KENNESAW STATE UNIVERSITY

Efficient Vision Knowledge Pipeline Systems on Resource-Constrained Edge Devices

Authors: Joshua Scarpinato, Bobin Deng, Md Romyull Islam, Xinyue Zhang, Kun Suo

Presented by

Md Romyull Islam

Kennesaw State University

Introduction & Motivation

Why Edge AI?

Deploying AI locally on edge devices offers critical advantages over cloud-based solutions:

- **Lower Latency:** Immediate processing without network round-trips.
- **Privacy:** Data stays local, reducing exposure risks.
- **Cost:** Reduces expensive network bandwidth usage.

The Challenge with VLMs

Vision Language Models (VLMs) are powerful but resource-heavy foundation modules.

- **Hardware Limits:** Mid-end edge devices lack the RAM and compute power for full VLMs.
- **Bottleneck:** Tightly coupled image encoders consume massive resources, making real-time response difficult.

The Solution: Vision Knowledge Pipeline System (VKPS)

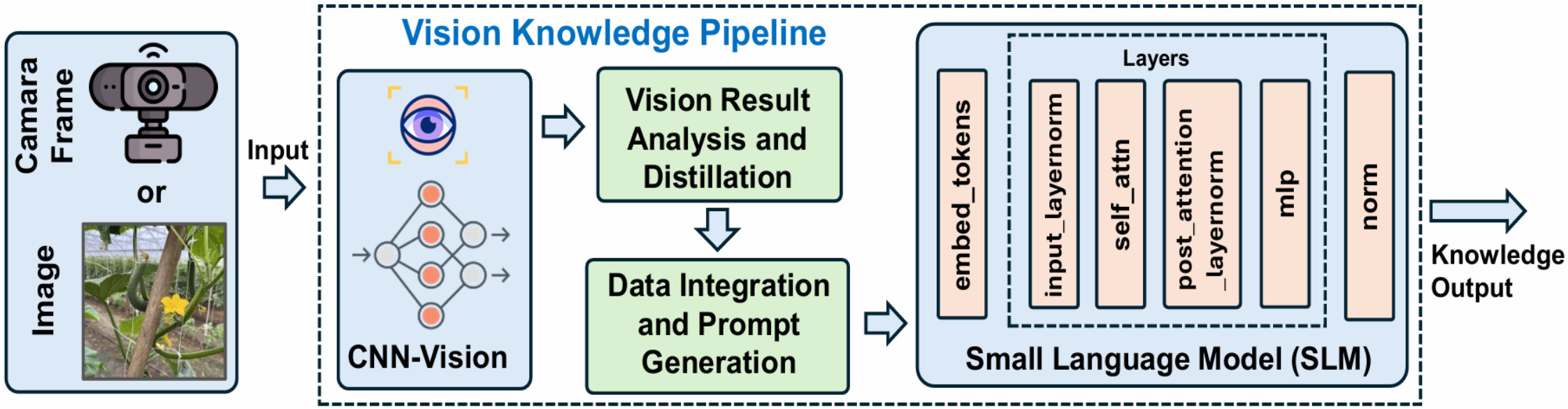
Instead of a monolithic, tight-coupled VLM, we propose a flexible, **loose-coupling approach**.

Modular Design: Replaces the heavy image encoder with a lightweight Vision-CNN (e.g., YOLO).

Efficient Reasoning: Connects the vision module to a Small Language Model (SLM) for reasoning.

Flexibility: Allows customization of vision tasks while maintaining knowledge capabilities.

VKPS Framework



The system operates in a streamlined sequence to maximize efficiency:

Input: Camera frame or static image.

Vision-CNN: Lightweight model (e.g., YOLO) detects objects and extracts visual features.

Data Integration: Analyzes vision results and generates a prompt.

SLM: Small Language Model processes the prompt to generate knowledge-based output.

