

Unlocking Your Potential: The Proof is in the Results

Danilo Pau

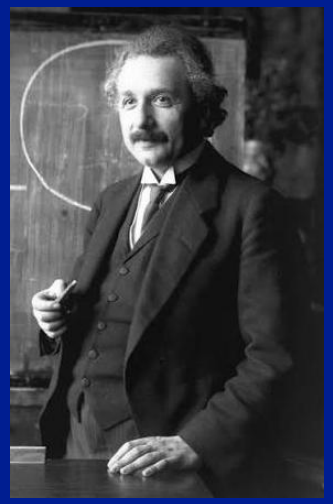
Technical Director, System Research and Applications

IEEE, AAIA and ST Fellow

Agrate Brianza

Students can achieve great things without realizing how much latent power they truly possess





"If we did all the things we are capable of, we would literally astound ourselves.“

(Often also attributed to Albert Einstein in slightly different variations)



"Our deepest fear is not that we are inadequate. Our deepest fear is that we are powerful beyond measure. It is our light, not our darkness, that most frightens us. We ask ourselves, who am I to be brilliant, beautiful, talented, and fabulous? Actually, who are you not to be?"

Niks Kordjukovs



symmetry

MDPI

Article

Complexity Analysis for Categorized Edge Language Models

Niks Kordjukovs and Danilo Pietro Pau *

System Research and Applications, STMicroelectronics, 20864 Agrate Brianza, Italy; niks.kordjukovs@st.com
* Correspondence: danilo.pau@st.com

Abstract

Edge generative artificial intelligence (AI) increasingly combines language, perception, reasoning, audio, and action on resource-constrained devices. This paper profiles public GPT-Generated Unified Format (GGUF) checkpoints from the Hugging Face Hub (HFH) across conversational, instruct, thinking, audio, vision-language (VL), and vision-language-action (VLA) categories using a shared parser-based deployment-envelope workflow. The main category-specific run retained 21,039 profiled entries and estimated the minimum memory bandwidth, compute throughput, and unified-memory architecture (UMA) footprint needed to satisfy category-specific target throughput values. The resulting measurement protocol was symmetric, but the deployment envelopes were asymmetric: VL and thinking workloads were the heaviest on the compute–bandwidth axis, VLA formed a smaller elevated multimodal branch, and audio, instruct, and conversational workloads were lighter on average. A unified 10-tokens-per-second (TPS) sensitivity run compressed the compute–bandwidth gaps, showing that service-rate assumptions contributed strongly to cross-category separation. Welch/Games–Howell and Kruskal/Dunn analyses confirmed large category effects for bandwidth and compute in the category-specific regime, but only small memory effects. The results show that edge-model feasibility cannot be inferred from parameter count alone; throughput target, backbone family, modality, and memory budgeting must be considered jointly.

Keywords: symmetry/asymmetry; edge AI; GGUF; memory bandwidth; multimodal inference; workload complexity; UMA; backbone heterogeneity

1. Introduction

Edge AI has moved from primarily text-centric assistants toward heterogeneous workloads that combine language, perception, and action, especially on mobile, embedded, and robotic platforms [1–3]. On such platforms, deployment feasibility is constrained not only by model size, but also by memory bandwidth, compute throughput, and available memory [4–6]. In that context, downstream capability alone is insufficient to characterize feasibility, and deployment-oriented complexity profiling is needed alongside task accuracy. The target deployment abstraction adopted here was a simple UMA, namely a central compute element connected to RAM through a shared bus.

The motivation has become more pressing as recent VL and VLA systems such as Helix, OpenVLA, and SmolVLA have shown that multimodal perception, language understanding, and action generation are increasingly being pushed toward generalist embodied operation [7–9]. That trend increased the importance of format-level compatibility, parser-level observability, and hardware-envelope estimation for models intended to run at the edge.

Check for updates

Received: 21 March 2026
Revised: 23 April 2026
Accepted: 25 April 2026
Published: 29 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

computers

MDPI

Article

VibroAgent: Multiresolution Soft-Token Fusion with Gemma 4 for Reference-Conditioned Structural Vibration Monitoring

Niks Kordjukovs [1], Alfredo Cigada [2] and Danilo Pietro Pau [1,*]

[1] System Research and Applications, STMicroelectronics S.r.l., 20864 Agrate Brianza, Italy; niks.kordjukovs@st.com
[2] Dipartimento di Meccanica, Politecnico di Milano, 20156 Milan, Italy; alfredo.cigada@polimi.it
* Correspondence: danilo.pau@st.com

Abstract

VibroAgent represented ten-second acceleration windows from one reference and five monitored locations as 84 continuous soft tokens for Gemma 4 E2B, letting one model decide and potentially explain; only its decisions were evaluated. Tasks were global healthy/changed classification and affected-zone flags for released or loosened connections on a lattice mast (LUMO), grandstand (QUGS), and scaled offshore jacket (UPC). Against tuned comparators on the paper's 984-window split, published Gemma had the highest overall global accuracy (0.9787) but pooled affected-target F1 0.1802. Its target predictions tracked slot position. An Abdeljaber-style CNN adaptation had higher pooled target F1 (0.3107) but LUMO global accuracy below 0.47. In an even three-seed comparison (equal data, pretraining, tuning, selection) on 1483 held-out windows, mean LUMO/QUGS target F1 was 0.290/0.407 (binary Gemma) and 0.623/0.813 (CNN); finer supervision gave no benefit. With more data and healthy/changed-balanced sampling for all methods, binary Gemma matched the CNN in healthy/changed detection on 792 never-used damaged and 72 healthy UPC windows (balanced accuracy 0.949 versus 0.935 under equal loss weighting; paired difference 0.014 ± 0.030); LUMO/QUGS target F1 did not improve. Results established neither localization advantage, unseen-structure generalization, nor field validity. Six STWIN.box nodes on a Qualcomm IQ-9075 EVK demonstrated local integration only.

Keywords: structural health monitoring; vibration monitoring; multiview representation learning; soft tokens; Gemma 4; edge AI; dataset ablation; reference-conditioned sensing; partial-label learning

Check for updates

Academic Editor: Xueyi Li

Received: 22 August 2026
Revised: 15 October 2026
Accepted: 17 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Vibration-based structural health monitoring (SHM) uses changes in measured dynamic response as evidence that can warrant inspection. Such a change alone does not establish a damage mechanism: excitation, temperature, operating state, mounting, timing, and sensor quality can generate similar changes [1,2]. Monitoring systems should consequently retain the measurements, document unavailable information, and limit their outputs to observations supported by that evidence.

