AI Accelerators For Machine Vision Competitive Analysis And Review: Developer Experience And Performance Evaluation

HOT TECH | VISION AND ANALYSIS

June 2025 | Commissioned By Axelera AI

Table Of Contents

Key Takeaways

- The Axelera Metis PCI-Express AI accelerator showed the highest performance in our multi-stream video inference tests.
- The Axelera Metis PCI-Express and M.2 modules consistently had the best energy efficiency.
- In combination with their overall cost, cost of operation and performance, Axelera Metis accelerators exhibited the best TCO value proposition in this use case.
- The developer setup experience with the Axelera platform was by far the easiest among the manufacturers represented in our study.
- Our accuracy observations favored Axelera's accelerators, which found more objects and placed more detection boxes around subjects versus competing devices.

Research Overview: AI Accelerator Multi-Stream Computer Vision Performance

This research report encompasses the results of our setup and testing experiences with an AI accelerator-enabled workstation, for development in a computer vision use case. We employed an x86 host platform with AI accelerators from leading silicon manufacturers for our evaluation purposes. The design goal was to simulate multiple Full HD 1080p video input streams while utilizing stable, mature existing AI inference models to provide object detection, identification and intent observation.

This is a common application designed to simulate multi-camera security systems that monitor retail stores, warehouses, schools, government buildings and other secure facilities. On the pages ahead, we will discuss the setup and development experience, inference performance and accuracy, and power efficiency across a variety of devices installed on our test platform under the Ubuntu Linux operating system.

AI Inferencing Test Setup And Methodology

Our testing was performed using Ubuntu 22.04 LTS, which is a current version of the operating system with long-term support until 2027. To test our target hardware, we used the methodologies and applications outlined above on our test platform. The workstation machine had the following configuration:

- Intel Core i7-13700K 16-Core/24-Thread Host Processor, standard air cooling
- ASUS Z790 TUF Wi-Fi motherboard
- 64 GB (2 x 32 GB) DDR5-6000 CL30 memory
- 2TB WD SN770 PCI-Express 4.0 M.2 NVMe solid state drive
- Corsair RM750e 80-Plus Gold power supply

Multi-Stream Computer Vision Performance Testing

Our test workload consisted of 14, 30 fps video streams. Under inference workloads, the host Intel Core i7 -13700K CPU in our test system never exhibited more than roughly 30% utilization, indicating that there were CPU resources to spare. As such, our video decoding and re-encoding pipeline was not a processing bottleneck.

This was an important part of our test. Even though the GPUs have dedicated hardware encode and decode engines, we observed performance decreases caused by context switching. Each process that would access the GPU's encoding and decoding resources (28 processes in total, decoding and re-encoding 14 video streams) also required RAM. In this scenario, the CPU provides the best performance.

Power-efficient AI inferencing is critical for a myriad of machine vision use cases in the field. As such, we chose the following AI accelerators based on typical workstation or server installation requirements, to provide offload of object detection and analysis workloads for relatively low power edge inference solutions. All listed prices are accurate as of June 2025 using first-party listings or reputable retailers.

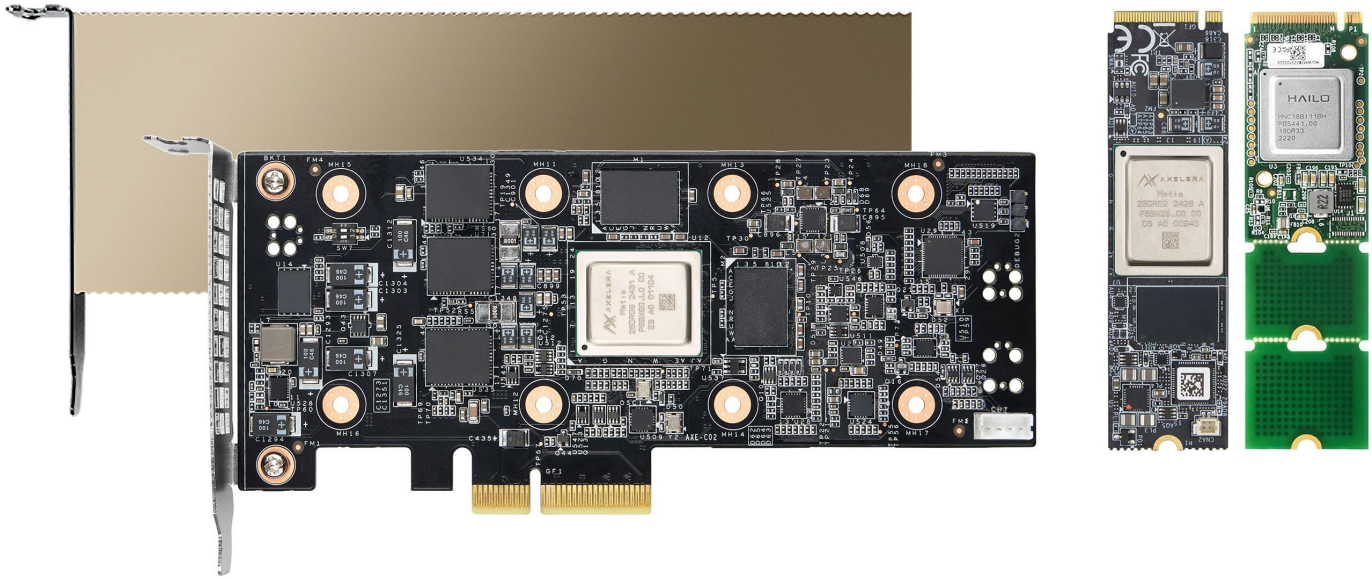
- NVIDIA RTX A4000 16 GB Workstation Graphics Card - $1,000
- NVIDIA RTA A2000 6 GB Workstation Graphics Card - $619
- NVIDIA L4 24 GB Tensor Core GPU - $2,540
- NVIDIA GeForce RTX 3060 12 GB Consumer Graphics Card - $329
- Axelera Metis PCI Express AI Inference Acceleration Card - $384
- Axelera Metis M.2 Inference Acceleration Card - $274
- Hailo-8 M.2 AI Accelerator - $199

We initially intended to include additional AI accelerator configurations in our testing, but were unable to acquire samples from DeepX or higher-end accelerators from Hailo, such as the Hailo-10.

The complete list of Object Detection AI models tested consists of:

- SSD MobileNet v1 & v2
- YOLOv5s & YOLOv5m
- YOLOv7
- YOLOv8s & YOLOv8l

For each model, we collected performance and power consumption data, which allows us to measure not only absolute inference throughput, but also energy cost per frame. Raw performance is reported in frames per second, while energy efficiency is expressed as a function of Joules per frame.

