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

Sofia Marilina Glorioso[1][0009−0000−8371−5543], Andrea Pietro Arena[1][0009−0001−6138−615X], Ivana Guarneri[2], and Salvatore Sorce[1][0000−0003−1976−031X]

[1] Università Kore di Enna, Enna, Italy
[sofia.glorioso|andrea.arena]@unikorestudent.it
salvatore.sorce@unikore.it
[2] STMicroelectronics, Catania, Italy ivana.guarneri@st.com

Abstract. Eye tracking can reveal which object a user is likely attending to, but most vision systems still process the entire camera frame before using gaze information. This is inefficient for smart glasses, assistive devices, and other always-on edge devices, where memory, energy, and latency budgets are extremely limited. We propose a microcontroller-oriented framework that uses gaze as an early attention cue: the estimated fixation point selects a small region of interest around the likely attended object, and the vision model performs single-object inference on that cropped region instead of running full-frame detection. To support multimodal reasoning without deploying a runtime language model, the system uses a compact table of precomputed class prototypes and combines them with visual features through a gaze-confidence-aware fusion gate. The architecture is designed for STM32N657X0-class hardware[3] and exploits the Image Signal Processor, Direct Memory Access, and Neural Processing Unit data path to avoid storing full camera frames in on-chip memory. The main contribution is a hardware-aware architectural design together with a pre-registered evaluation protocol. The expected outcome is to reduce memory and computation substantially while preserving predictive accuracy within a closed class set, with explicit tests for gaze-noise robustness, modality dropout, and on-device feasibility.

Keywords: Multimodal Fusion · Gaze-Conditioned Inference · Cross-Modal Feature Extraction · Robustness under Modality Dropout · Edge AI · Microcontroller-Class Inference · Hardware-Algorithm Co-Design.

1 Introduction

Eye tracking is central to understanding user attention, intention, and cognitive processes, and has long served in human–computer interaction as both an explicit input modality and an implicit cue of user focus.

[3] https://www.st.com/en/microcontrollers-microprocessors/stm32n657x0.html

At the same time, microcontroller-class platforms with lightweight neural accelerators are enabling always-on embedded AI in wearables, appliances, and assistive devices. These platforms impose strict memory, energy, and latency constraints: a smart-glasses system, for instance, must process a camera stream, run inference, and return a result within a sub-Watt envelope. This makes it impractical to simply port desktop or cloud-based gaze-aware perception pipelines to the edge.

The key question is therefore how gaze-aware perception should be redesigned for ultra-low-power hardware. While most approaches focus on compressing the model through pruning, quantisation, distillation, or architectural redesign, this still assumes that every frame must first be analysed in full. We instead treat gaze as an attentional signal that reduces the perception problem before model compression: the system restricts inference to a localised region of interest around the user's fixation, turning multi-object scene analysis into single-object inference over a small foveated patch. Related work has explored gaze-guided detection [9,7] and saliency-driven recognition [2], but gaze is typically used as a *post-hoc* re-ranking signal on top of a full-frame backbone.

Three obstacles make this difficult on microcontroller-class hardware. *Noise:* consumer trackers report 1–2° angular errors [6], so ROI policies must account for gaze uncertainty. *Fragility:* when vision degrades, runtime text encoders such as TinyBERT [5] exceed the 4.2 MB SRAM limit of target devices [19]. *Data path:* naïve implementations copy full sensor frames before inference, creating a multi-megabyte memory cost that can erase the benefit of cropping.

We propose an attention-driven perception framework that uses gaze-related cues to identify the region of the scene corresponding to the user's current focus. It deals with each of the obstacles above with as many design moves: (i) a σ-aware ROI policy sizes the crop as a function of the gaze covariance Σ_g, with a closed-form crossing σ_{cross}; (ii) a class-prototype backstop replaces the runtime text encoder with an offline-precomputed INT8 table, fused with the visual head by an inner-product lookup and a gaze-confidence-aware scalar gate that adds zero NPU operators and supports cross-modal feature exchange under variable input quality; (iii) an ISP→DMA→NPU data path (Fig. 1) which streams only the gaze-anchored ROI tile to the Neural-ART NPU.

The contribution is fourfold: *(i)* a formal cross-modal problem formulation of gaze-conditioned single-object inference under a microcontroller envelope (Section 3); *(ii)* a closed-form σ-crossing analysis with the corrected factor-of-two algebra (Eq. (5)); *(iii)* an offline INT8 class-prototype fusion mechanism replacing runtime text encoders within the 4.2 MB SRAM constraint by three orders of magnitude (Section 4); and *(iv)* a pre-registered empirical protocol with five falsifiable hypotheses and explicit refutation criteria, executable on STM32N657X0 silicon (Section 7).

The rest of the document is organised as follows: Sections 2–7 present the framework (position, §2; problem formulation §3; method and corrected σ-crossing §4; STM32N657X0 resource budget, §5; discussion of fusion-gate limits, load-bearing conjunction, and threats §6; pre-registered protocol, §7),

whereas §8 sums up with the conclusions and shortly introduce some further developments.

2 Related Work

Three bodies of work intersect at this proposal: gaze-guided visual recognition, multimodal fusion under tight memory budgets, and attention mechanisms on edge hardware.

Gaze-guided visual recognition. Human gaze has long been used as an auxiliary signal for detection, segmentation, and image understanding [6,15]. Early head-mounted-tracker systems used fixation points to gate object recognition [23,2,7], while recent gaze-following and gaze-target-detection methods rely on deep transformer pipelines and large-scale learned encoders [16,9,22,24,18], including benchmarks such as GOO [21]. Barz et al. [1] further treat gaze noise as a first-class interaction signal. Across these systems, however, the visual backbone typically still runs on the full frame, with gaze used for re-ranking or learned attention. Our framework instead uses gaze *upstream* of the backbone to crop the input tensor itself, so computation is spent only on the gazed neighbourhood. The σ-aware sizing rule of Section 4.1 is the technical core of this departure.

