Axelera Edge 232p

AI on your terms, powered by Europa

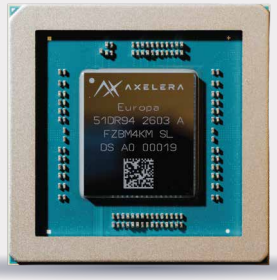
Purpose-built inference for enterprises and system builders running Physical AI, private generative AI on their own infrastructure, or computer vision at scale. It all runs from a standard PCIe slot, without rebuilding your stack.

The Axelera Edge 232p brings Europa, Axelera's second-generation AI processing unit, to a standard half-height, half-length PCIe card. It is purpose-built for high-performance inference in a small energy footprint: real-time answers, a system you own outright, and a lower total cost of ownership.

One card runs the full spectrum of your inference needs: advanced vision, reasoning, language, and action models, all in one easy-to-deploy card. Pre-processing, post-processing and video decode stay on board, focusing the CPU utilization on business logic instead of image processing. And because it is a standard air-cooled PCIe card in a standard power envelope, inference deploys where your servers or systems already are: on-premises, air-gapped if required, with your models and data never leaving your infrastructure.

Key Features

- **Any modern AI workload.** CNNs, vision transformers, VLMs, VLAs, LLMs and agentic pipelines, at INT4, INT8 and INT16.
- **160+ pre-optimized models** across 16 task families, 24 ready-to-run reference pipelines, and support for your own weights and models.
- **The whole pipeline on board.** Dedicated vector processors and a hardware HEVC decoder keep pre/post-processing and video decode off the host allowing for a cheaper host CPU, and multi-camera deployments that needed two cards may need only one.
- **No infrastructure rebuild.** Single-slot, air-cooled, standard PCIe. Same system, same cooling, same power budget.
- **AI you control.** Hardware Root of Trust and Secure Boot. Inference runs entirely on-card. For legal, financial services, healthcare, government and defense workloads where data cannot leave the building. Tokens are not spent in the cloud, but utilized directly within your control.
- **Zero lock-in.** RISC-V throughout, Voyager® SDK on GitHub, same toolchain, same developer workflow.



Key Technical Specifications

AIPU	1x Europa
AI Performance (INT8)	629 TOPS
Precision	INT4 / INT8 / INT16
AI Cores	8x 2nd-Gen, D-IMC with VPU, Improved transformer support
Memory	LPDDR5, 32 or 64 GB, 200 GB/s
Video Decode	HEVC/H.265, 4K120FPS, 16x FHD or 4x 4K@30FPS
Security	Secure Boot, Root of Trust
Form Factor	PCIe Single Slot HHHL (Half-Height, Half-Length)
Host Interface	PCIe Gen4 x4 (8 GB/s)
Thermal Solution	Active cooling (Passive variant available)
Power Consumption	45 W per card (30-40 W typical)
Operating Temperature	-10 to 60°C ambient

Easy to Integrate

VOYAGER TOOLCHAIN

Axelera Wingman Builds it for You

- **Describe the pipeline, get a running code.** Axelera Wingman builds the pipeline, compiles it, runs it on the card and benchmarks the result, drawing on a curated set of workflows, tools and knowledge for Axelera hardware.
- **Need to port your existing project?** Axelera Wingman adapts your existing inference stack, including for non-Axelera hardware, and builds your Axelera pipeline, often from a single prompt.

Voyager SDK Gives You Direct Control

- **Fast end-to-end integration.** A high-level pipeline framework builds optimized AI applications with pre-optimized models, datasets, models and business logic in very few lines of code.
- **Measure before you commit.** Built-in tools and turnkey Model Zoo networks let you evaluate performance, accuracy and power consumption on the card in minutes.
- **Direct hardware control.** For teams running their own pipelines, low-level APIs give direct access to the inference hardware.

A Simple, Feature-Rich SDK

160+ pre-optimized models across 16 task families, with 24 end-to-end reference pipelines as templates.

