An Architecture For Adaptive Split Computing In Vision-Language Models

Tech ID: 34594 / UC Case 2026-662-0

BRIEF DESCRIPTION

An intent-aware, dual-stream AI architecture that adapts compute allocation and inference depth on embedded platforms, balancing rapid triage and detailed analysis for real-time visual understanding.

FULL DESCRIPTION

This invention introduces a novel split computing architecture for deploying multi-billion-parameter Vision-Language Models (VLMs) on embedded platforms such as drones and autonomous vehicles. An onboard controller adjusts compute placement and reasoning depth based on operator intent, mission context, and system constraints, including energy, thermal, and bandwidth limits. This enables human-in-the-loop, context-aware autonomy, allowing the system to shift between rapid triage and deeper analysis as mission priorities change. The architecture uses a cognitive-inspired dual-stream pipeline. A lightweight stream provides real-time situational awareness, while a heavier stream is invoked on demand for higher-confidence reasoning, enabling explicit “triage → escalate” behavior. Learned compression at the split boundary reduces the amount of data sent over unreliable wireless links while maintaining semantic and pixel-level accuracy. This allows large-scale multimodal AI to run efficiently in mission-critical applications such as disaster response, military reconnaissance, and autonomous inspection.

SUGGESTED USES

» Autonomous Vehicles, ADAS, and Mobility Stacks:Enables real-time, adaptive vision-language AI for navigation, perception, and decision-making in autonomous and assisted driving systems.

» Drones, Aerospace, and Defense Autonomy:Supports robust, intent-driven visual reasoning for UAVs, defense platforms, and aerospace applications in constrained and contested environments.

» Robotics & Industrial Automation:Drives smart automation and inspection in industrial robots, warehouse logistics, and field robotics with energy-efficient AI inference.

» Agriculture & Field Autonomy:Enhances autonomous farming and environmental monitoring through precision AI on agricultural and field robotics platforms.

» Inspection, Critical Infrastructure & Energy:Delivers scalable vision-language AI for remote inspection, anomaly detection, and safety compliance in infrastructure and energy sectors.

» Edge AI Platforms and Deployment Tooling:Provides hardware and software solutions to efficiently deploy large vision-language models on resource-constrained edge devices.

» Telecom / Edge Networks:Enables low-latency, bandwidth-adaptive AI inference supported by edge network infrastructure for mission-critical applications.

ADVANTAGES

» Up to 93.98% reduction in edge-side energy consumption, extending operational time on battery-powered devices.

CONTACT

Edward Hsieh
hsiehe5@uci.edu
tel: 949-824-8428.

OTHER INFORMATION

CATEGORIZED AS

» **Computer**
» Software
» **Security and Defense**
» Screening/Imaging
» **Sensors & Instrumentation**
» Environmental Sensors
» Position sensors
» **Transportation**
» Aerospace
» Automotive

RELATED CASES

2026-662-0

» Maintains inference accuracy within 1% of full cloud processing and surpasses naive image compression by over 11% in segmentation accuracy

» Enables mission-driven control over latency and accuracy trade-offs through intent-coordinated, dynamic compute placement and reasoning depth.

» Demonstrates graceful degradation in performance under constrained or fluctuating network connectivity (8–20 Mbps), ensuring robust, real-time operation.

» Produces operator-aligned outputs, adapting inference behavior to human intent and shifting mission priorities for actionable insights.

» Employs a functional dual-stream design generalizes across VLM architectures and supports multiple tasks including segmentation, captioning, and retrieval.

» Requires minimal on-device modification with no need for custom hardware or extensive model retraining, facilitating easy integration.

PATENT STATUS

Patent Pending

RELATED MATERIALS

» Bhattacharjya, R., et al. (2025). AVERY: adaptive VLM split computing through embodied self-awareness for efficient disaster response systems. arXiv.

UCI Beall Applied Innovation

5270 California Avenue / Irvine,CA 92697-7700 / Tel: 949.824.2683