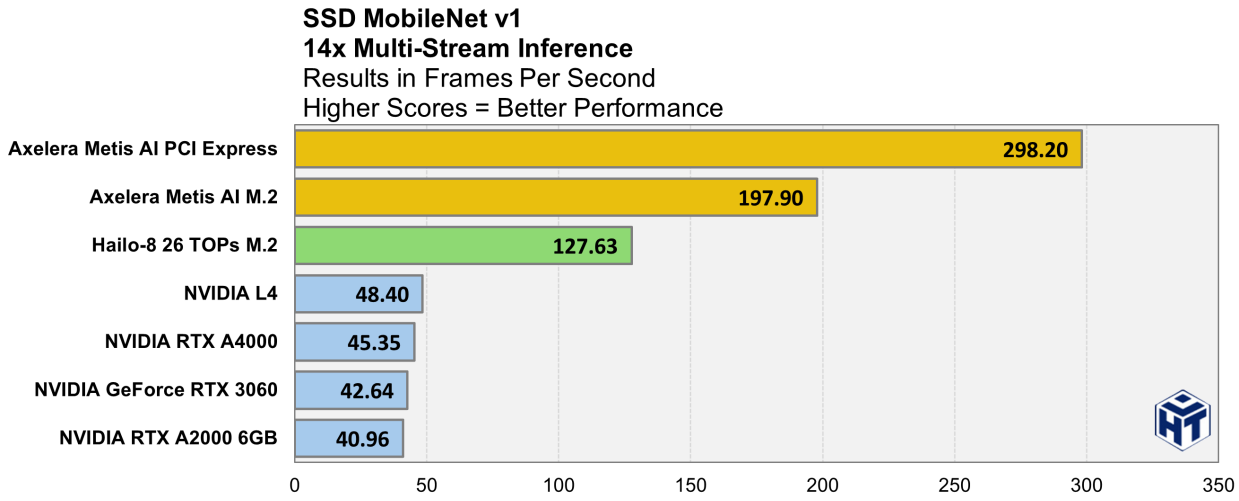


We tested a variety of edge AI accelerators from NVIDIA, Axelera AI, and Hailo in diffrent form factors.

SSD MobileNet Performance Testing

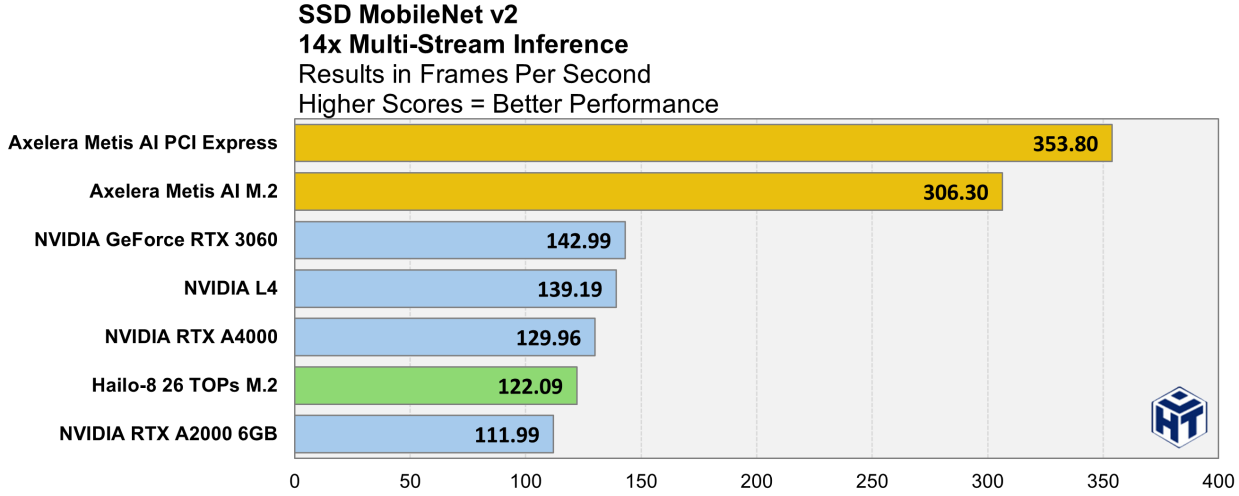
SSD MobileNet v1 and v2 utilize a target resolution of 300x300. Our preprocessing pipeline decoded videos and resized the images to fit those dimensions, including letterboxing as needed to preserve the video's aspect ratio for best accuracy.

SSD MobileNet v1



The Axelera Metis PCI Express cards offered both the highest performance and power efficiency in this test. The Metis M.2 module outperformed the Hailo-8 M.2 module by approximately 50% and showed better efficiency as well. NVIDIA's hardware struggled on this test, although we will see in the remainder of our results that performance on GPUs greatly increases beyond this specific test condition.

SSD MobileNet v2



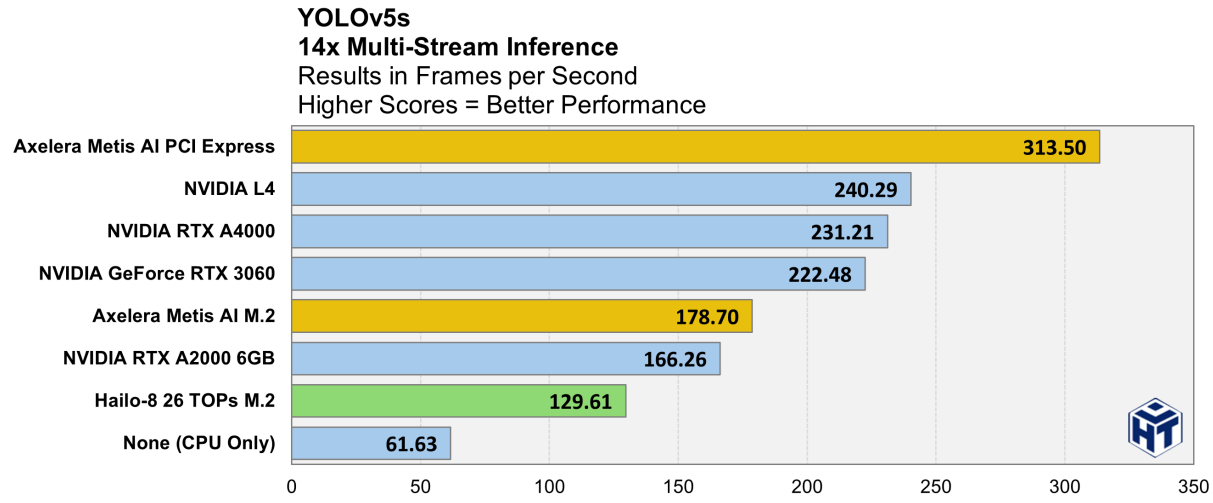
Most of the test platforms saw greatly increased performance in this test. Axelera's Metis accelerators were the best performers by more than double versus any competing platform. NVIDIA's hardware showed the greatest improvements by moving to MobileNet v2 over v1.

The Hailo-8 accelerator experienced a small but repeatable reduction in performance when moving from SSD v1 to v2, and as a result it is surpassed by most of the GPUs in our test group.

YOLO MobileNet Performance Testing

The YOLO models we tested all employ a target resolution of 640x640. Like our SSD MobileNet testing, our preprocessing pipeline decoded videos and resized the images to fit those dimensions, including letterboxing.

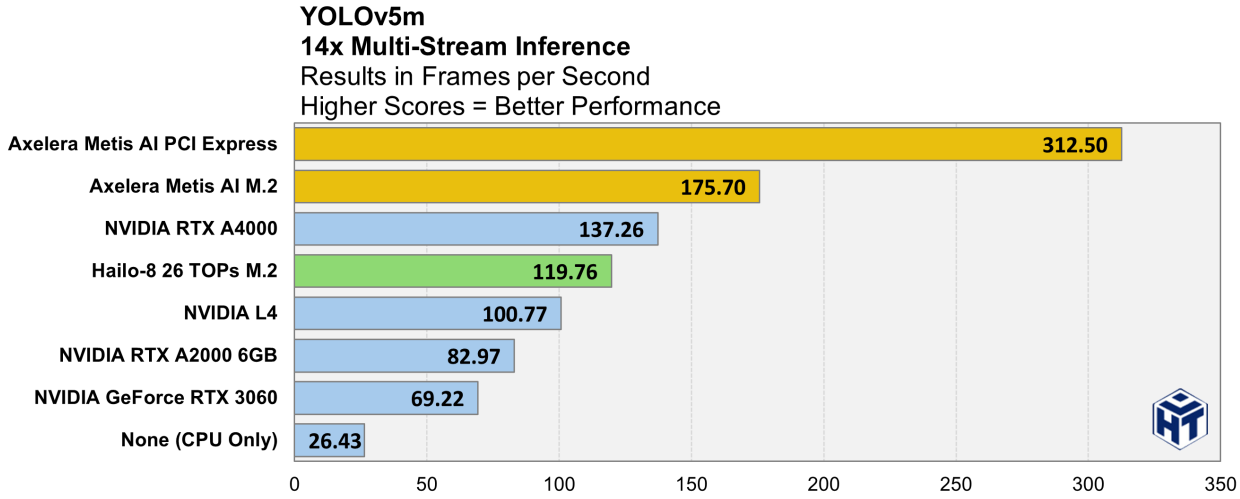
YOLOv5s



For our YOLO models, we are including CPU-only inference results in addition to the AI accelerator results. All of the accelerators surpassed the performance of CPU-only inference by a factor of two at a minimum. The test GPUs offered around four times faster performance than the CPU.

