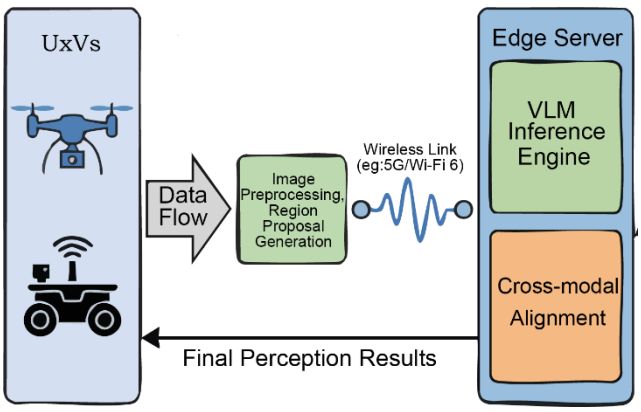

Figure 1 Proposed system architecture

IV. Research Plan and Expected Contributions

The proposed study includes the following components: (1) implementation of a baseline using GLIP on standard datasets augmented with object state annotations; (2) simulation of edge offloading via a custom Python-based wireless emulator; and (3) systematic evaluation of the trade-offs between perception accuracy and end-to-end latency under varying bandwidth and energy constraints.

Expected contributions encompass: (1) a novel integration of fine-grained vision-language model (VLM) perception with wireless cooperative offloading for unmanned systems; (2) design guidelines for VLM-aware edge inference under communication constraints

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024-00437190) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation, 50%) This research was supported by the MSIT(Ministry of Science and ICT), Korea under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2023-00259061) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation, 50%)

REFERENCES

[1] Li et al., "GLIP: Learning Open-Vocabulary Object Detection via Grounded Language-Image Pre-training," CVPR 2022.

[2] Yao et al., "FILIP: Fine-grained Interactive Language-Image Pre-Training," ICLR 2022.

[3] Shah et al., "VLMaps: Building a Visual-Language Map for Robot Navigation," RSS 2023.

[4] Liu et al., "Grounded Language-Image Pretraining for Zero-Shot Robotic Manipulation," CoRL 2023.

[5] Chen et al., "Deep Decision Making for Mobile Edge Computing," IEEE TMC 2021.

[6] Wang et al., "Joint Computation and Communication Optimization for DNN Inference in Vehicular Edge Networks," IEEE JSAC 2022.

[7] Abidi et al., "SafeVision: Vision-language reasoning for context-aware safety monitoring," Neurocomputing 2026.