Multimodal fusion under embedded budgets. Vision-language and class-prior fusion is well established at cloud and mobile-GPU scales, where runtime text encoders are common [5,14,20]. Embedded systems usually either deploy a distilled text encoder [5], whose INT8 footprint can still exceed microcontroller SRAM budgets, or drop the text stream entirely. In contrast, we precompute textual/class-prior embeddings offline into a compact INT8 lookup table and perform only a single inner-product fusion step at inference time. The full fusion path must fit within the on-chip SRAM of an STM32N6 and add zero NPU operators to the deployed graph.

Edge attention and embedded ML deployment. Compact INT8 backbones such as MobileNetV3 [4] and MCUNet [12,11] target microcontroller-class memory budgets, while CMSIS-NN [8], TFLite Micro [3], and NPU-aware dataflow work [10] support deployment. These works mainly improve *intra-model* efficiency. Our position is complementary: gaze can reduce the inference *problem itself* before model compression, a claim tested falsifiably in Section 7.

Differentiation. The contribution is not gaze-based cropping alone, which dates back to Toyama et al. [23] and appears in later host-class pipelines [2,7,22,24,18,9]. Rather, it is the *conjunction* of three design moves: (i) an uncertainty-aware ROI policy with a closed-form containment guarantee parameterised by gaze covariance, adapting Barz et al. [1]; (ii) replacement of a runtime BERT-class text stream by an offline CLIP-style class-prototype

Table 1. Notation used throughout the formulation and method.

Symbol	Meaning	Symbol	Meaning
$I_t \in \mathbb{R}^{H\times W\times 3}$	RGB scene frame	$R_t \in \mathbb{R}^{96^2\times 3}$	ISP-cropped ROI patch
$g_t \in \mathbb{R}^2$	Estimated gaze point	c_t	Optional context cue
$\Sigma_g = \mathrm{diag}(\sigma_x^2, \sigma_y^2)$	Per-frame gaze covariance	$f_v : R_t \mapsto z_v \in \mathbb{Z}_8^d$	INT8 encoder, d=128
$q_g \in [0,1]$	Gaze-confidence summary	$E_c \in \mathbb{Z}_8^{K\times d}$	Prototype matrix (flash)
$\hat{s}_o$	Class-prior object full side	$z_c \in \mathbb{Z}_8^d$	Semantic prototype $E_c[c_t]$
$\rho = \hat{s}_o/2$	Per-axis half-extent	$\lambda_t \in [0,1]$	Fusion gate, (6)
α, β	ROI sizing (α=β=2)	$z_t \in \mathbb{Z}_8^d$	Fused embedding, (7)
r	ROI side length, (2)	$\hat{y}_t$, $p_{\max}$	Predicted label / confidence

matrix [14], far smaller than the MobileBERT lower bound [20]; and (iii) MCU-class deployment co-designed with the STM32N657X0 ISP→DMA→NPU data path [19]. Each component has precedent, but only their combination reaches the deployment envelope of Section 5: $\leq 8\%$ of SRAM and a 5.5-MMAC operating point with closed-form guarantees.

3 Problem Formulation

From now on, we will follow the notation in Table 1. We consider an embedded perception system observing a scene through a camera attached to or co-located with a microcontroller-class processor, and equipped with an external gaze-tracking module estimating the user's fixation point in the same image plane. Let $I_t \in \mathbb{R}^{H\times W\times 3}$ be the RGB scene frame at time t, $g_t = (x_t, y_t) \in \mathbb{R}^2$ the estimated gaze point, $\Sigma_g = \mathrm{diag}(\sigma_x^2, \sigma_y^2)$ a per-frame estimate of the gaze covariance, and $q_g \in [0,1]$ a scalar gaze-confidence summary derived from Σ_g (for instance, $q_g = \exp(-\mathrm{tr}(\Sigma_g)/\tau_q)$ for a calibration constant τ_q). The mapping from tracker-specific telemetry onto q_g is deployment-dependent; we treat q_g as the system-level summary consumed by the fusion gate of Section 4.5. The system additionally has access to a possibly empty contextual cue $c_t \in \{1,\dots,K\} \cup \{\varnothing\}$ that, when available, biases the categorical prior toward a subset of the K candidate classes.

3.1 Inference and deployment objective

The system must produce a label $\hat{y}_t \in \{1,\dots,K\}$ and a confidence $p_{\max} \in [0,1]$, degrading gracefully toward the more reliable channel when either signal is uninformative; the binding geometric correctness condition is the per-frame containment of the attended-object box inside the cropped region of interest. The deployment objective combines task accuracy with hard envelope constraints:

$$\begin{aligned} &\min_{\theta,\phi}\ \mathbb{E}_{(I,g,y)}\big[\ell\big(f_\phi(\rho_\theta(I,g,\Sigma_g)),\, y\big)\big] \\ &\text{s.t. } \mathrm{MAC}(f_\phi) \leq M_{\mathrm{NPU}}, \quad \mathrm{SRAM}(f_\phi) \leq M_{\mathrm{SRAM}}, \quad T_{\mathrm{e2e}} \leq T_{\mathrm{deadline}}. \end{aligned} \tag{1}$$

The constraints are not soft penalties but binding limits: $M_{\mathrm{SRAM}} = 4.2\,\mathrm{MB}$ on the STM32N657X0 [19], and $T_{\mathrm{deadline}} = 33\,\mathrm{ms}$ for a 30 fps interaction loop. A configuration that violates any of them is non-deployable, regardless of accuracy.