Data-driven SHM has moved from hand-crafted damage-sensitive features with statistical classifiers to deep networks that learn features directly from the measured response [3–5]. When examples of damage are unavailable, novelty detection against a

Niks Kordjukovs



DEEP GENERATIVE MODELING: A COMPREHENSIVE REVIEW AND A COMPARATIVE STUDY OF GANS, VAES AND DIFFUSION MODELS

Danilo Pietro Pau[1], Niks Kordjukovs[2], and Claudio Turchetti[3]

[1]System Research and Applications, STMicroelectronics, I-20864 Agrate Brianza, Italy; danilo.pau@st.com
[2]System Research and Applications, STMicroelectronics, I-20864 Agrate Brianza, Italy; niks.kordjukovs@st.com
[3]Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 12, I-60131 Ancona, Italy; c.turchetti@univpm.it

ABSTRACT

Deep Generative Models (DGMs) aim to learn complex data distributions and generate realistic samples across modalities such as images, audio, and text. Among the most prominent approaches are Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DMs), each offering distinct trade-offs in quality, stability, and interpretability. This chapter presents a comprehensive overview and comparative analysis of these three model families, evaluating performance in terms of fidelity, diversity, training stability, and computational efficiency. It further examines practical deployment aspects, including cost–performance trade-offs, GPU requirements, and inference latency. Special attention is given to generative modeling at the edge, where constraints on memory, compute, and energy demand efficient architectures. Finally, the work analyzes real-world model formats (GGUF and safetensors) under standardized hardware assumptions, providing insights into scalability, quantization, and deployment feasibility.

Keywords Deep neural network (DNN), Deep generative models (DGM), Variational autoencoder (VAE), Generative adversarial network (GAN), Diffusion model (DM), Denoising diffusion probabilistic model (DDPM), Noise conditioned score network (NCSN), Stochastic differential equation (SDE), Edge-oriented generative architecture (EOGA).

1 Introduction

Deep generative models (DGMs) aim to learn complex data distributions from finite samples and to synthesize novel, realistic instances from these learned distributions [1–5].Formally, we have data points x (e.g. images) drawn from an unknown distribution $p_{data}(x)$ and a generative model that tries to learn a distribution $p_\theta(x)$ (parametrized by a neural network with parameters θ) which

A Heterogeneous Edge-to-Actuator Architecture for Real-Time Visual Servoing Coupled with Hardware-Level Field-Oriented Control

Giulio Tamburini[1], Niks Kordjukovs[2], Marco Marcon[3], Danilo Pietro Pau[4*]

[1, 2, 4] System Research and Applications, STMicroelectronics, Agrate Brianza, Italy .
[3] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy .

*Corresponding author(s). E-mail(s): danilo.pau@st.com;

Abstract

This paper investigated how a lightweight vision-language model can drive the field-oriented control of an electric motor, derating its speed when a human operator approaches by predicting short-term speed commands from camera frames. A frozen SmolVLM2-256M backbone, run as a multimodal encoder through `llama.cpp` without autoregressive generation, fed a task-specific action head; single-layer, multilayer-perceptron, temporal convolutional network, and linear-attention heads were compared. Evaluation used 64 purpose-built video episodes (79,393 frames), episode-level splits, a four-fold rotation, and the commands played at 75 Hz after the 205 ms latency. On eight held-out episodes, the deployed single-frame linear-attention head reached a mean absolute error of 0.062 (95% confidence interval 0.049–0.078) and an R^2 of 0.912. Command-history feedback degraded closed-loop accuracy in 24 of 25 configurations, the instruction tokens carried no measurable information, and a frozen MobileNetV3-Small encoder reached 0.056 at one tenth (CPU) or one hundredth (NPU) of the encoder latency. The INT8 heads were profiled on STM32 microcontrollers, and an Arduino VENTUNO Q ran the encoder on its Qualcomm NPU and the head on its STM32H5. In live motor trials, the played commands tracked the target with an error of 0.056, and an interrupted head link stopped the motor within 0.44 s. SmolVLM2 feature extraction dominated the latency (50 ms per frame on a Jetson Orin Nano, 132 ms on the VENTUNO Q), whereas MobileNetV3-Small ran at the 30 Hz camera rate. The results come from one laboratory scene with

Niks Kordjukovs



Focalized Thinking

FOCALISER VERSION | 0.8B | 2B | Run Both | Run Baseline | Run Focalized

Show Both | Show Baseline | Show Focalized | Shared | Vanilla Qwen | Traces

VERBOSITY SELECTOR | Level 1 | Level 2 | FOCALIZER REASONER | Preload

Ask a question

Send | Stop | Clear

MODEL RUNTIME

FOCALIZER	REASONER	GPU / RUNTIME
Qwen3.5-2B	**Qwen3.5-4B**	**NVIDIA GeForce RTX 5070 11.9 GB**
loaded on cuda:0; requested cuda	loaded on cuda:0; requested cuda	device cuda; dtype float16; device_map none; 4-bit

Baseline — Done first

PROMPT

Think of a well-known U.S. city that, when spelled out, contains exactly 13 letters, and every single letter must be different

RESPONSE

Thinking Process:

1. **Analyze the Request:**
   * Target: A well-known U.S. city.
   * Constraint 1: When spelled out, it must contain exactly 13 letters.
   * Constraint 2: Every single letter must be different (no repeated letters).

2. **Brainstorm U.S. Cities:**
   * New York (8)
   * Los Angeles (10)
   * Chicago (7)
   * Houston (6)

Overthinking, Confabulating

Focalized — Done seco

PROMPT

Think of a well-known U.S. city that, when spelled out, contains exactly 13 letters, and every single letter must be different

RESPONSE

San Francisco

DONE SECOND — 100%

-86% **-90%**

METRICS

MODEL	APPLICATION TIME	REASONER INPUT TOKENS	REASONER OUTPUT TOKENS
Baseline	155.2 s	29	4,036
Focalized	21.8 s	29	387

Niks Kordjukovs





WORKSHOP: When Buildings Speak in Codes

Neural Codec & On Device Qwen3 for Vibration Diagnostics

Hosts: N. Kordjukovs & D. Pau (STMicroelectronics), A. Nagendra (Qualcomm)

START TIME: 10:45AM PST

AUDIENCE: Edge AI developers and engineers building private, sensor driven diagnostics without sending raw data off device.

GOALS:

1. PIPELINE: Learn how the codes_v3 pipeline converts 10 second sensor windows into 250 single token symbols per sensor via a residual VQ codec.
2. COMPARISON: Compare 1 reference STWIN.box (STEVAL-STWINBX1) against 5 targets using codec symbols and broadband levels.
3. SCENARIOS: Explore live 6 board USB monitoring (running Qwen3-4B on the IQ-9075 NPU), immutable recorded replay, and operator triage.
4. PRODUCTION: Keep raw data local, hash verify files, apply strict schema validation, and use fail closed error handling.

