# Gaze-based Attention-driven Multimodal Single-Object Inference for Microcontroller-Class Edge Vision

**Sofia Marilina Glorioso**[1] , Andrea Pietro Arena[1] , Ivana Guarneri[2] , Salvatore Sorce[1]

[1] Università degli Studi di Enna “Kore”, Enna, Italy
[2] STMicroelectronics, Catania, Italy

August 21st, 2026, Lyon, France - MCMI Workshop, co-located with ICPR 2026

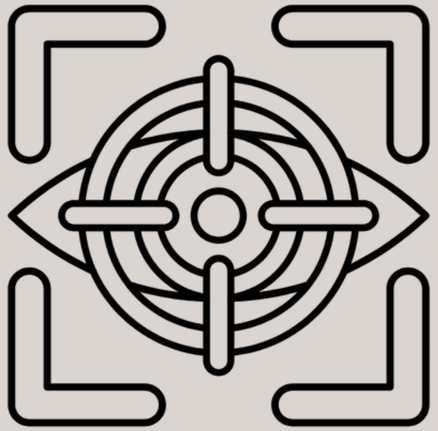
# How should gaze-aware perception be redesigned for ultra-low power edge hardware?

The idea is to treat **gaze**, the point toward which the user's attention is directed, **as a computational compression mechanism** applied before inference. Instead of processing the entire scene and using gaze as ancillary information, the system limits analysis to the region of interest identified by visual fixation, **significantly reducing computational cost** and making multimodal object recognition feasible on microcontrollers with extremely limited resources.

**Compress the problem before compressing the model**

# Overview

# 1 Introduction



Current approaches still treat gaze as a post-hoc re-ranking signal applied on top of a full-frame visual backbone.

**Three key obstacles**

1. Noise → gaze-tracker uncertainty may cause the Region of Interest (ROI) to miss the target.
2. Fragility → occlusions and low-light conditions can degrade the visual stream.
3. Data Path → repeatedly buffering and copying full-resolution frames introduces a substantial "memory tax".

We invert this logic by treating **gaze as a computational attentional prior** that guides perception before inference.

Gaze-centered ROI transforms full-frame object detection into lightweight single-object recognition, reducing memory and computation for MCU-class edge devices.

# 2 Problem Formulation

The system observes an RGB scene and receives an estimate of the user’s gaze, together with its uncertainty and, when available, an additional contextual cue. Its objective is to identify the object currently attended by the user and return both a class label and a confidence score.

**Key problems**

- Define the smallest gaze-centered ROI that still contains the attended object.
- Compensate for gaze uncertainty without losing the benefit of local inference.
- Meet strict deployment constraints in terms of SRAM, NPU workload, and latency.

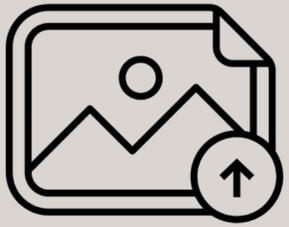
# 3 Method

## σ-Aware ROI Selection

Estimation of the gaze point and its covariance is used to define a σ-aware Region of Interest (ROI). The selected patch is then cropped and resized to 96×96 before being processed by the inference pipeline.

## Visual Encoder

An INT8 MobileNetV3-Small extracts a compact 128-dimensional visual embedding from the gaze-centered ROI.



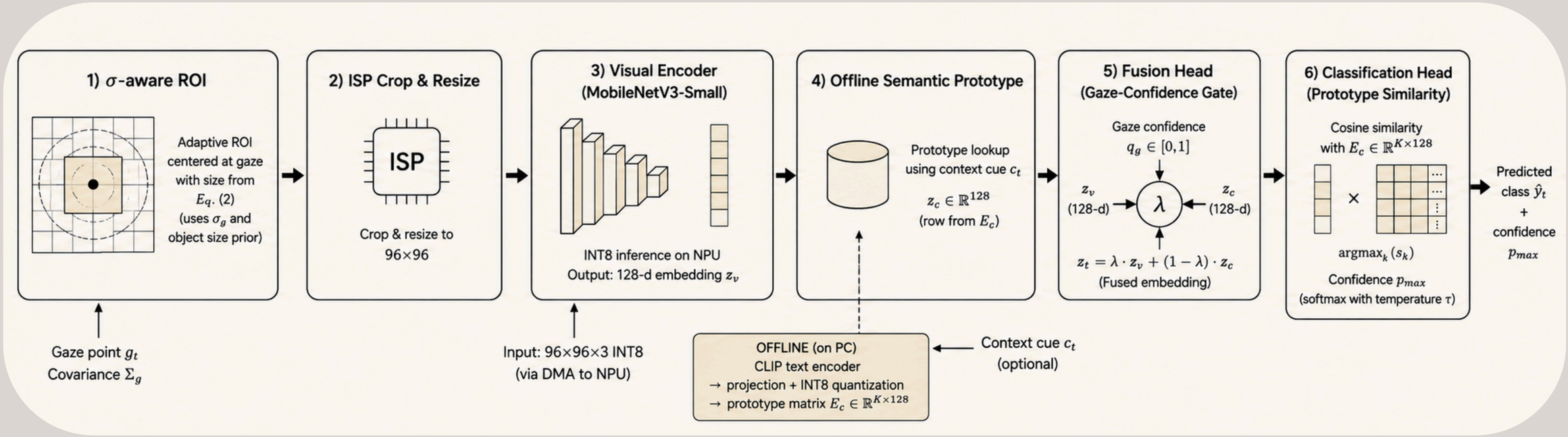
## Offline Semantic Prototypes

Instead of running a text encoder on the microcontroller, class-level semantic embeddings are precomputed offline, quantized to INT8, and stored in a compact prototype matrix.