3.2 Gaze-uncertainty model and ROI objective

We model the gaze tracker as producing an unbiased estimate of the true attended-object centroid $g^\star$ with additive Gaussian residual $g_t = g^\star + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \Sigma_g)$; calibration drift and blink-induced gaps are deferred to §6.3. Treating the attended object as an axis-aligned bounding box of full side $\hat{s}_o$, the per-axis half-extent is $\rho = \hat{s}_o/2$. The ROI objective is to choose, given g_t and Σ_g, the smallest crop $R(g_t, r)$ such that $\Pr(B_o \subseteq R(g_t, r)) \geq 1 - \delta$.

4 Method

The proposed method turns gaze into an early filtering mechanism for visual inference. Instead of first analysing the full camera frame and then using gaze to interpret or re-rank the result, the system uses the estimated fixation point to select a small region of interest that is likely to contain the object currently attended by the user. This region is then resized to a fixed input tensor and processed by a lightweight visual encoder. The resulting visual representation is combined with a compact semantic prior, stored as an offline class-prototype table, so that the system can retain some benefits of multimodal inference without executing a text encoder on the microcontroller. Finally, a confidence-aware fusion gate balances the visual and semantic streams according to gaze reliability, and a prototype-similarity head produces the final class prediction.

Each design choice follows from the deployment target. The uncertainty-aware ROI policy is used because gaze estimates are noisy and the crop must remain large enough to contain the attended object. The 96×96 visual input is chosen to keep the visual encoder within the memory and computation envelope of the target device. Offline semantic prototypes replace runtime language models because the latter exceed the available on-chip memory. The scalar fusion gate is deliberately simple so that multimodal fusion adds negligible computational cost. Algorithm 1 summarises the resulting single-frame inference pipeline; the following subsections describe each component in detail.

4.1 Uncertainty-aware ROI policy

The ROI side length is parameterised as

$$r(\sigma) \;=\; 2\alpha\,\sigma + \beta\,\hat{s}_o, \tag{2}$$

where $\hat{s}_o$ is a class-prior estimate of the attended-object full side, α is a dimensionless gaze-tail-coverage factor (so that $\alpha = 2$ absorbs a 2σ tail per axis), and $\beta \geq 1$ is a margin factor over the bare object extent. The defaults $\alpha = \beta = 2$ size the ROI to absorb a 2σ gaze tail and a one-object-diameter margin. With half-extent $\rho = \hat{s}_o/2$, the per-axis containment margin is

$$M \;=\; r/2 - \rho \;=\; \alpha\,\sigma + (\beta - 1)\,\hat{s}_o/2, \tag{3}$$

yielding the per-axis containment factor $2\Phi(M/\sigma)-1 = 2\Phi\big(\alpha+(\beta-1)\hat{s}_o/(2\sigma)\big)-1$ and the per-frame lower bound

$$\Pr\big(B_o \subseteq R(g_t, r)\big) \;\geq\; \prod_{a\in\{x,y\}} \left(2\Phi\left(\alpha + \frac{(\beta-1)\hat{s}_o}{2\sigma_a}\right) - 1\right). \tag{4}$$

For $\alpha = \beta = 2$, the bound (4) simplifies to $(2\Phi(2)-1)^2 \approx 0.911$ in the gaze-noise-dominated limit $\hat{s}_o/\sigma \to 0$ (per-axis margin $\alpha = 2$, $\Phi(2) \approx 0.97725$), and rises toward 1 as $(\beta-1)\hat{s}_o/\sigma \to \infty$. The deployment can therefore expose a tunable miss-rate δ and choose α, β such that $\Pr(B_o \subseteq R_t) \geq 1-\delta$ for the calibrated Σ_g and $\hat{s}_o$.

4.2 σ-crossing analysis (corrected closed form)

At what gaze-noise level σ does the gaze-tail term dominate the object-margin term in M, switching the ROI from object-dominated to noise-dominated sizing? Setting the two terms in (3) equal,

$$\alpha\,\sigma_{\text{cross}} \;=\; (\beta-1)\,\hat{s}_o/2 \quad\Longrightarrow\quad \boxed{\sigma_{\text{cross}} \;=\; \frac{(\beta-1)\,\hat{s}_o}{2\alpha}}. \tag{5}$$

Equation (5) is the corrected closed form. With representative parameters $\alpha = 2$, $\beta = 2$, and $\hat{s}_o = 100\,\text{px}$ (a small object on a VGA-class sensor), the crossing falls at $\sigma_{\text{cross}} = (\beta-1)\hat{s}_o/(2\alpha) = (1)\cdot 100/(2\cdot 2) = 100/4 = 25\,\text{px}$. Below σ_{cross} the ROI is determined by the object's prior size and the gaze contribution is essentially free; above it, the ROI grows linearly with σ and the compute saving over full-frame inference shrinks. The factor of two in the denominator follows directly from the half-extent $\rho = \hat{s}_o/2$; we re-derive the closed form independently to fix a factor-of-two algebraic slip in our prior emulation draft.

Translating to angular error at approximately $21\,\text{px}/^\circ$, calibrated screen-based and head-mounted devices fall in the object-scale-limited regime; uncalibrated webcams and post-drift mobile devices fall in the gaze-noise-limited regime [1,6]. The same closed-form rule covers both. Anisotropic gaze noise decouples per axis: (2)–(5) apply independently with σ_x, σ_y in place of σ.