PREPARATION REQUIRED: None mandatory. Live replay requires: Qualcomm IQ-9075 EVK, 6 STEVAL-STWINBX1 boards, a USB hub, and the provided repo.

When Buildings Speak about their Vibrations



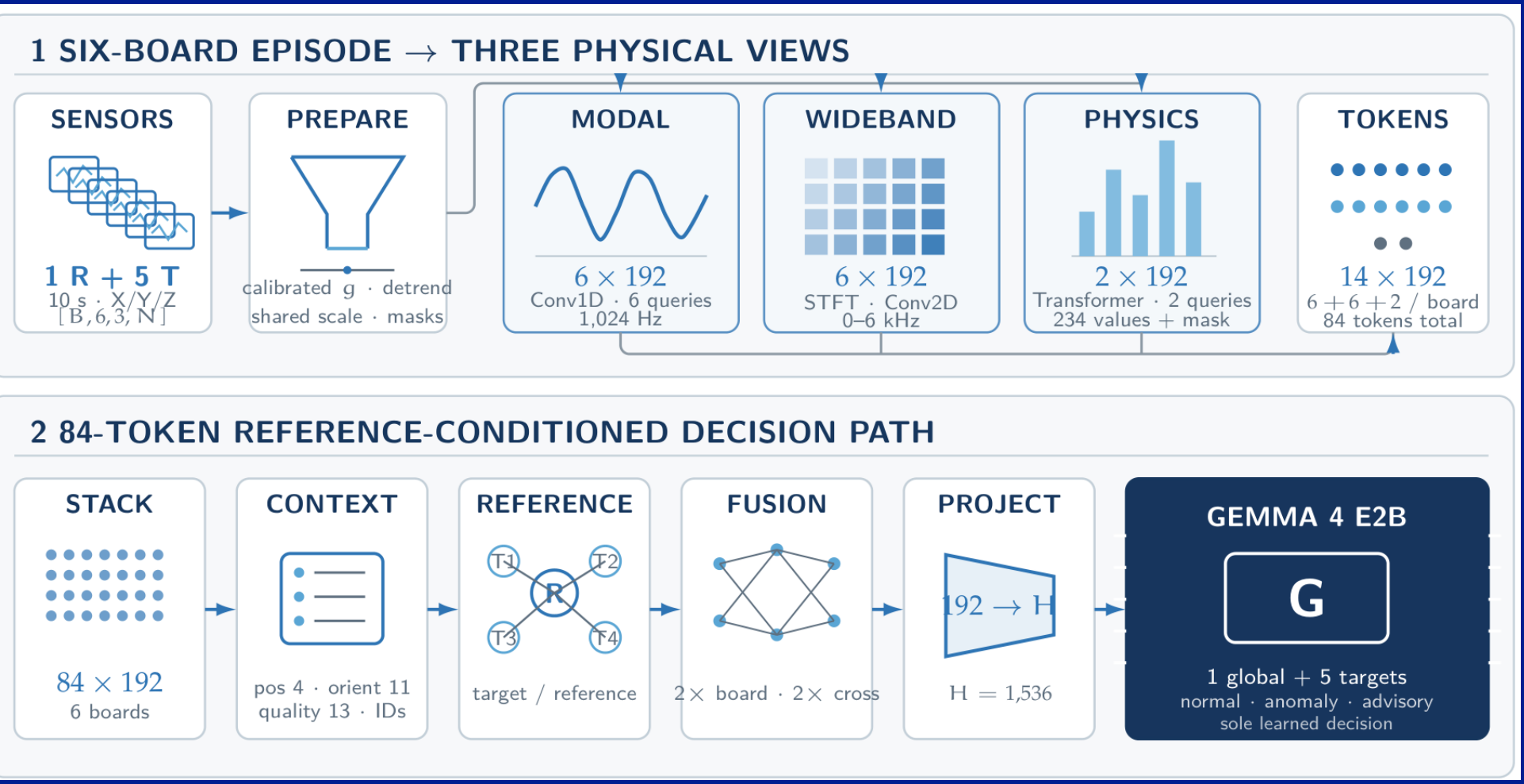




POLITECNICO MILANO 1863

DIPARTIMENTO DI MECCANICA

Three complementary representations: one edge language model



THREE VIEWS

1. Modal low-freq motion
2. Time-freq content
3. Physics features

84

learned vectors per episode

Structured physical evidence → projection → Gemma global and target decisions.

Luigi Zappia



electronics

MDPI

Article

Heterogeneous Low-Power Agentic Architecture for Edge Distributed Home Surveillance

Luigi Lucio Zappia, Marco Lattuada, Marco Bianchessi and Danilo Pietro Pau *

System Research and Applications, STMicroelectronics, 20864 Agrate Brianza, Italy; luigilucio.zappia@st.com (L.L.Z.); marco.lattuada@st.com (M.L.); marco.bianchessi@st.com (M.B.)
* Correspondence: danilo.pau@st.com

Abstract

Recent advancements in edge hardware enabled increasingly complex artificial intelligence workloads to be executed directly on resource-constrained devices. However, single-edge devices remain computationally limited compared with hybrid or cloud alternatives. This work presents a fully edge-based architecture implementing agentic AI on heterogeneous devices for home surveillance with a natural language interface, in which analysis tasks are parallelized across the distributed devices. The system is tested on a two-node prototype: a Raspberry Pi 5 with a Hailo-10H accelerator and an STM32MP257FDK board. Dynamic leader election allows recovery from the loss of any single node, including the leader, with graceful degradation. Qwen3-1.7B achieved 93% tool-selection accuracy over 24 tools and maintained robust summarization performance, outperforming FunctionGemma-270M. The jina-reranker-v2-base-multilingual cross-encoder performs comparably to the tested LLMs at tool selection on small pool sizes (i.e., with few tools to choose from). Hardware accelerators improved inference energy efficiency by 2.0× to 17.3×, while leader execution incurred less than 2 W of additional power. The main bottleneck is model latency rather than network overhead, and KV-cache reuse proved essential to meet the latency constraint on the STM32MP257FDK. The proposed prototype coordinates inference across heterogeneous edge devices to answer queries, with the weaker node exhibiting scalability limitations.

Keywords: distributed networking; edge artificial intelligence; tiny machine learning; edge language models; agentic frameworks; microprocessors; low power

Check for updates

Academic Editor: Fernando De la Prieta Pintado

Received: 10 June 2026
Revised: 15 July 2026
Accepted: 17 July 2026
Published: 19 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In recent years, the adoption of surveillance technologies has accelerated. The market reflects this trend: the global video surveillance market was valued at USD 73.75 billion in 2024 and is projected to reach USD 147.66 billion by 2030 with a compound annual growth rate (CAGR) of 12.1% between 2025 and 2030 [1] (Refs. [1,2] are market-research reports (Grand View Research and Mordor Intelligence), while refs. [3,4] are vendor reports published by physical-security technology companies (Genetec and Axis Communications). They are not peer-reviewed studies and are cited here solely to provide market and industry context). Advancements in technology and rising security concerns across multiple sectors are key drivers of this growth [1,2].

