Hailo Announces General Availability of Hailo-10H Edge AI Accelerator with Generative AI Capabilities

TEL AVIV, ISRAEL, July 22, 2025 /EINPresswire.com/ -- Hailo, the leading chipmaker of edge AI processors, today announced the commercial availability of the Hailo-10H, its second-generation AI accelerator featuring powerful generative AI capabilities. Starting today, customers worldwide can download the software from Hailo's website and place orders for the Hailo-10H, marking the highly anticipated arrival of the first market-available and orderable discrete AI processor with native support for generative AI workloads at the edge.

Building on the proven success of the company's first-generation AI accelerator, Hailo-8, the Hailo-10H complements the industry-leading performance in vision AI tasks with new generative AI capabilities and introduces, for the first time, the ability to run large language models (LLMs), vision-language models (VLMs), and other generative AI models entirely on-device, without relying on cloud connectivity. This leap in functionality brings powerful generative AI to edge devices with unmatched cost and power efficiency, making the Hailo-10H a game-changing solution for product developers across multiple sectors including personal compute, automotive, telecommunication, retail, security and more.

"With the Hailo-10H now available for order, we're taking another major step toward our mission of making AI accessible to all," said Orr Danon, CEO and Co-Founder of Hailo. "This is the first discrete AI processor to bring real generative AI performance to the edge, combining high efficiency, cost-effectiveness, and a robust software ecosystem."

The Hailo-10H is fully compatible with Hailo's mature and widely adopted software stack and benefits from the support of a vibrant global developer community with more than 10,000 users each month. It empowers developers to run state-of-the-art vision and generative AI models directly on edge devices, delivering real-time responsiveness with ultra-low latency.

By processing data locally, the Hailo-10H ensures strong data privacy, since personally identifiable information remains on the device, while significantly reducing overall system costs by minimizing both cloud bandwidth usage and the need for expensive cloud-based AI service subscriptions. Just as importantly, the AI operates independently of cloud connectivity, ensuring consistent availability even in environments with limited or no internet access.

Specifically designed for edge devices across consumer, enterprise, and automotive markets,

including media centers, home gateways, and automotive cockpit systems, the Hailo-10H enables advanced use cases like natural language human-machine interaction, visual awareness, and multi-modal AI to run seamlessly within the power and cost constraints typical of edge environments.

In performance benchmarks, the Hailo-10H has demonstrated outstanding results across generative workloads. For example, achieving a first-token latency of under 1 second and over 10 Tokens per Second on a variety of 2B language and vision-language models. For video analytics, the Hailo-10H enables state-of-the-art object detection (e.g., YOLOv11m) on a real-time 4K video stream. All of these come at a typical power consumption of just 2.5W, making it ideal for compact, efficient AI-enabled systems. The Hailo-10H is automotive-qualified to AEC-Q100 Grade 2 standards and is aimed at automotive designs with 2026 start of production.

With the Hailo-10H now available for order, Hailo continues to lead the way in redefining what's possible with AI at the edge, delivering the future of intelligent computing in the here and now. To start integrating Hailo-10H into your next product, contact Hailo.

For hi-res imagery, click here.

About Hailo
Hailo, an edge AI-focused chipmaker, is developing specialized AI processors that enable data center-class performance on edge devices. Hailo's processors are the product of a rethinking of traditional computer architecture, enabling smart devices to perform sophisticated deep learning tasks such as object detection and segmentation in real-time, with minimal power consumption, size, and cost. The processors are designed to fit into a multitude of smart machines and devices, impacting a variety of sectors including compute, automotive, security, industry 4.0, and retail.

Marta Majstorovic
Griffin360
marta@griffin360.com