4.3 Visual encoder

The cropped patch R_t is ISP-resized to $96 \times 96 \times 3$ INT8 and consumed by an INT8 MobileNetV3-Small backbone [4] at width-multiplier 0.75, with the final 1×1 convolution replaced by a width-d projection producing $z_v \in \mathbb{Z}_8^{128}$. Separable-convolution scaling rescales the published 224^2 cost: $\text{MAC}(f_v) \approx 30 \cdot (96/224)^2 \approx 5.5\,\text{MMAC}$, with about 1.5 MB of INT8 weights in external flash. MCUNetV2 [11] is the retreat path if 96^2 accuracy proves insufficient.

4.4 Offline semantic prototypes

A runtime BERT-class encoder [5] is the canonical multimodal-fusion alternative; the smallest credible mobile variant [20] is approximately 25 MB at INT8, infeasible alongside the visual encoder, camera path, and OS in 4.2 MB of SRAM. We therefore relocate the textual encoder to deployment time and replace it with a frozen class-prototype matrix $E_c \in \mathbb{Z}_8^{K\times d}$ stored in external flash; for $K = 80$ COCO classes [13] and $d = 128$, this is 10,240 B, four orders of magnitude smaller than MobileBERT.

Each prototype e_k is computed offline as the L2-normalised mean of CLIP ViT-B/32 text-encoder embeddings [14] over a small prompt set, passed through a learned projection $P \in \mathbb{R}^{128\times 512}$ (trained jointly with f_v, per-tensor symmetric INT8-quantised). On-device retrieval is a single 128-byte read: $z_c = E_c[c_t]$ when $c_t \neq \varnothing$, and the precomputed uniform mean $\bar{e} = \frac{1}{K}\sum_k E_c[k]$ otherwise. The fallback approximately vanishes from the cosine-similarity argmax of (8) under approximately class-invariant inner products $\langle \bar{e}, E_c[k]\rangle$, satisfied empirically for the $K = 80$ COCO prototypes; §6.1 treats the limit. Both z_t and the rows of E_c are L2-normalised so that $z_t^\top E_c[k]$ is exactly cosine similarity.

4.5 Fusion head

The visual and semantic embeddings are fused by a gaze-confidence-aware sigmoid gate:

$$\lambda_t \;=\; \sigma\big(W_g\,[\,z_v;\, z_c;\, q_g\,] + b_g\big), \tag{6}$$

with $W_g \in \mathbb{Z}_8^{1\times(2d+1)}$ and $b_g \in \mathbb{Z}_8$, followed by a convex combination at the embedding level:

$$z_t \;=\; \lambda_t\, z_v + (1-\lambda_t)\, z_c. \tag{7}$$

At $d = 128$ the gate occupies 257 INT8 bytes and 257 MACs, two orders of magnitude cheaper than a concat-and-linear $W \in \mathbb{Z}_8^{d\times 2d}$. The convex combination of INT8 vectors uses an INT16/INT32 accumulator with a single re-quantisation step at the output, a standard NPU pattern. Beyond cost, (6)–(7) expose gaze confidence as an explicit lever: when training has set the gate's q_g component with the appropriate sign, $q_g \to 1$ drives $\lambda_t \to 1$ (visual-led) and $q_g \to 0$ drives $\lambda_t \to 0$ (prototype-led). The three canonical limits are analysed in §6.1.

4.6 Classification head

The classification head is the cosine similarity between the fused embedding and each prototype, divided by a temperature τ:

$$\hat{y}_t \;=\; \arg\max_{k\in\{1,\dots,K\}} \frac{z_t^\top E_c[k]}{\tau}, \qquad p_{\max} \;=\; \mathrm{softmax}\big(z_t^\top E_c/\tau\big)_{\hat{y}_t}. \tag{8}$$

This head is parameter-free and shares the representation space with the prototypes, so training-time alignment is automatically respected at inference. A linear head $W_h \in \mathbb{Z}_8^{d\times K}$ would add $d \cdot K = 10{,}240$ INT8 bytes and MACs; the similarity head saves both at no loss in the alignment-trained regime.

Algorithm 1 Gaze-conditioned multimodal ROI inference (single frame).

Require: frame I_t; gaze g_t with covariance Σ_g and confidence q_g; optional context c_t
Require: frozen f_v, W_g, b_g, E_c, P; class-prior scale $\hat{s}_o$
$r_x \leftarrow 2\alpha\,\sigma_x + \beta\,\hat{s}_o$; $\quad r_y \leftarrow 2\alpha\,\sigma_y + \beta\,\hat{s}_o$ ▷ (2)
$R_t \leftarrow \mathrm{ISP\_crop\_resize}(I_t,\, g_t,\, r_x,\, r_y,\, 96 \times 96)$ ▷ ISP → DMA → NPU
$z_v \leftarrow f_v(R_t)$ ▷ INT8 on Neural-ART NPU
$z_c \leftarrow E_c[c_t]$ **if** $c_t \neq \varnothing$ **else** $\frac{1}{K}\sum_k E_c[k]$ ▷ flash lookup, §4.4
$\lambda_t \leftarrow \sigma\big(W_g[z_v;\, z_c;\, q_g] + b_g\big)$ ▷ (6)
$z_t \leftarrow \lambda_t z_v + (1-\lambda_t) z_c$ ▷ (7)
$\hat{y}_t \leftarrow \arg\max_k \big(z_t^\top E_c[k]\big)/\tau$; $\quad p_{\max} \leftarrow \mathrm{softmax}(\cdot)_{\hat{y}_t}$
return $(\hat{y}_t,\, p_{\max})$

4.7 Inference algorithm and temporal smoothing

The single-frame algorithm (Algorithm 1) runs at the camera's native frame rate. Across consecutive frames within a fixation, we apply a sliding-window majority vote of length $W = 3$: at 30 Hz, three frames span 100 ms, inside the 200–300 ms minimum-fixation envelope reported for natural-scene viewing [15]. The window therefore stays inside a single fixation and inherits the spatial-stability envelope of the attentional system. Saccade-onset frames (detected by the gaze tracker as a velocity transient and tested independently in F2) trigger a one-frame compute skip rather than committing to an inference whose ROI may straddle two fixations.