According to an industry survey by security firm Genetec, as of early 2026, an increasing number of surveillance systems rely on hybrid (using both edge and cloud hardware) or cloud-based approaches [3] that may use company servers or serverless architectures; for

Heterogeneous Low-Power Agentic Architecture for Edge Distributed Home Surveillance

- Video Demo

Luigi Lucio Zappia · Marco Lattuada · Marco Bianchessi · Danilo Pietro Pau

System Research and Applications, STMicroelectronics - Agrate Brianza, Italy

Electronics 2026, 15(14), 3175 · doi:10.3390/electronics15143175

Date: 20/09/2026

Andrea Spagnolo



electronics

MDPI

Article

Edge-Constrained Food Retrieval System

Andrea Spagnolo [1], Simone Bianco [1], Raimondo Schettini [1] and Danilo Pau [2,*]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy; andreaspagnolo99@gmail.com (A.S.); simone.bianco@unimib.it (S.B.); raimondo.schettini@unimib.it (R.S.)
[2] STMicroelectronics, 20864 Agrate Brianza, Italy
* Correspondence: danilo.pau@st.com

Abstract

Fine-grained food image and text retrieval requires models that can distinguish subtle differences in ingredients, preparation methods, and visual appearance. While vision language models provide strong multimodal reasoning capabilities, their computational requirements can make direct deployment on edge devices impractical. This paper proposes a training free retrieval framework that combines efficient candidate retrieval with selective vision language reasoning. For ingredient retrieval, Qwen3-VL performs closed choice selection over the retrieved candidates, while dense caption retrieval uses Qwen3-VL-Reranker as a second stage multimodal reranker. An efficiency oriented configuration, Stage 4-E, further introduces confidence based execution, allowing the VLM components to be skipped when the initial retrieval is sufficiently confident. The proposed approach is evaluated on the Dishcovery challenge and on auxiliary food retrieval datasets, with particular attention to the tradeoff between retrieval quality and computational cost. The final pipeline is also deployed on two different edge AI platforms, the NVIDIA Jetson AGX Orin and the Qualcomm Dragonwing IQ-9075 EVK, using platform specific model representations and execution backends. The results show that selective multimodal reasoning, confidence based execution, and hardware aware optimization can substantially reduce inference cost while preserving most of the retrieval performance, making the proposed architecture suitable for practical edge deployment.

Keywords: food image–text retrieval; edge AI; vision–language models; training-free retrieval; SigLIP2; Qwen3-VL; multimodal reranking

Check for updates

Academic Editors: Galina Ilieva, George A. Tsihrintzis and Maria Virvou

Received: 18 September 2026
Revised: 27 October 2026
Accepted: 29 October 2026
Published: 1 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Fine-grained food image and text retrieval requires matching food images with structured ingredient labels or rich textual descriptions. Although recent vision language models have significantly improved cross modal retrieval, food remains a difficult domain because many dishes share similar visual appearances while differing in small but important details. Retrieving the correct ingredient set or dense textual description therefore requires models to capture subtle visual and semantic cues.

The CVPR Dishcovery VLM MetaFood Challenge [1] provides a realistic benchmark for this problem by evaluating fine grained food retrieval across two complementary tasks: multi label ingredient retrieval and single label dense caption retrieval. High performing multimodal systems often rely on large VLMs [2], model ensembles, or expensive inference pipelines, which can be difficult to deploy when memory, latency, and power consumption are constrained. This is particularly relevant for edge AI platforms, where inference must be adapted to the available compute budget rather than assuming server grade hardware.

Andrea Spagnolo



EDGEAI FOUNDATION LEARN EVENTS LIVESTREAMS PARTNERS SOLUTIONS ABOUT

THREE NEW WORKING GROUPS AND STRATEGIC PARTNERSHIP WITH LF EDGE READ TH

WORKSHOP: Training Free Food Understanding with Vision Language Models at the Edge

START TIME: 9:30AM PST

Hosted by D. Pau and A. Spagnolo (STMicroelectronics) and A. Nagendra (Qualcomm).

TARGET AUDIENCE: Edge AI developers, computer vision engineers, embedded developers, and interested in running multimodal AI pipelines locally on constrained hardware.

GOALS: By the end of the session, attendees will understand:

1. How a training free food understanding pipeline combines image text retrieval and vision language models without task specific fine tuning.
2. How image text retrieval reduces large ingredient and caption vocabularies to a small candidate set before invoking a more expensive VLM.
3. How confidence based gates skip unnecessary model calls to reduce latency while preserving prediction quality.
4. How this architecture supports three concrete tasks:
    1. Ingredient recognition – retrieve likely ingredients and refine the final selection with Qwen3-VL-4B-Instruct.
    2. Dish caption retrieval – retrieve candidate descriptions and rerank them with Qwen3-VL-Reranker-2B.
    3. Calorie estimation – estimate ingredient quantities or portion sizes and combine them with nutritional information.
5. How the full pipeline was optimized and deployed locally on the Qualcomm Dragonwing IQ-9075 EVK.

PREPARATION REQUIRED:

1. Nothing mandatory to follow along.
2. To reproduce the demo: a Qualcomm Dragonwing IQ-9075 EVK, plus the project repositories and setup instructions shared with attendees.

DURATION: 1 hour (technical presentation + live on device demo)

Gen EdgeAI helps mindful eating

Input and Resources

Food image query and precomputed databases

Query image



Ingredient vocabulary
(text embeddings precomputed)

Caption database
(text embeddings precomputed)