Experimental Hardware

1. Raspberry Pi 5

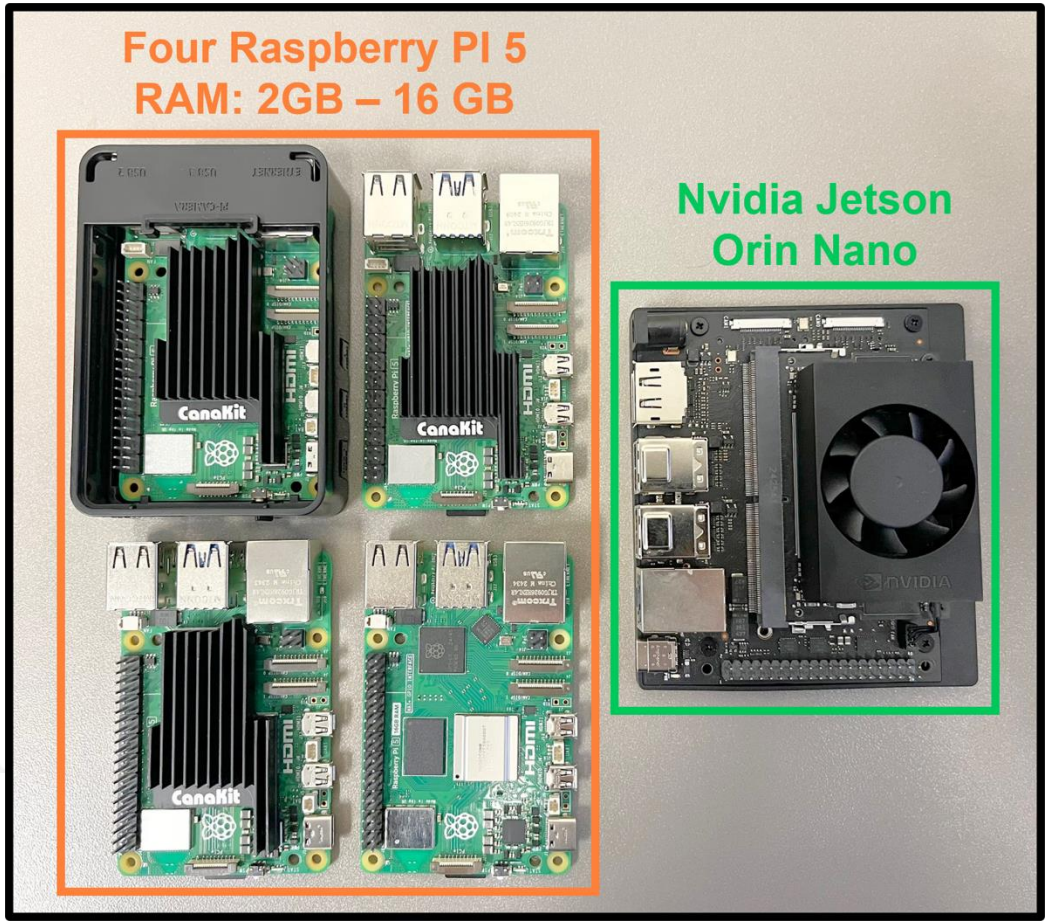
Represents common, resource-constrained mid-end edge devices.

- **CPU:** 4 cores @ 2.4 GHz
- **RAM Variants:** Tested across 2GB, 4GB, 8GB, and 16GB.
- **Constraint:** Lack of dedicated GPU makes memory management critical.

2. NVIDIA Jetson Orin Nano

Represents AI-accelerated edge hardware.

- **Specs:** 6 CPU cores, 1024 GPU cores, 8GB RAM.



AI Models Evaluated

Category	Model Name	Size / Params	Role
Vision-CNN	YOLOv8n	6.25 MiB (2.6M)	Object Detection (VKPS)
SLM	Gemma-3 / Llama-3.2	~1.2 - 1.8 GiB (1B)	Reasoning (VKPS)
VLM	SmolVLM2	2.34 GiB (2.2B)	Baseline Comparison
VLM	nanoLLaVA-1.5	2.0 GiB (1B)	Baseline Comparison