Meanwhile, the Axelera Metis PCI Express AI accelerator was a full five times faster than a CPU-only configuration. As before, the Axelera card was the best performer. Of our M.2 modules, the Axelera Metis was faster than the Hailo-8 by approximately 38%.

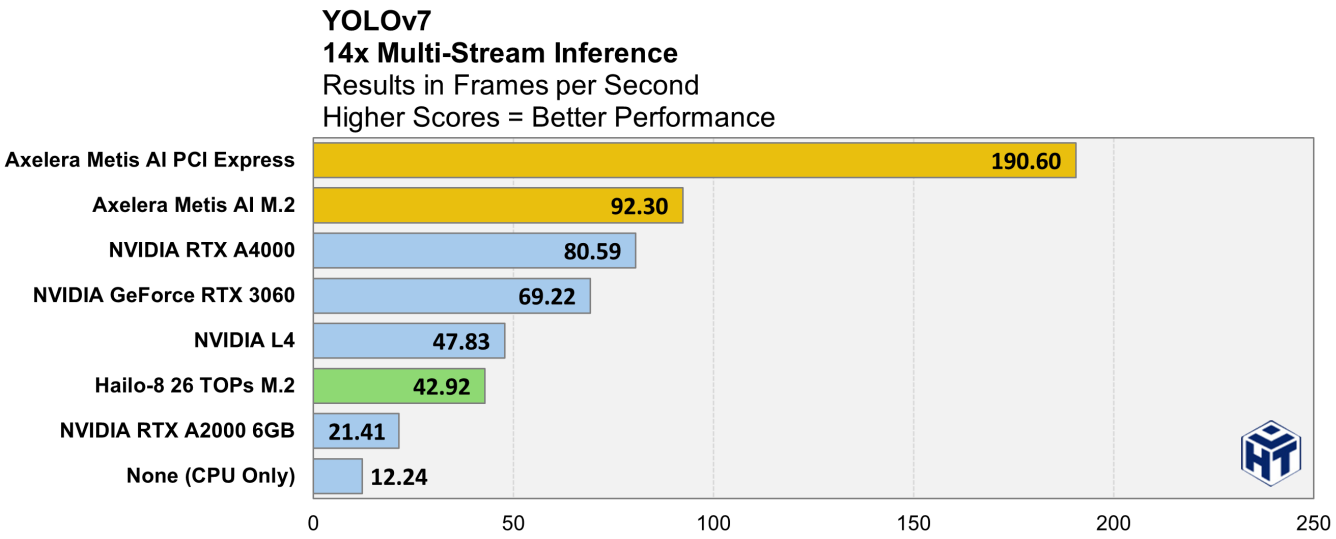
YOLOv5m



On the larger YOLOv5m model, the Axelera M.2 module outperformed all of the competing cards by 15% at a minimum and offered three times the performance of the GeForce RTX 3060.

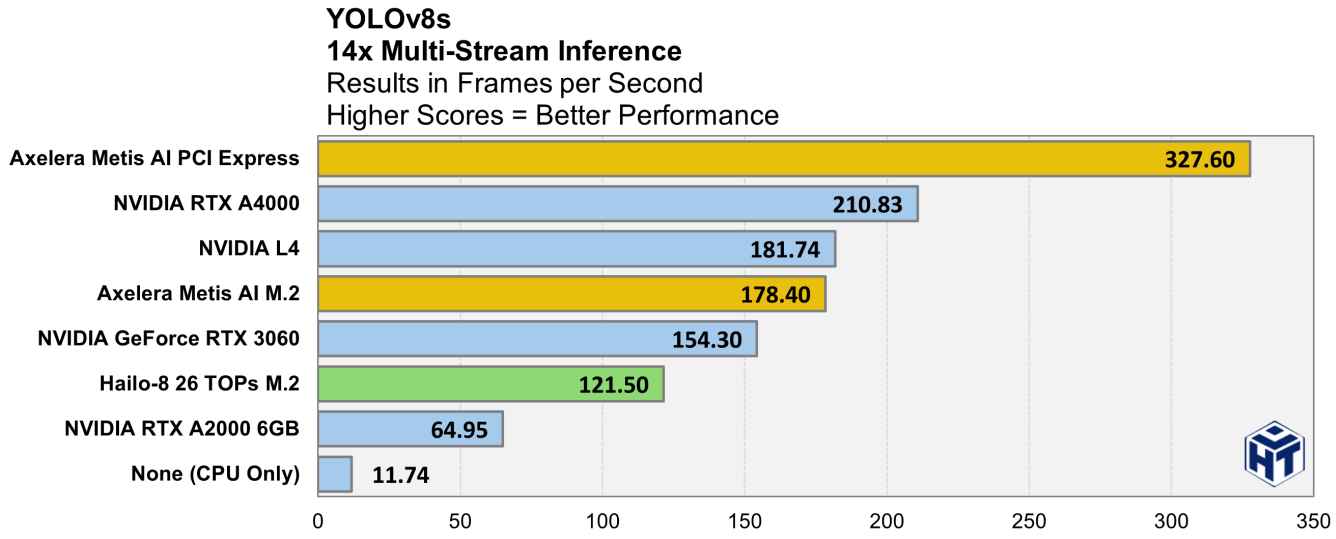
The Axelera Metis PCI Express accelerator, however, was more than twice as fast as any competing device with this heavier model workload. Both Axelera accelerators experienced effectively no drop-off in performance moving from the smaller YOLOv5s model, whereas the competing devices saw a performance reduction. Because the M.2 module did not experience a performance drop-off, we do not believe that the CPU is the limiting factor.