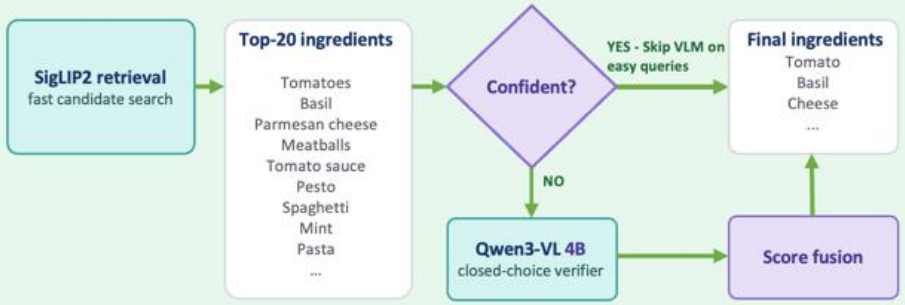
Text embeddings cached offline

Stage 4-E: Edge-Optimized Dual-Task Pipeline

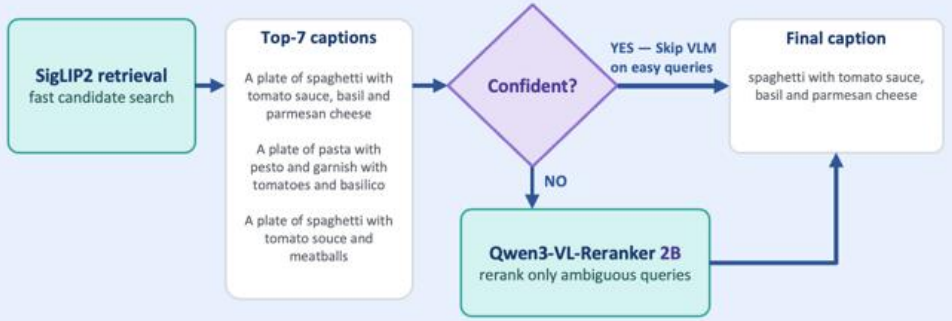
Selective multimodal reasoning only when needed

Confidence-gated execution

Task 1: Ingredient Retrieval



Task 2: Dense Caption Retrieval



Edge Deployment and Results

Same pipeline, optimized for different platforms

Hardware-portable deployment
with platform-specific model representations and runtimes

Jetson AGX Orin 64 GB



Dragonwing IQ-9075 EVK

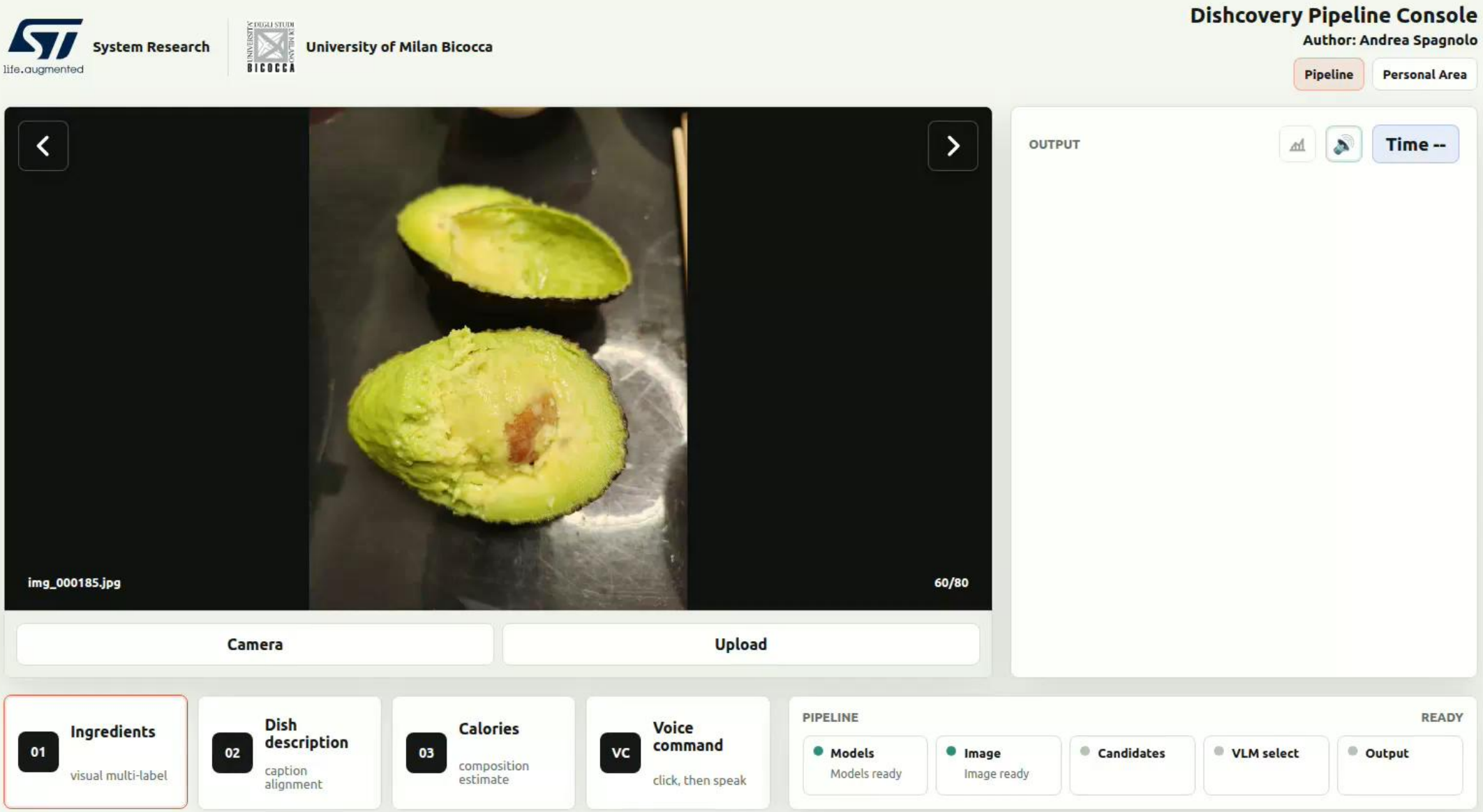


Jetson AGX Orin 64 GB	Dragonwing IQ-9075 EVK
Task 1 (Ingredients)	Task 1 (Ingredients)
F1 0.773 | 0.923 s	F1 0.732 | 2.795 s
Task 2 (Captions)	Task 2 (Captions)
Top-1 0.7004 | 1.311 s	Top-1 0.694 | 9.108 s

Hardware-aware optimization preserves retrieval quality while reducing inference cost