# 3 Method

## Fusion Head

A gaze-confidence-aware scalar gate **combines** the visual embedding with the selected semantic prototype, adapting their contribution according to the reliability of the available signals. The fused representation is compared with all class prototypes using **cosine similarity**, producing the predicted object label and its confidence score.

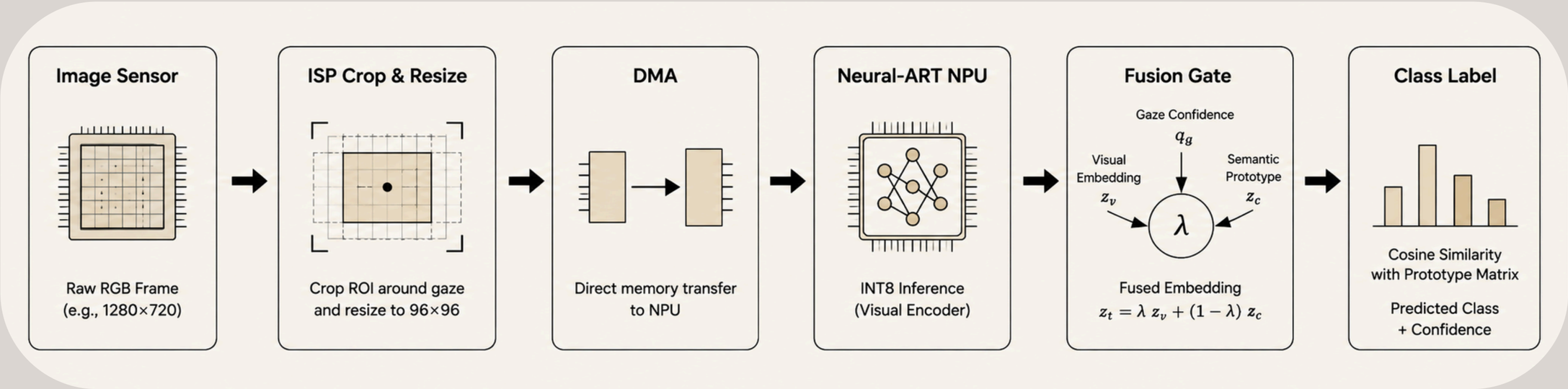


# 4 Deployment Architecture

The system is designed for the **STM32N657X0**, which integrates:

- Arm Cortex-M55 processor
- Neural-ART INT8 accelerator
- an on-chip Image Signal Processor
- ~4.2 MB of SRAM.

A direct **ISP → DMA → NPU** pipeline performs hardware ROI extraction, streaming only the 96 × 96 gaze-centered region to the NPU.



# 5 Resource Analysis

Full frame (1280 × 720)
2.76 MB SRAM → ~66% of on-chip memory.

By processing only the gaze-centered
Gaze-centered ROI (96 × 96)
27 KB buffer → ~2.74 MB saved.

The estimated resource budget of the complete pipeline is:

- Peak SRAM: ~332 KB, about 8% of the available memory
- External flash: ~1.56 MB
- Computational cost: 5.51 MMAC
- Latency target: < 33 ms

ROI-only inference + offline prototypes + ISP → NPU streaming enable MCU deployment.

# 6 System behaviour in extreme cases

## Low gaze confidence

When gaze confidence is low, the fusion gate suppresses the noisy visual embedding and falls back to the semantic prototype.

## Missing contextual cue

Without contextual information, the system collapses to visual-only inference.

## Both modalities are weak

When both signals are weak, the system produces a low-confidence prediction, allowing an upstream module to reject it.

# 7 Limitations

**Gaze Noise:** real trackers may suffer from drift, blinks, and non-Gaussian errors.

**Object Size:** ROI sizing relies on an estimated object size.

**Multimodal Gain:** multimodal benefit remains to be validated.

**Hardware Validation:** memory and latency benefits must be confirmed on the target device.

**Limited Scope:** proposed method is limited to closed-set, single-object inference.

# 8 Conclusion

This work presented a gaze-conditioned multimodal framework for microcontroller-class object recognition, based on three coupled design moves: a σ-aware ROI policy, offline INT8 semantic prototypes, and a direct ISP → DMA → NPU data path.

Future work will validate the proposed framework through a pre-registered on-silicon protocol but our central object is clear:

**Compress the problem before compressing the model**

# Thank You
for your attention

## Gaze-based Attention-driven Multimodal Single-Object Inference for Microcontroller-Class Edge Vision

Sofia Marilina Glorioso, Andrea Pietro Arena, Ivana Guarneri, Salvatore Sorce

**Sofia Marilina Glorioso**
PhD Student - University of Enna “Kore”, Enna, Italy

sofiamarilina.glorioso@unikorestudent.it