YOLOv7

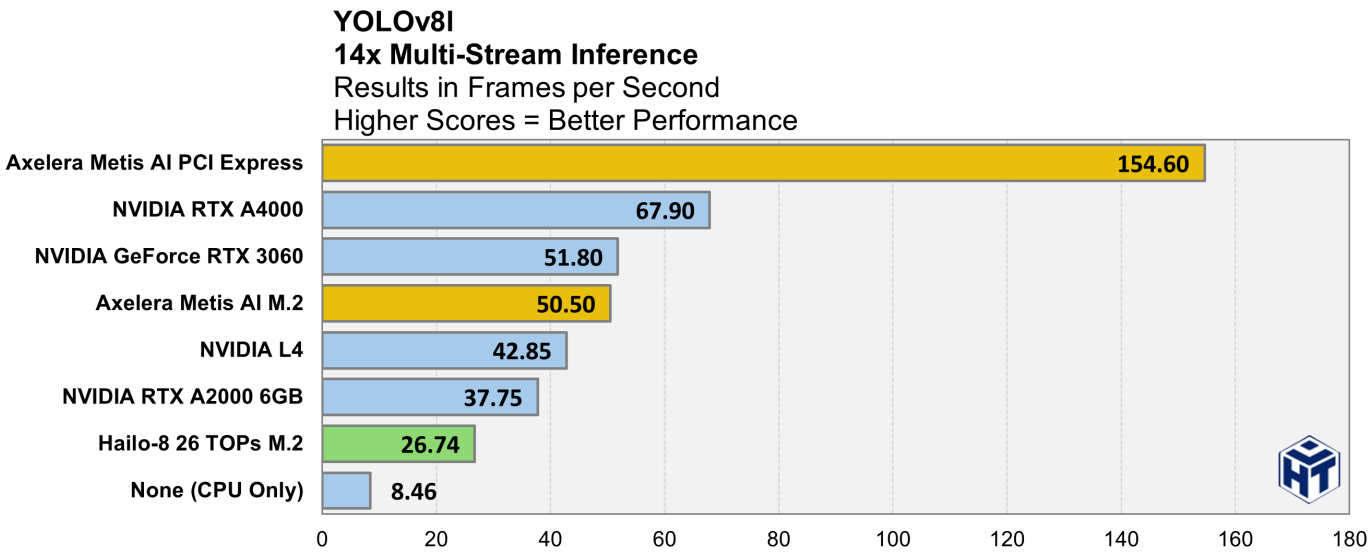


YOLOv7 performance is lower than either YOLOv5 model on all platforms, and efficiency is reduced as a result. Again, Axelera has the fastest accelerators for both PCI Express cards and M.2 modules among our test subjects. This is the first YOLO-based test in which all NVIDIA cards outperformed the Hailo-8. As with our other tests, CPU-only inference is relatively slow and power-inefficient.

YOLOv8s



On the lightweight YOLOv8s model, Axelera's PCI Express card was again the fastest overall by a wide margin. The Metis PCIe card was 50% faster than the closest NVIDIA competitor, the RTX A4000. Among the M.2 modules, Axelera's smaller Metis was 50% faster than the Hailo-8, and was also faster than two of our test GPUs.



The YOLOv8l module is the largest and most taxing AI model of our YOLO test suite. Across the board, each accelerator lost at least 50% of its performance relative to YOLOv8s, and efficiency suffered as a result. Despite this, our test system with the Axelera Metis PCI Express card never used more than two Joules per frame to perform this video AI inference task.

Inference Performance Summary

Across all seven test models, three factors were always true:

- The Axelera Metis PCI Express card was the fastest AI accelerator tested.
- The Axelera Metis M.2 module was the faster of the two M.2 modules tested.
- None of the competing AI accelerators were capable of matching either Axelera Metis accelerators' energy efficiency.
- All accelerators were between 2X and 30X faster than CPU-only inference.

In three tests, including both YOLOv8 tests, the workstation-class NVIDIA RTX A4000 had a performance advantage over the Axelera Metis M.2 module of 20% or more. The largest advantage NVIDIA had over the lower powered M.2 Axelera card was 32% in YOLOv5s.

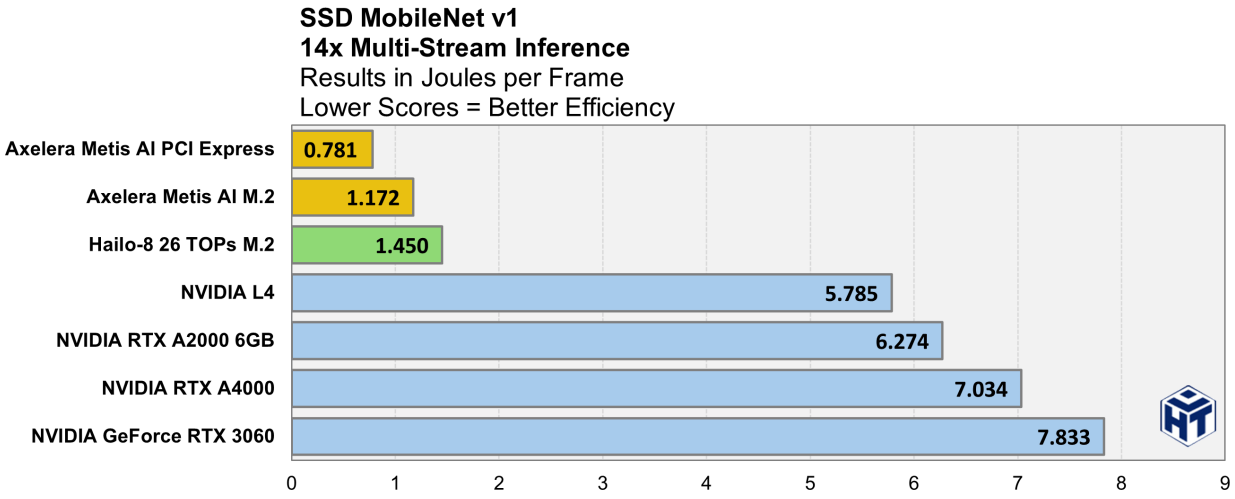
It's important to note that CPU-based inference tasked all available hardware threads, showing between 2,200% and 2,300% utilization in the Linux command line utility top. On the other hand, CPU-based video decode and encode rarely utilized more than 600-800% on the CPU (out of a total of 2400% available for a 24-thread processor). It's likely that the CPU could have been aided by a GPU doing video decode. However, our other results consistently show that the better use of that GPU resource is for the CPU to handle video transcoding and the GPU to do inference.

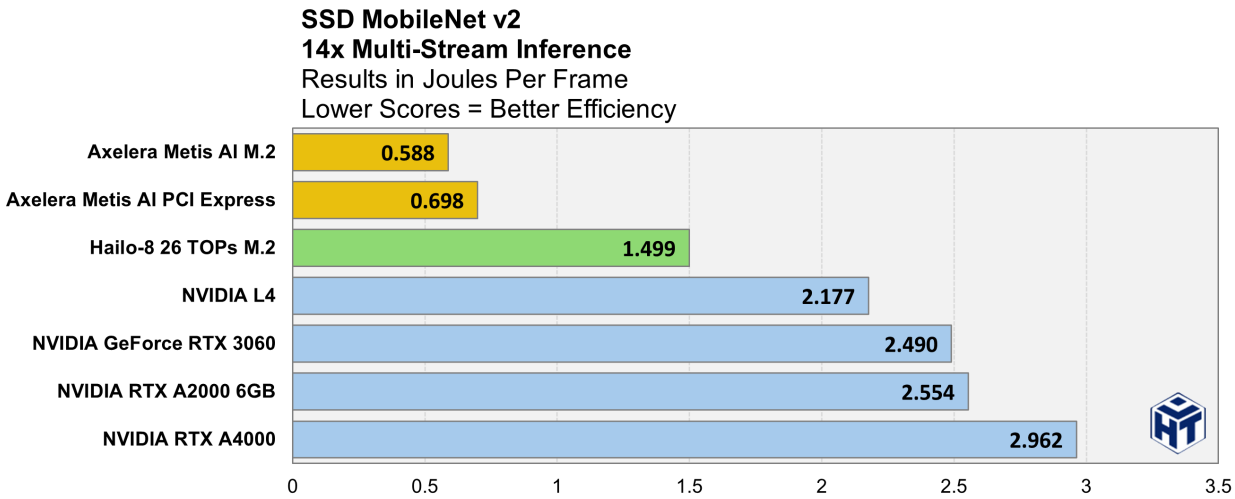
Overall, we found that Axelera Metis PCI-Express and M.2 devices had the highest performance compared to accelerators with the same connectivity among those we were able to acquire at retail or through industrial distribution channels.