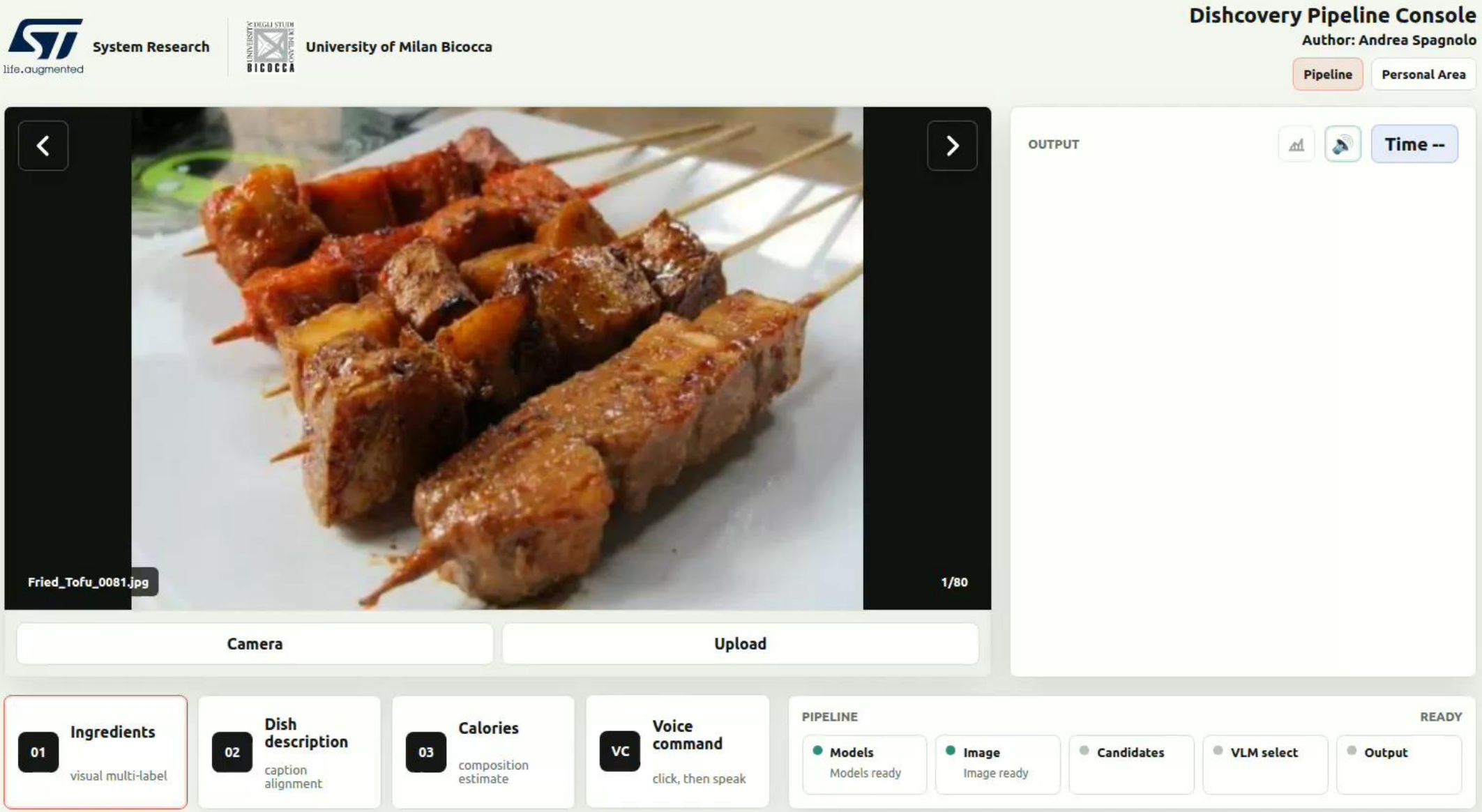
Frozen pretrained models – no FineTuning

Ingredient Recognition Pipeline (Task 1)



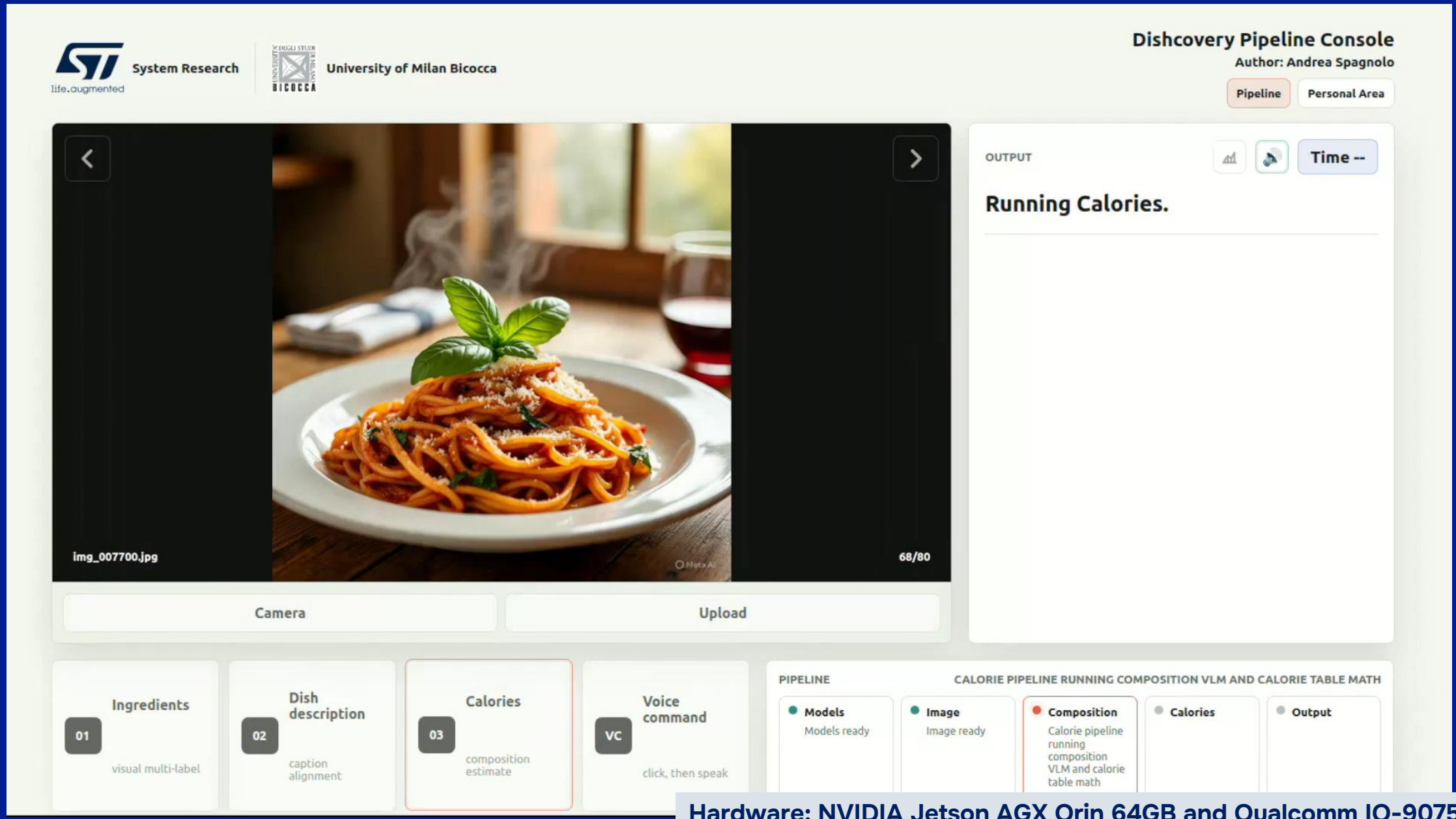
Hardware: NVIDIA Jetson AGX Orin 64GB and Qualcomm IQ-9075