4.8 Training

Only the visual encoder f_v, the projection P, the gate parameters (W_g, b_g), and the temperature τ are learned; the prototype matrix E_c is frozen throughout training and at deployment, with gradient flowing only through z_v and P. Training samples carry a context cue $c_t = y_t$ so the KD term in (9) is always well-defined; at inference, c_t may be $\varnothing$ and the system falls back to the uniform-mean prototype as in Algorithm 1. The training loss combines a standard cross-entropy term over (8), a distillation-style alignment between visual-only and text-only posteriors, and a cosine alignment that pulls each z_v toward its ground-truth prototype:

$$\mathcal{L}_{\text{train}} \;=\; \mathcal{L}_{\text{CE}}(\hat{y}_t, y_t) \;+\; \lambda_{\text{KD}}\,\mathrm{KL}\big(p_v \,\|\, \mathrm{sg}\, p_c\big) \;+\; \lambda_{\text{align}}\big(1 - \cos(z_v, E_c[y_t])\big), \quad (9)$$

with $p_v = \mathrm{softmax}(z_v E_c^\top/\tau)$, $p_c = \mathrm{softmax}(z_c E_c^\top/\tau)$, and sg a stop-gradient. We default to $\lambda_{\text{KD}} = 0.5$ and $\lambda_{\text{align}} = 0.1$. During the final tenth of training we apply per-tensor symmetric quantisation-aware training to f_v, which the reference characterisation of MobileNetV3-Small reports to absorb essentially the full INT8 accuracy gap relative to FP32 [4].

5 Deployment Architecture and Resource Analysis

We instantiate the architecture against the STMicroelectronics STM32N657X0 [19], which couples an Arm Cortex-M55 core at up to

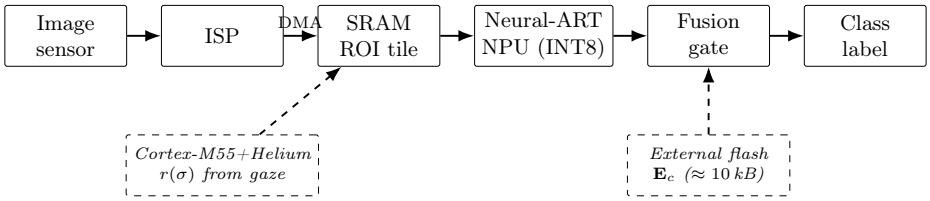

Fig. 1. Hardware-aware ISP→DMA→NPU data path on the STM32N657X0.

800 MHz (Helium MVE, FPU) with a 600 GOPS Neural-ART INT8 accelerator and approximately 4.2 MB of contiguous on-chip SRAM shared by weights, activations, frame buffers, and stacks. On-chip peripherals include an ISP with hardware crop and downsize, MIPI CSI-2 / parallel DCMI cameras, and a DMA path streaming ISP output in place to the accelerator; external Octo-SPI XIP flash ($\geq$ 128 MB in reference designs) holds code, weights, and constant tables.

5.1 The image-signal-processor data path is structural

A CPU-managed crop would buffer the full sensor frame in on-chip SRAM before the NPU touches the image. At $1280 \times 720 \times 3$ INT8 the frame buffer alone is

$$B_{\text{frame}} \;=\; 1280 \cdot 720 \cdot 3\,\text{B} \;\approx\; 2.7\,\text{MB}, \tag{10}$$

roughly 64% of the available SRAM. The architecture instead parameterises the on-chip ISP crop and downsize by (g_t, r) and DMA-streams the resulting patch directly into the NPU tensor arena (Fig. 1). The gaze-anchored ROI tile is streamed directly from the on-chip ISP into the Neural-ART NPU's tightly-coupled activation arena, eliminating the multi-megabyte sensor frame-buffer copies of a naïve port. The offline class-prototype table $\mathbf{E}_c$ resides in external flash and feeds the fusion gate via a single inner-product lookup; the Cortex-M55 with Helium MVE computes the ROI parameters of Eq. (2) in well under one millisecond. The only frame-related cost charged to on-chip SRAM is the ROI buffer

$$B_{\text{roi}} \;=\; 96 \cdot 96 \cdot 3\,\text{B} \;\approx\; 27\,\text{KB}, \tag{11}$$

saving $B_{\text{frame}} - B_{\text{roi}} \approx 2.74$ MB (about 65% of on-chip SRAM). The ISP path is therefore the principal reason the pipeline fits N6-class hardware. Two constraints follow: (g_t, r) must be expressible as an axis-aligned ISP crop downsized to 96×96, and the input crop must be at least 96×96 in sensor pixels because the ISP cannot upscale.

5.2 Component-by-component resource budget

Table 2 reports the per-component flash and on-chip SRAM (in kB), MAC count (x10^6), and analytic NPU latency floor (in μs). The encoder activation peak excludes the 27 KB input ROI buffer (allocated separately). The total row's FIT verdict is conditional on the conjunction of three coupled moves: offline prototypes, ROI-only classification, and the ISP→DMA→NPU path. INT8 storage

Table 2. Per-component resource budget on the STM32N657X0.