Power Efficiency

To test power efficiency, we monitored power consumption at the wall using a power analyzer, logging energy consumed by the second. We used a five-minute period of logged energy tracking, taken after the initial spike caused by filling up the preprocessing queue. We then calculated the Joules used each second. For the Joules Per Frame metric below, we divided average Joules Per Second by the average FPS to calculate constant energy consumption.

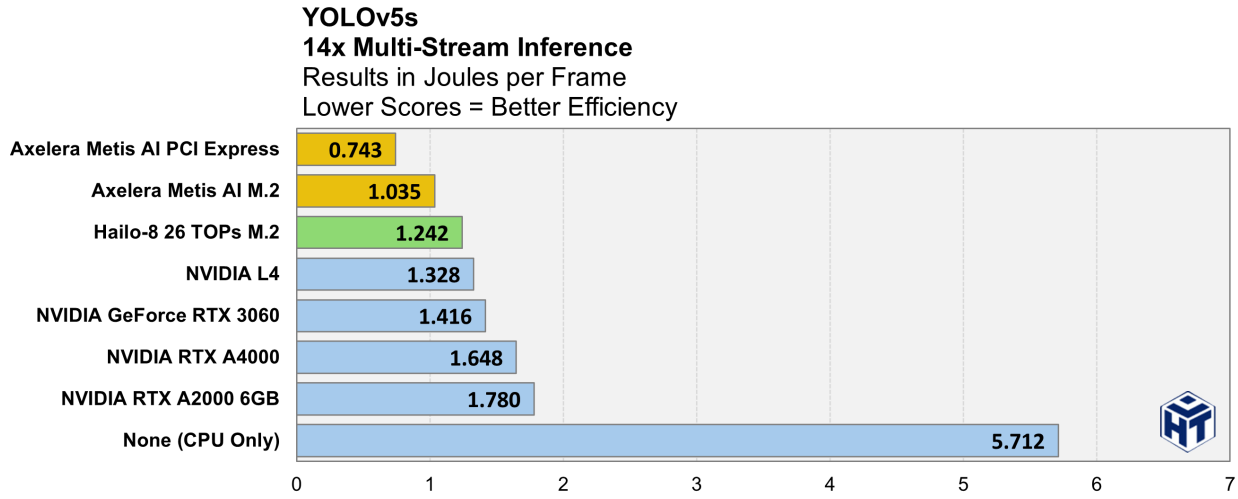
As the charts below demonstrate, the Axelera solutions were consistently the most efficient across every tested inferencing model. The Axelera Metis AI PCI Express Accelerator was the most efficient of the solutions tested. In five of our seven test models, total consumption resulted in an average of less than one Joule per frame of inference, complete with detection boxes and labels.

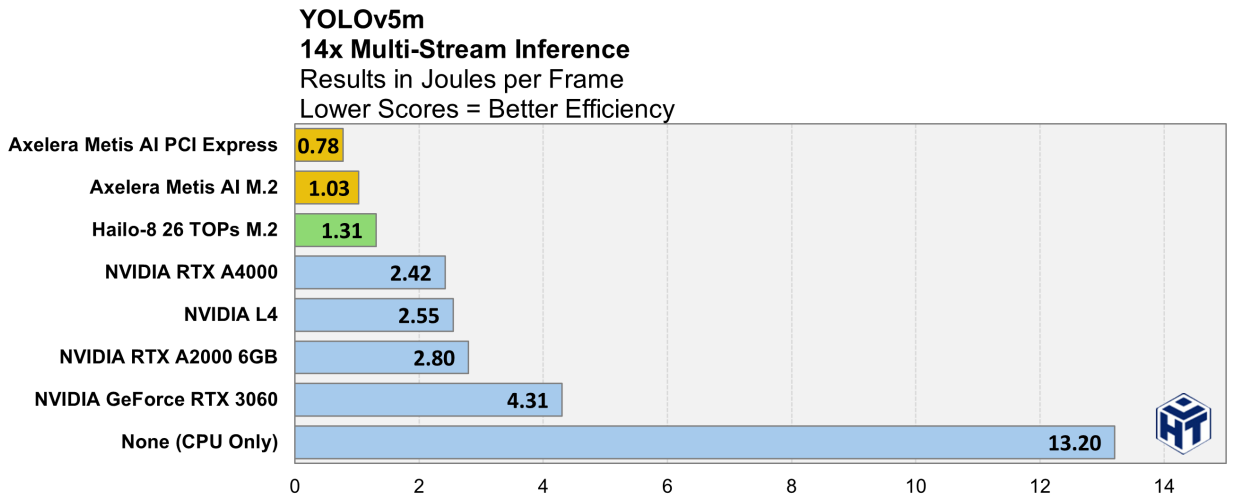




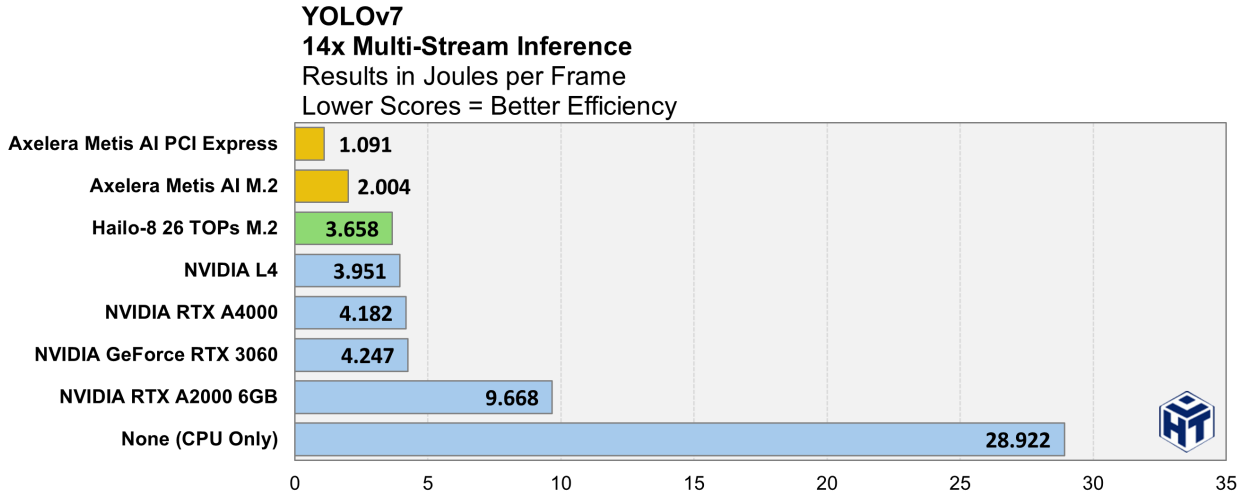
Our testing in SSD MobileNet v1 shows that both the Hailo-8 and the two Axelera Metis accelerator cards have a large efficiency advantage over the GPUs. This is not surprising, since they also offer much higher performance with this test.

The gap narrows somewhat in SSD MobileNet v2, as the GPUs are approximately 2.5x more efficient than with the first model. This is because the GPUs saw a large increase in performance without consuming much more power. Still, all of the accelerators saw efficiency gains to match their performance improvements, with the notable exception of the Hailo-8.