Caption Retrieval & Reranking Pipeline (Task 2)



Hardware: NVIDIA Jetson AGX Orin 64GB and Qualcomm IQ-9075

Ingredient Recognition & Calorie Estimation Pipeline



Hardware: NVIDIA Jetson AGX Orin 64GB and Qualcomm IQ-9075

Luigi Mazzone

applied sciences

MDPI

Article

Performance Analysis of a Modular Framework for Edge-Based Generative Conversational AI

Lorenzo Mazzone and Danilo Pau *

System Research, STMicroelectronics, 20864 Agrate Brianza, Italy; mazzonelorenzo1@gmail.com
* Correspondence: danilo.pau@st.com

Abstract

This study presents a multi-tier framework for deploying multi-modal Conversational AI on edge devices, spanning from constrained ultra-low-power systems to high-performance edge workstations. Utilizing an automated model discovery process and a modular benchmarking testbed, the research demonstrates that real-time, fully edge AI execution is feasible through strategic model selection and hardware acceleration. Key outcomes from the performance analysis are as follows. Speech-to-Text: Fun-ASR-Nano achieved the highest transcription accuracy with a Word Error Rate of 0.026, while Moonshine Tiny was the most efficient, recording a Real-Time Factor of 0.036 on the CPU. Scaling up to the high-performance tier, Whisper Large-V3 Turbo demonstrated high speed and robustness on a dedicated GPU, achieving an RTF of 0.093. Language Modeling: The Qwen 2.5 (1.5B Instruct) model, optimized for the Intel edge NPU, delivered robust constrained edge performance with an average generation speed of 20.15 tokens per second and a high semantic accuracy score of 0.86. The non-transformer Liquid LFM-24B model showcased server-level reasoning capabilities on the high-performance edge, reaching an impressive 39.2 tokens per second when fully offloaded to a dedicated GPU, despite its massive VRAM requirements. Text-to-Speech: Piper TTS emerged as the most efficient model for constrained environments (RTF of 0.034). However, Kokoro TTS redefined high-fidelity zero-shot synthesis on the GPU tier, achieving a groundbreaking RTF of 0.024 and far outperforming larger autoregressive audio models like OuteTTS, which remained too slow for real-time use without significant acceleration. Hardware Acceleration and Energy Efficiency: The use of Intel OpenVINO 2026.0 for hardware offloading significantly reduced energy consumption; for example, Whisper Large-V3 Turbo's energy per audio second
s on the CPU to just 3.24 Joules on the integrated GPU. Fur-
acceleration revealed a critical "race-to-sleep" paradigm, where
et by drastically reduced processing times. The study concludes
cascaded pipelines: a constrained edge tier (Moonshine, Qwen
Khadas NUC (Khadas Technology, Shenzhen, China powered
l Corporation, Santa Clara, CA, USA) maximizing energy effi-
ance tier (Whisper V3 Turbo, Liquid LFM-24B, Kokoro) running
ivering uncompromising accuracy and subsecond latency for
ced edge AI.

g; speech-to-text; text-to-speech; small language models; GGUF;
hmarking; modular framework

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

716 different pipelines, 3389 considering the deployability of Intel CPU, NPU, GPU hardware.

https://github.com/mazzonelorenzo1/audio-models-testbench



Selecting The Winning Trio

STT: Moonshine Tiny (CPU) | RTF: 0.036 | 0.8 J/s | Pwr: ~22 W | Max Temp: 77°C

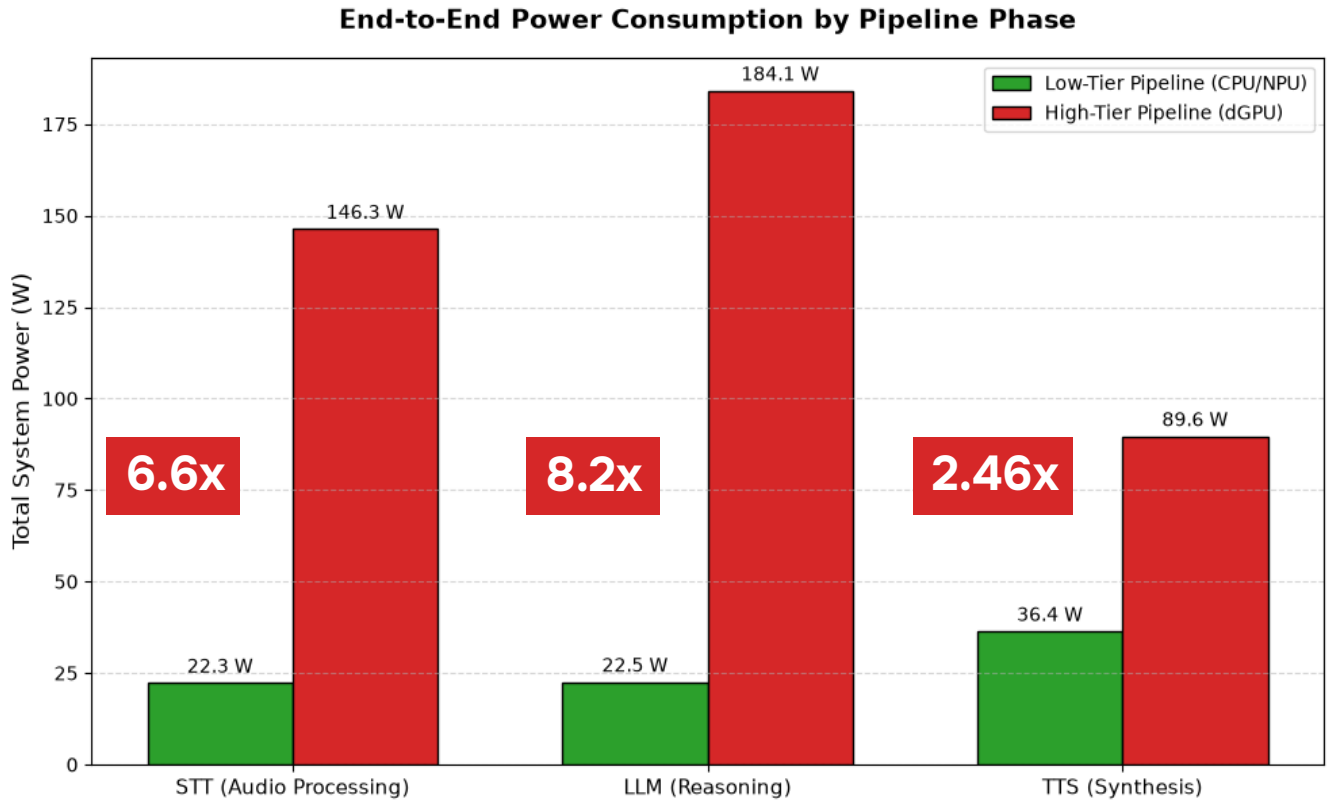
ELM: Qwen 2.5 1.5B (NPU) | Speed: 20.15 Tk/s | 1.12 J/tok | Max Temp: 66°C