Component	Flash (KB)	Peak SRAM (KB)	MMAC	NPU floor (μs)	Source / fit
Camera + ISP path	0	0	0	:	hardware path [19]
ROI buffer (96×96×3, INT8)	0	27	0	:	(11)
f_v: MobileNetV3-S, w=0.75, 96^2 INT8	$\approx 1\,500$	≤ 220	≈ 5.5	≈ 9	[4], §4.3
E_c: prototype matrix, 80×128 INT8	10	0.13	0	< 0.1	§4.4, [14]
Fusion gate ϕ	< 1	< 1	< 0.001	negligible	(6)–(7)
Similarity head (no extra params)	0	0.32	0.01	negligible	(8)
X-CUBE-AI runtime + descriptors	≈ 50	≈ 64	0	0	vendor runtime [3]
Total (proposed pipeline)	$\approx 1\,561$	≈ 332	≈ 5.51	≈ 9	**FIT** (8% of 4.2 MB)
For comparison (off-chip-class baselines, listed for context only):					
Full-frame YOLOv8-class detector [17]	$\geq$ 12,000	$\geq$ 50,000	$\sim$ 8,700	:	not deployable on STM32N6
Runtime TinyBERT-4 [5] (text path)	$\geq$ 14,700	$\geq$ 14,700	:	:	exceeds 4.2 MB SRAM by $\sim 6\times$

and per-tensor symmetric quantisation are assumed throughout; each cell follows from the cited primary source or simple arithmetic on it.

The encoder activation peak is bounded at 220 KB by the worst-case MobileNetV3-Small intermediate tile at 96^2 INT8 with double-buffered ping-pong; the X-CUBE-AI tensor arena adds approximately 64 KB. Under the rated $R_{\mathrm{NPU}} = 600$ GOPS INT8, the NPU-floor latency is $T_{\mathrm{NPU,min}} \geq 2 \cdot \mathrm{MMAC} \cdot 10^6 / R_{\mathrm{NPU}} \approx \mathrm{MMAC}/300$ ms, i.e. about 9 μs for f_v. The ≈ 5.5 MMAC figure follows the reference scaling $30\,\mathrm{M} \cdot (96/224)^2 \approx 5.5$ MMAC for MobileNetV3-Small at width w=0.75 when the input is rescaled from 224^2 to 96^2 [4]. CMSIS-NN-class utilisation envelopes [8] place realised end-to-end latency one to two orders of magnitude above this floor, putting the expected per-frame deployment in the 0.1–1 ms range, well within the 33 ms budget for 30 Hz interaction.

5.3 Verdict and design-move ledger

The pipeline fits the STM32N657X0 envelope: peak on-chip SRAM is approximately 332 KB out of the 4.2 MB ceiling, i.e. $332/4{,}300 \approx 7.7\% \approx 8\%$; external flash is approximately 1.5 MB; and per-inference cost is approximately 5.5 MMAC. The fit rests on the conjunction of three design moves, not on tuning, and removing any one pushes the budget out of envelope. The conjunction's mechanistic role is analysed in detail in §6.2; here we summarise the per-move contribution in budget terms. *Offline prototypes* collapse a TinyBERT-4 path [5] that exceeds the SRAM ceiling roughly 3.5× into a single 10 KB flash read per inference. *ROI-only classification* replaces a full-frame detector [17] that exceeds the activation budget at typical input resolution. The *ISP→DMA→NPU data path* (Fig. 1) keeps the 2.7 MB sensor-frame buffer of (10) off the SRAM arena, charging only the 27 KB ROI buffer of (11).

6 Discussion

6.1 Fusion-gate behaviour at q_g extremes

The fusion equations (6)–(7) admit three canonical limits, composed without any branch in the inference graph. *First*, when $q_g \to 0$ the gate-trained weight

on q_g drives $\lambda_t \to 0$, so $z_t \to z_c$: the system falls back to the textual prior and suppresses the noisy visual embedding. *Second*, when $c_t = \varnothing$ the uniform-mean $z_c = (1/K)\sum_k E_c[k]$ is a constant additive shift across all classes, vanishes from the cosine-similarity argmax of (8), and the system collapses to visual-only inference without an architectural switch. *Third*, when both signals are weak, $z_t \to z_c$ and z_c is invariant under the argmax, so $p_{\max} \to 1/K$:the signal an upstream consumer needs to drop the prediction. The architecture transitions monotonically from *abstain* to *high-confidence multimodal classification* via a single sigmoid scalar, addressing the workshop's robustness-under-modality-dropout topic analytically rather than by an engineering switch.

6.2 Conjunction analysis: why no pair of moves suffices

The three moves are mechanistically complementary. The σ-aware ROI policy supplies the analytical regime characterisation through Eq. (5): without it, deployment offers no analytical handle on the gaze-noise / object-scale trade-off and reduces to the MCUNet single-modality regime [12,11]. The offline prototypes supply the textual stream within the budget: without them, a runtime BERT-class encoder exceeds the SRAM ceiling by an order of magnitude (§4.4) or the textual stream is dropped, forfeiting the graceful-degradation property of §6.1. The ISP path supplies the memory saving on which the rest of the pipeline depends: without it, the MobileNetV3-class encoder cannot fit alongside the 2.7 MB sensor-frame buffer. The novelty is therefore the envelope-binding conjunction, not any single move.

6.3 Limitations and threats to validity

Five assumptions bound the analysis, each tested by F1–F5 or acknowledged as residual. *(i)* Gaze noise is zero-mean isotropic Gaussian; real trackers exceed this under blinks, drift, and head motion [1], so (4) is best-case (F4 spans anisotropic and Student-t Monte-Carlo regimes; the synthetic Gaussian σ-injection of §7.3 does not cover blink/drift dropouts, only the real GOO/GazeFollow subset does). *(ii)* $\hat{s}_o$ is the median object scale, so $\sigma_{\text{cross}} \approx 25\,\text{px}$ is an order-of-magnitude statement; the COCO taxonomy is itself a stand-in for assistive/wearable use cases, with residual class-coverage error not directly tested by F1–F5. *(iii)* The multimodal gain over visual-only is the subject of F3, not asserted analytically. *(iv)* The ISP path argument cites the STM32N657X0 product description [19] and is re-verified at camera-ready. *(v)* The scalar q_g conflates covariance with blink-dropout invalidity. We do not claim measured silicon performance, multi-object inference, or open-set generalisation; the scope is single-object inference under the closed taxonomy of §4.4 and the deployment envelope of §5.