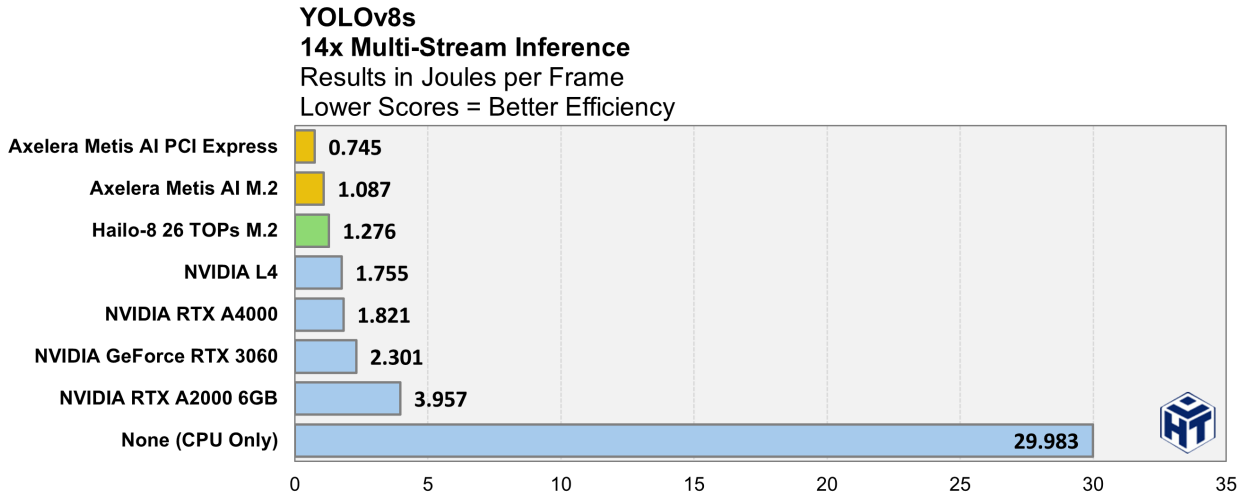


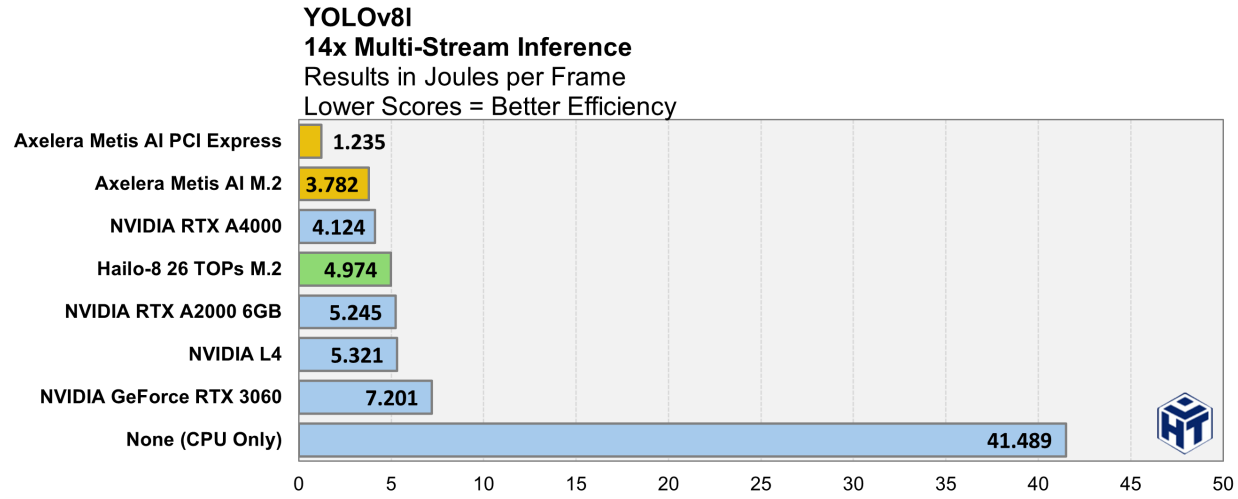
With the YOLOv5 models, the Axelera Metis PCI Express card was by far the most efficient with a 40% advantage over the nearest competitor in both v5s (small) and v5m (medium). We added our CPU performance to these charts to illustrate that even the weakest AI accelerator, which is a consumer-focused GPU with a larger power limit, improved processing efficiency by nearly 75%.



In YOLOv7, each accelerator requires more Joules per frame, as overall performance is lower. In our testing, the total power consumed over a five-minute sample is basically the same for each configuration as it was in YOLOv5. This model is far less performant than either YOLOv5 model above, resulting in lower power efficiency overall.

Power Efficiency (Continued)



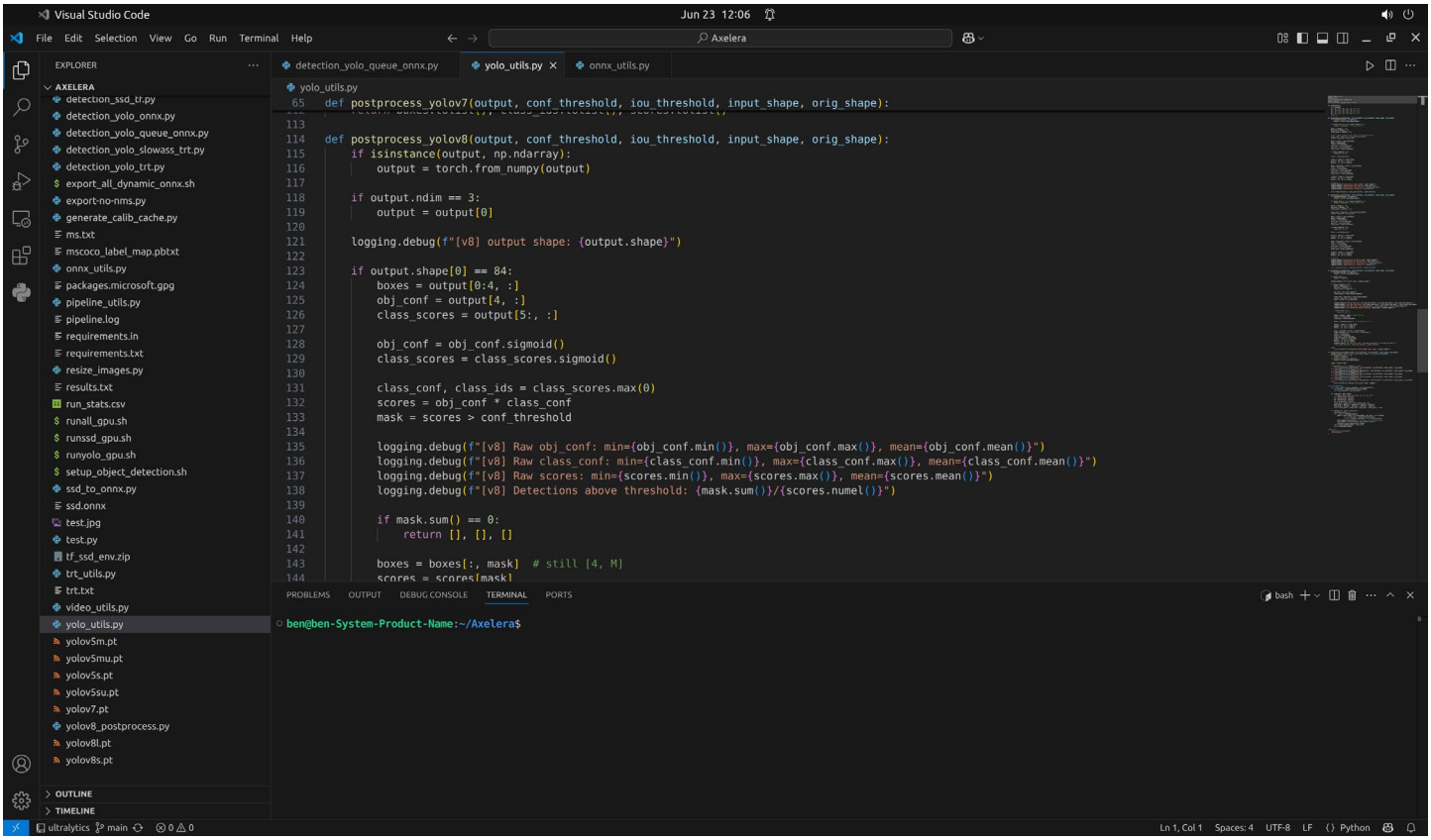


Finally, the two YOLOv8 tests showcase the extremes between small and large models, as every accelerator lost performance without saving much power. Because performance was cut approximately in half, the power consumed per frame of inference workload is approximately double for YOLOv8l compared to YOLOv8s. This resulting efficiency loss illustrates that it may not always be better to choose the more complex model, if your detection results are sufficiently accurate.