TTS: Piper TTS (CPU) | RTF: 0.034 | 1.24 J/s | Pwr: ~36 W | Max Temp: 81°C

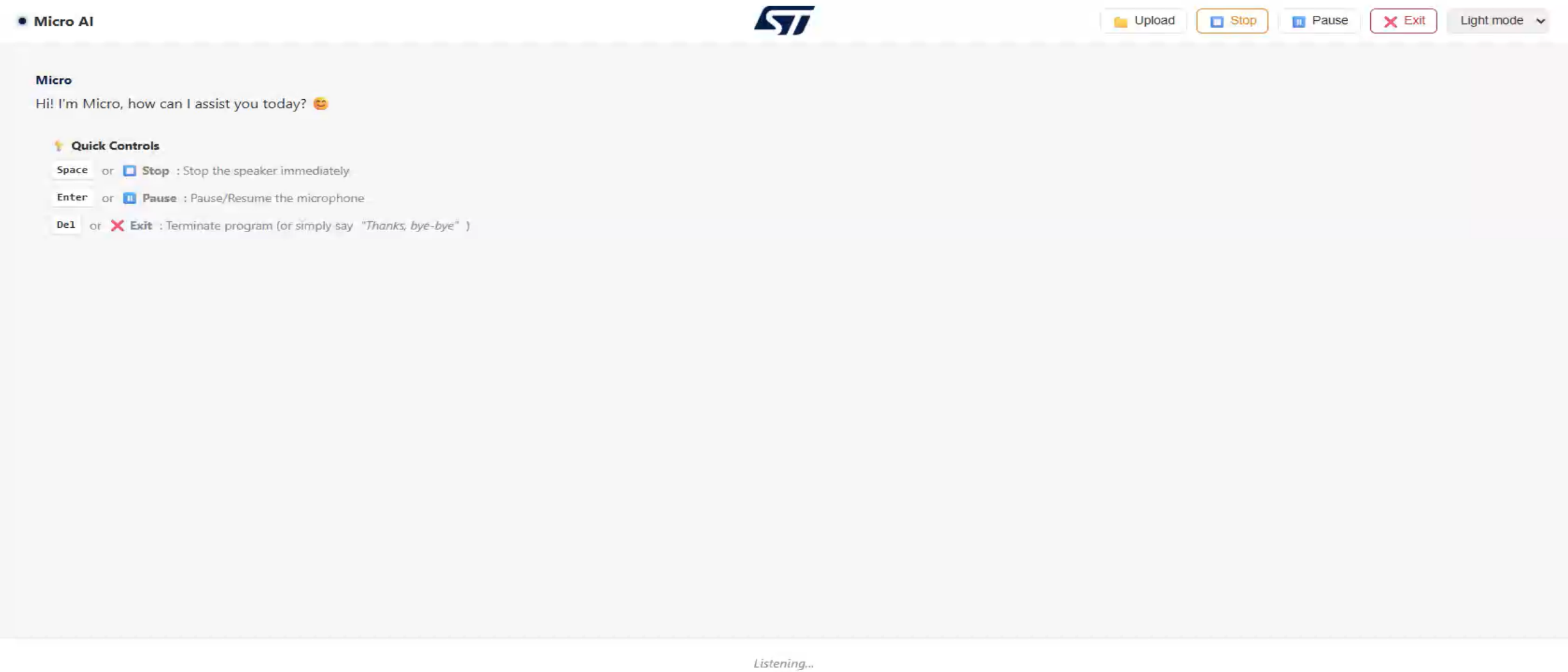
Khadas Mind 2 AI Maker Kit



NVIDIA RTX 5060 Ti

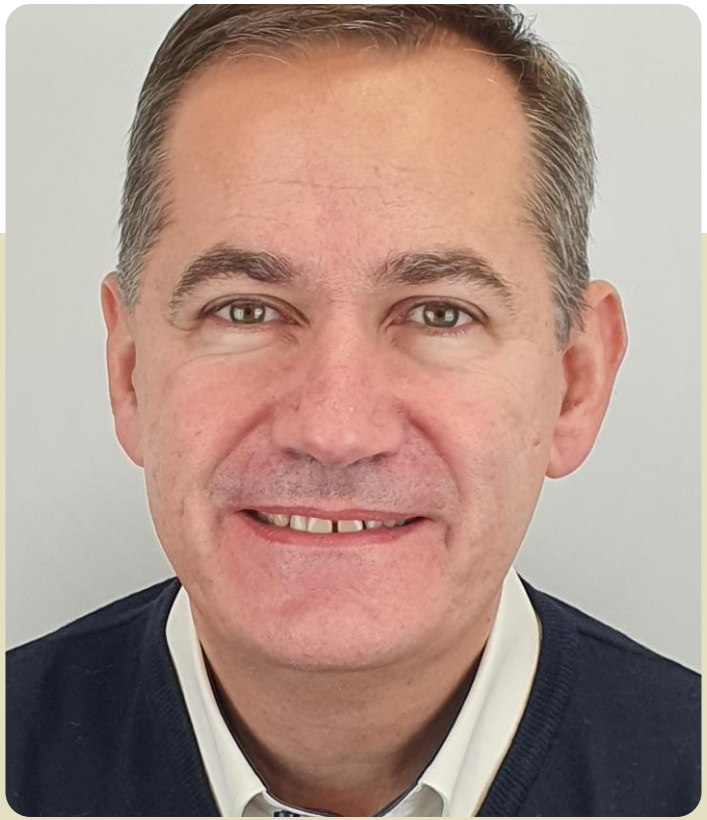


Conversational edge agent



Authors: Lorenzo Mazzone, Danilo Pau
Property: STMicroelectronics

Danilo PAU



Technical Director, System Research
IEEE, AAIA, ST Fellow
Full professor ASN 09/E3 HES Italy
IEEE Distinguished Industry Lecturer
Scientific Committee 2026-2028 IEEE SPS Italy
Global Governance and AI Safety NAAI Committee
Editorial Board Member MDPI Electronics
APSIPA Life Member

danilo.pau@st.com
+393485318664
https://www.linkedin.com/in/danilopietropau/
https://scholar.google.com/citations?user=nRKj7EMAAAAJ&hl=en