7 Empirical Protocol & Future Work

The contribution of this paper is an architectural commitment paired with a *pre-registered* empirical protocol. Treating the protocol as a first-class object lets us

answer three standard objections to non-empirical edge-AI proposals (absence of measured validation, novelty differentiation against gaze-guided prior art, and under-specification of ROI / gaze-noise / fusion mechanics) by binding each to a quantitative refutation criterion, so that follow-on work can cite results that *will* falsify or confirm the architecture rather than merely describe it.

7.1 Pre-registered hypotheses

The protocol is organised around the five hypotheses summarised in Table 3. F1 demands that gaze conditioning beat both a saliency-only baseline and the trivial centre-bias floor; the 8 pp gap at $\sigma_g = 30\,\text{px}$ ($\approx 1.4^\circ$ at $21\,\text{px}/^\circ$) is the smallest gap that survives an F-test under per-class accuracy variance reported for natural-scene viewing [15,6]. F2 separates the σ-conditional spatial switch from the saccade-onset temporal skip into two independent sub-criteria, each individually sufficient to refute. F3 is the central claim of §§4.4–4.5: the offline prototype table preserves accuracy against a TinyBERT-4 INT8 host-emulation comparator (the structurally infeasible runtime alternative of §4.4) within 2 pp top-1 at matched prompt set, and deployed-table size measured on-silicon stays within tolerance. F4 is a Monte-Carlo validation of (5) under three regimes: isotropic Gaussian, anisotropic Gaussian with $\sigma_x \neq \sigma_y$, and Student-t with $\nu = 3$ for saccade-onset jitter; the relative-error tolerance is 0.15, justified by the sharpness of the corrected closed form. F5 is a non-emptiness existence claim on the deployable Pareto frontier: at least one P-variant must satisfy $T_{\text{e2e}} \leq 33\,\text{ms}$ and SRAM $\leq 4.2\,\text{MB}$ simultaneously on STM32N6 silicon. The full frontier (including non-deployable variants) is reported regardless.

7.2 Design space and acceptance envelope

Seven configurations span the design space. **B0**: full-frame MobileNetV3-Small INT8, no gaze (no-attention baseline). **B1**: B0 plus a soft attention map (full-frame compute). **B2**: centre-crop classifier, no gaze (isolates gaze-conditioned cropping versus mere cropping). **B3**: full-frame backbone with gaze as a post-hoc re-ranking signal over dense outputs (the comparator pattern §2 differentiates against). **P1**: proposed gaze-cropped visual-only stream at $r \in \{64, 96, 128\}$. **P2**: P1 plus the class-prototype backstop, the full proposed system. **P3**: oracle variant of P2 with ground-truth gaze, upper-bounding performance under perfect attention. The acceptance envelope is binding: configurations outside the deployable region ($T_{\text{e2e}} \leq 33\,\text{ms}$ and peak SRAM $\leq 4.2\,\text{MB}$) are retained for analysis but flagged non-deployable. The on-silicon campaign on the STM32N657X0 evaluation board uses 32 warm-up frames and software-triggered current sampling, and is the subject of follow-on work.

7.3 Datasets and synthetic benchmark

The primary surface combines a real gaze-labelled subset (GOO [21] and GazeFollow [16], downsampled to VGA with subject-disjoint splits) with a procedurally generated synthetic benchmark addressing three gaps: controlled gaze-noise

Table 3. Pre-registered hypotheses, refutation criteria, and the standard objection class each addresses. A failed criterion is reported as a *null* outcome rather than absorbed into a sweep average.

Tag	Claim	Refutation criterion	Objection addressed
F1	Gaze-anchored ROI beats saliency-only and the centre-bias floor	IoU margin < 0.05 at any σ_g; OR gap to centre-bias < 8 pp at σ_g=30 px	Novelty vs. gaze-guided prior art
F2a	σ-conditional spatial switch dominates fixed-r policies	ΔAcc < 2 pp at every σ_g vs. fixed r=96 baseline	ROI sizing
F2b	Saccade-onset temporal skip reduces compute without accuracy loss	Skip reduction $< 15\%$ at matched accuracy, OR accuracy drop > 3 pp at any skip rate	Temporal robustness
F3	Offline class-prototype table preserves accuracy vs. TinyBERT-4 INT8 host emulation [5]	Top-1 drop > 2 pp at matched prompt set; OR fusion latency > 1 ms on-silicon	Embedded fusion feasibility
F4	Empirical σ-crossing matches Eq. (5) within tolerance	$\lvert\hat{\sigma}_{\text{cross}} - \sigma_{\text{cross}}\rvert/\sigma_{\text{cross}} > 0.15$ in Monte-Carlo (isotropic, anisotropic, Student-t ν=3)	Analytical correctness
F5	Existence of a deployable P-variant under $T_{\text{e2e}} \leq 33$ ms and SRAM ≤ 4.2 MB	No P-variant satisfies both constraints simultaneously on STM32N6 silicon	Frontier honesty

injection at $\sigma \in \{0, 1, 2, 5\}^\circ \approx \{0, 21, 42, 105\}$ px at the 21 px/° conversion of §4.2 (so the σ-crossing of Eq. (5) is bracketed:$\sigma = 0$ is the oracle, $\sigma = 21$ px is below the predicted 25 px crossing, $\sigma = 42$ px is above, $\sigma = 105$ px is well above); controlled clutter density (separating regimes where the prototype backstop helps from those where it is neutral); and scene-template-disjoint splits (preventing background-statistic leakage). Splits, INT8 calibration sets, and procedural seeds are released alongside the code.