The Development Experience

Our chosen models included two Single Shot Detection (SSD) MobileNet models and five different models based on You Only Look Once (YOLO). In all tests, we chose to use publicly available versions of each model. For Hailo and Axelera devices, the manufacturers maintain their own public model repos, and all our targets were available in each.

We had intended to use NVIDIA DeepStream, but our selection of models does not have versions available in NVIDIA's TAO Toolkit. Instead, we opted to use the ONNX Model Zoo.



Development Environment Setup Workflow

Each accelerator platform we tested required its own proprietary setup, and each platform provides its own drivers and dev libraries. However, they all share some commonalities: the need for drivers, an SDK with supporting libraries, and the Python programming language.

The setup experience was easiest by far with Axelera's Voyager SDK. The only requirement was to clone the SDK repo from GitHub and run the included installer script. Everything we needed was installed, and after rebooting we were ready to begin. The SDK includes a utility to download models from Axelera's repository, which happened automatically the first time each model was used.

Hailo's SDK for our Hailo-8 M.2 device runs in a Docker container that has been pre-compiled by Hailo. We needed to install drivers manually, but the rest of the setup process was straightforward. The SDK includes all of the tools to download, optimize, and compile models from the Hailo Model Zoo. We did have to do this manual step for each model, and the process ranged from 10-15 minutes for the lighter models to over an hour for YOLOv8l. This is not something we had to do for other models.

Setup with NVIDIA was more manual since our selection of models was not compatible with DeepStream. We installed drivers, CUDA, and ONNX Runtime with the TensorRT Execution Provider for best performance. For each model, we used official models from the ONNX Model Zoo, rather than building our own.

Configuring Inference Processing On Multiple Video Streams

Our test scenario is a common one among retail stores and secure facilities: running object detection on a collection of video streams. This simulates a security system to aid in retail loss prevention or to ensure that only authorized personnel have access to secure areas, for example.

NVIDIA GPU Inference Application

For our GPU tests, our choice to not train custom models meant that we could not rely on NVIDIA TAO or DeepStream sample code. Instead, we built an application with multi-threaded video encode and decode blocks and a queue system to feed batches of frames into a shared inference provider running on the GPU. This follows best practices including a reduction in context switching, and batch inference means fewer small transfers across the PCI Express bus.

Our custom application ensured that video with detection boxes and COCO labels is then saved for us to perform accuracy checks. The sample application required approximately 28-32 hours of development and testing time before performance monitoring could begin.

If all of our desired models were available in the TAO model catalog, much of this development effort could have been saved, since the DeepStream SDK includes a sample application for handling our use case of multiple input streams with annotated output. We may have needed to do some development work for recording performance metrics with that application, but we would not have had to start from scratch.

Hailo AI Accelerator Inference Application

The Hailo SDK includes a sample app for running inference using models from its model zoo directly out of the box, complete with annotated video encoding. However, the sample application cannot take more than one input at a time, and it only supports image files rather than video streams.

We spent around 12 hours converting the sample single-input application to handle the same multi-threaded video encode and decode blocks of our GPU-focused application and run inference in batches instead of one frame at a time. The Hailo sample application provided all the inference and postprocessing. However, if manual postprocessing is desired as part of a deeper integration with an existing application, the Hailo SDK supports that as well.

Axelera Metis AI Accelerator Inference Application

Axelera's SDK ships with sample application code that performs various inference tasks. One of those sample applications was a full-featured demo application that happens to support multi-stream inference. In addition, Axelera's sample application reported performance metrics and saved annotated output to a user-specified location. We elected to spend around two hours implementing our own measurements to ensure that reporting was consistent among platforms. We also used the app's command line arguments to turn off features that we did not need, like live video preview, for example.

Unlike the competing platforms, we did not have to do any additional development work to handle multiple simultaneous inputs, output postprocessing, or data saving. However, if developers desire deeper integration within an existing application, the Axelera SDK provides low-level hooks via AxRuntime to communicate with devices.

By using a common interface for a variety of hardware classes, all of our custom preprocessing and postprocessing logic should be compatible with not only the GPUs we tested, but all the AI accelerators tested as well.

Accuracy Checking Video Inferencing Results

To ensure that our sample applications and their associated models were detecting objects correctly, we saved all annotated frames to output files. All outputs had non-max suppression applied prior to drawing detection boxes. This postprocessing algorithm prevents duplicate, overlapping boxes being drawn around the same object. This can happen when multiple features of the same object are detected. The result should be one box per detected object, as long as at least one detection crosses the confidence threshold.

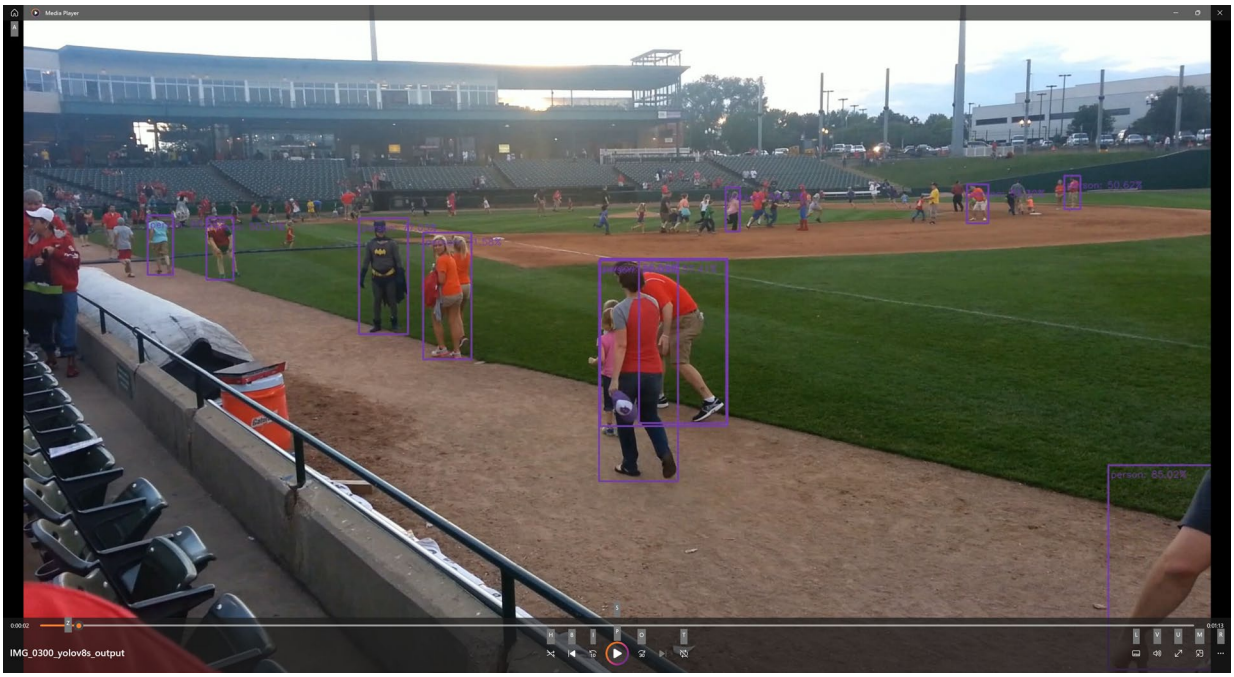
Confidence was calculated as {object detection confidence} X {class confidence} with a minimum threshold of 0.25. This means, for example, that any detection with a confidence score of 0.5 and a class confidence of 0.5 would meet the threshold, as 0.5 X 0.5 = 0.25. This is traditional best practice in determining accuracy.