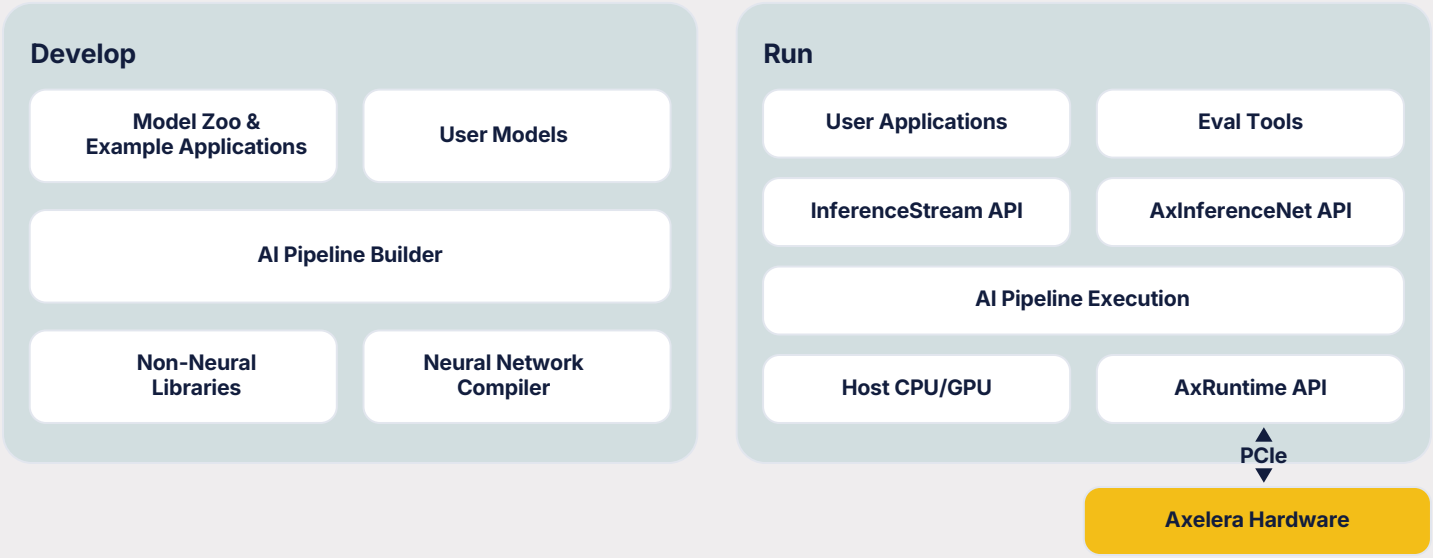
- **Detection** Ultralytics YOLO, YOLOX, YOLO-NAS, SSD-MobileNet, RetinaFace
- **Classification** ResNet, EfficientNet, MobileNet, DenseNet, RegNet, SqueezeNet, Inception, ResNeXt
- **Segmentation and Pose** Ultralytics YOLO for instance, semantic, pose and oriented bounding boxes, plus U-Net
- **Vision transformers** ViT-B-16, SigLIP2, DINOv2 for segmentation and depth, DepthPro
- **Vision language** Qwen3-VL, SmolVLM2, LFM2.5-VL
- **Language models** Qwen3, Llama 3.2, Llama 3.1, LFM2.5, Phi-3-mini
- **Specialist** Re-identification, depth estimation, license plate recognition, super-resolution, face recognition

Bring Your Own Weights / Bring Your Own Models

Pair your own weights with a standard architecture and deploy them through the same path every Model Zoo model uses. Axelera Wingman works alongside it, turning a plain-language description into the compile, run and validate loop on your card.

Compiler and Tooling

PyTorch and ONNX models compile with automatic quantization and graph optimization. Full pre- and post-processing libraries included, with accuracy and performance evaluation built in.



To order the Axelera Edge 232p,
visit store.axelera.ai/products
Part numbers are listed in the product datasheet.

AXELERA ARTIFICIAL INTELLIGENCE