** VKPS combines Vision-CNN and SLM to rival the functionality of the larger VLMs.*

Performance Breakthrough

36.8X

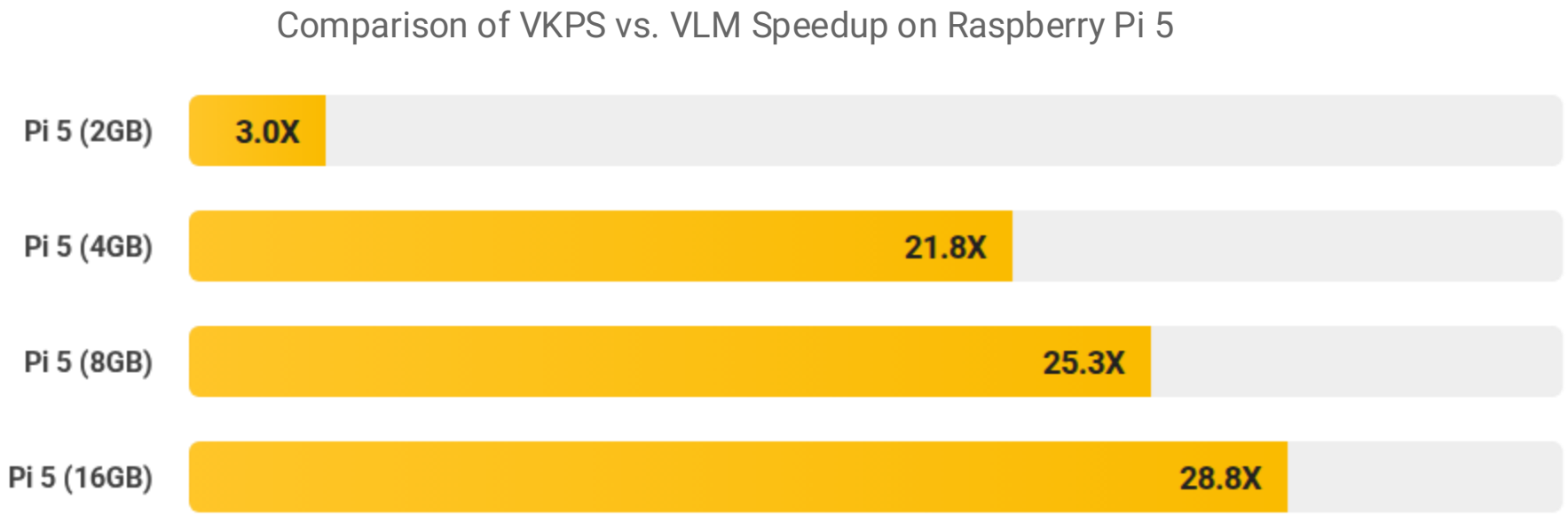
Maximum Speedup

Drastic Reduction in Latency

The VKPS approach achieves a speedup ranging from **2.37X to 36.86X** in the pre-processing stage compared to standard VLMs.

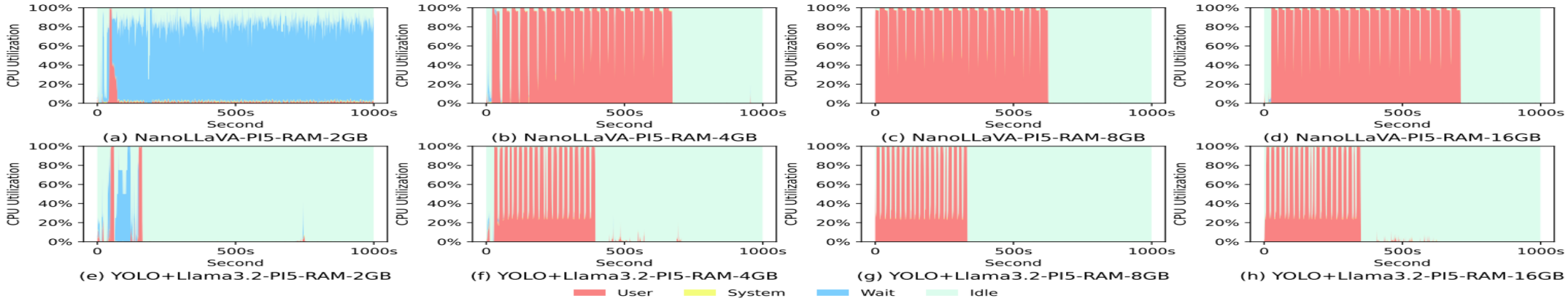
This allows edge devices to process visual data in near real-time, a feat previously unattainable for VLMs on this hardware.

Speedup Factor by RAM Size



*Data based on latency reduction in pre-processing tasks.

Resource Constraints & Analysis



RAM Bottleneck

2GB RAM is insufficient for most VLMs, causing constant disk swapping and high latency. 4GB+ is recommended for smooth operation.

CPU Efficiency

VKPS demonstrates higher active CPU utilization, indicating efficient processing rather than idle waiting times seen in VLMs.

Storage I/O

Heavy models on low-RAM devices trigger excessive SSD/Disk access, becoming a major performance killer.

Guidelines & Conclusion

- **Viable Alternative:** VKPS is a proven, high-efficiency alternative to VLMs for edge devices, offering massive speedups.
- **Hardware Design:** Future edge devices must prioritize RAM size to support vision-reasoning capabilities effectively.
- **Model Selection:** For applications requiring real-time response (e.g., Agentic AI), prioritizing Vision-CNN + SLM (VKPS) over monolithic VLMs is the optimal strategy.
- **Trade-off:** VKPS may sacrifice some generalizability for speed, making it ideal for specific, well-defined tasks.

Future Applications

The efficiency of VKPS unlocks advanced AI applications on the edge:

Smart Manufacturing: Digital twins can monitor production lines and predict failures in real-time.

Autonomous Systems: Faster processing enables quicker reaction times for robots and drones.

Privacy-First AI: Processing sensitive video data locally in hospitals or homes.

Questions?

Thank you for your attention.

Contact Me
mislam22@students.Kennesaw.edu

Kennesaw State University
College of Computing and Software Engineering