7.4 Roadmap to silicon and engineering attribution

Eq. (5) predicts that a 2σ-budget system on ~ 100 px objects switches regime at σ of order tens of pixels (F4 tests this under isotropic, anisotropic, and Student-t regimes). Three engineering risks separate silicon-level failures from architectural refutation, each pinned to a measurable signal: operator coverage on the Neural-ART NPU (X-CUBE-AI per-layer profiler, pass criterion: zero Cortex-M55 fallback); INT8 calibration drift host-vs-silicon (per-class top-1 delta at matched calibration set, pass: ≤ 1 pp); ISP/NPU DMA contention (per-frame latency 99th-percentile under sustained 30 Hz, instrumented per [10], pass: ≤ 33 ms).

8 Conclusion

Treating gaze as a computational attentive prior rather than a downstream reranking signal makes the perception problem smaller before the model is com-

pressed. The proposed framework turns this commitment into a deployable architecture for the STM32N6 through three coupled design moves: a σ-aware ROI policy with the corrected closed-form crossing of Eq. (5); an offline INT8 class-prototype backstop that fits on-chip SRAM by three orders of magnitude over a runtime text encoder; and an ISP→DMA→NPU data path (Fig. 1) that eliminates the sensor frame buffer. The fusion gate's three canonical limits realise robustness under modality dropout as a closed-form analytic property. The five falsifiable hypotheses (Section 7) bind each architectural claim to a quantitative refutation criterion, and the on-silicon execution of the protocol on the STM32N657X0 is the natural next step.

References

1. Barz, M., Daiber, F., Sonntag, D., Bulling, A.: Error-aware gaze-based interfaces for robust mobile gaze interaction. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications (ETRA '18). ACM (2018). https://doi.org/10.1145/3204493.3204536
2. Boujut, H., Benois-Pineau, J., Megret, R.: Saliency-driven object recognition in egocentric videos with deep CNN: Toward application in assistance to neuroprostheses. Computer Vision and Image Understanding **164**, 82–91 (2017). https://doi.org/10.1016/j.cviu.2017.03.001
3. David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Wang, T., Warden, P., Rhodes, R.: TensorFlow Lite Micro: Embedded machine learning on TinyML systems. In: Proceedings of Machine Learning and Systems (MLSys). vol. 3 (2021)
4. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019)
5. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. In: Findings of the association for computational linguistics: EMNLP 2020. pp. 4163–4174 (2020)
6. Khamis, M., Alt, F., Bulling, A.: The past, present, and future of gaze-enabled handheld mobile devices: Survey and lessons learned. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services. pp. 1–17 (2018)
7. Kim, J.B., Cho, H.i.: Human gaze-aware attentive object detection for ambient intelligence. Engineering Applications of Artificial Intelligence **106**, 104471 (2021). https://doi.org/10.1016/j.engappai.2021.104471
8. Lai, L., Suda, N., Chandra, V.: CMSIS-NN: Efficient neural network kernels for arm Cortex-M CPUs. arXiv preprint (2018)
9. Lall, V., Liu, Y.: Eyes on target: Gaze-aware object detection in egocentric video. arXiv preprint arXiv:2511.01237 (2025)
10. Li, R.: Dataflow and tiling strategies in edge-AI FPGA accelerators: A comprehensive literature review. arXiv preprint arXiv:2505.08992 (2025), https://arxiv.org/abs/2505.08992
11. Lin, J., Chen, W.M., Cai, H., Gan, C., Han, S.: MCUNetV2: Memory-efficient patch-based inference for tiny deep learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34 (2021)

12. Lin, J., Chen, W.M., Lin, Y., Cohn, J., Gan, C., Han, S.: MCUNet: Tiny deep learning on IoT devices. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 11711–11722 (2020)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021)
15. Rayner, K.: Eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology **62**(8), 1457–1506 (2009). https://doi.org/10.1080/17470210902816461
16. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems (NeurIPS) (2015)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
18. Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., Rehg, J.M.: Gaze-LLE: Gaze target estimation via large-scale learned encoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28874–28884 (2025), highlight
19. STMicroelectronics: STM32N657X0: High-performance microcontroller with Neural-ART accelerator. https://www.st.com/en/microcontrollers-microprocessors/stm32n657x0.html (2024), product page; specifications cross-referenced with the STM32N657X0 datasheet
20. Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., Zhou, D.: MobileBERT: A compact task-agnostic BERT for resource-limited devices. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 2158–2170 (2020). https://doi.org/10.18653/v1/2020.acl-main.195
21. Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., Guinto, R.: GOO: A dataset for gaze object prediction in retail environments. In: IEEE/CVF CVPR Workshops (GAZE) (2021)
22. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21860–21869 (2023). https://doi.org/10.1109/ICCV51070.2023.02000
23. Toyama, T., Kieninger, T., Shafait, F., Dengel, A.: Gaze guided object recognition using a head-mounted eye tracker. In: Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12). pp. 91–98. ACM (2012). https://doi.org/10.1145/2168556.2168570
24. Wang, B., Guo, C., Jin, Y., Xia, H., Liu, N.: TransGOP: Transformer-based gaze object prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10180–10188 (2024). https://doi.org/10.1609/aaai.v38i9.28883