Overall, we found the Axelera Metis devices to find the most objects within the same threshold. In nearly all of our tests, the Axelera cards had higher confidence scores and drew more detection boxes than the GPUs or the Hailo-8. Generally, all devices found people who were in closer proximity, therefore those who were larger in the inference images. But people in the distance were more challenging for the GPUs and the Hailo-8 accelerator.

Axelera Metis with YOLOv8s, Sample Scene

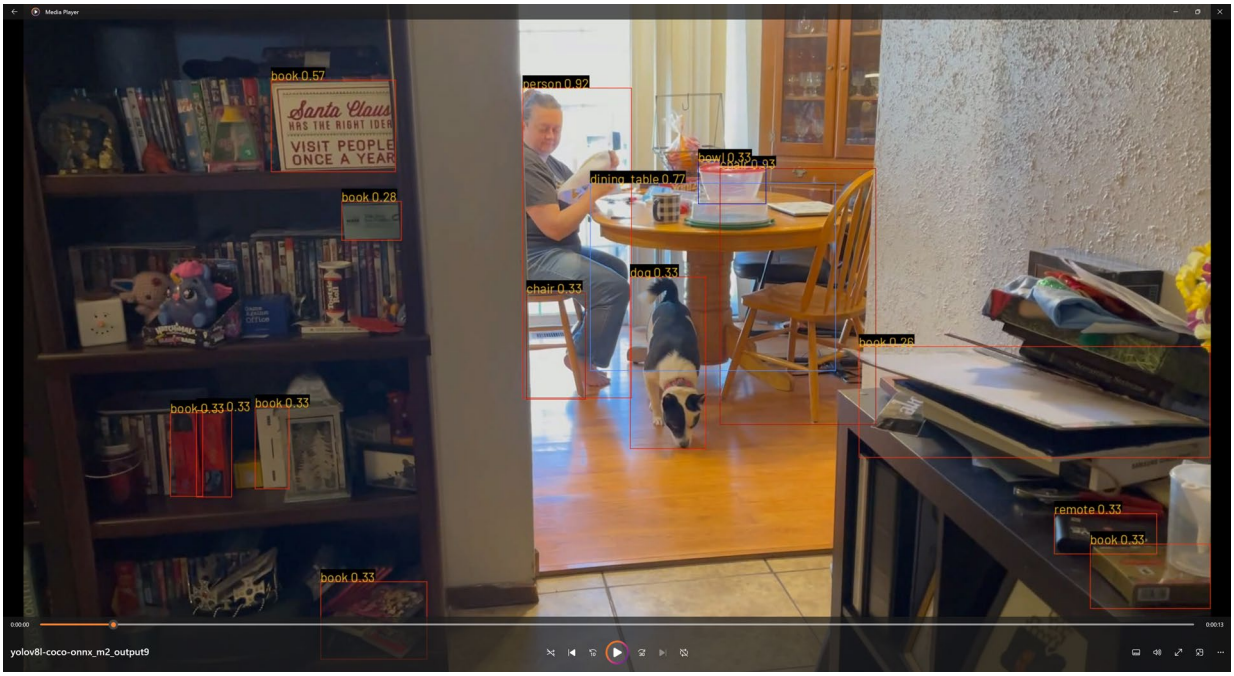
GPU Detection with YOLOv8s, Sample Scene

Accuracy Checking Video Inferencing Results (Continued)



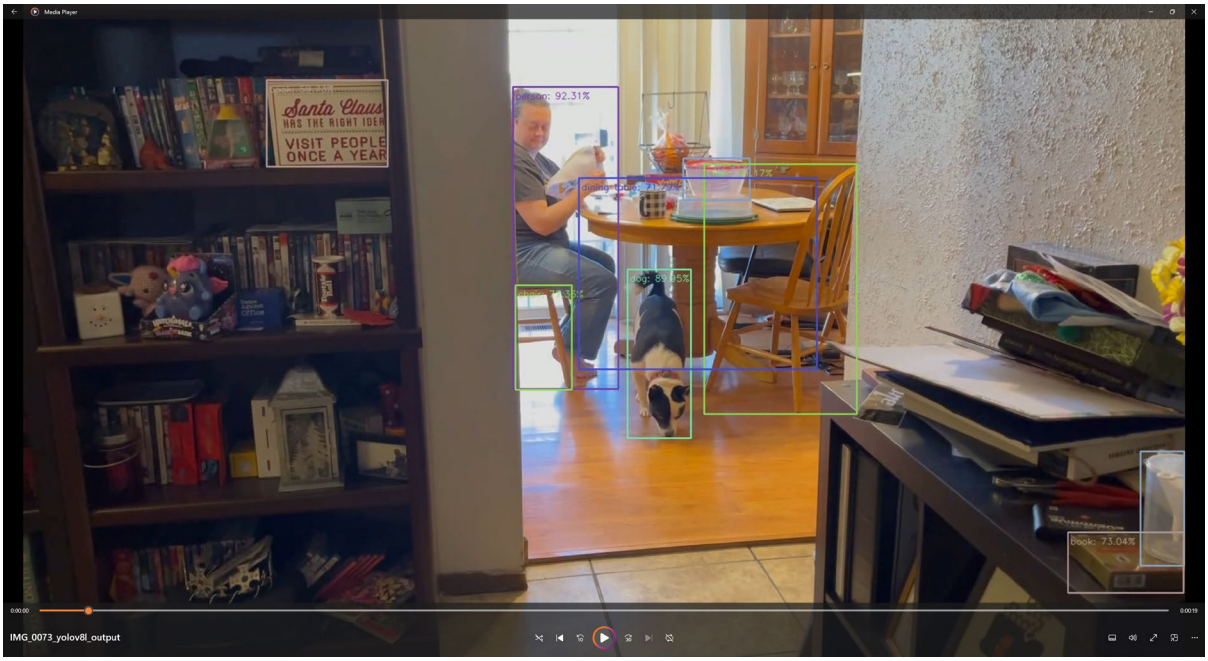
Hailo-8 with YOLOv8s, Sample Scene

In another test scene with a person up close in the frame, all AI inference accelerators found both the person and other objects in the area. However, Axelera found more up-close objects than the others. Again, we believe that when thresholds are adjusted, other devices may have had more detections.



Axelera Metis with YOLOv8l

GPU Detection with YOLOv8l



Hailo-8 with YOLOv8l

It is worth noting that on the bookshelf, most of the items identified as books by all of the devices are DVD and Blu-Ray cases or video game cases. However, given their rectangular shape and printing on the spines, we don't take issue with this classification. The Common Objects In Context class list only has 80 classes, so the cause is likely the list of available classes used by the models, not the AI accelerators themselves. All three devices also correctly identified the dog, table, and chairs in this scene, as well.

In summary, all of the test devices found objects within their confidence threshold and annotated the correct classes, but the Axelera Metis devices found more objects with more confidence than the others. Given the highly sensitive nature of security systems, we preferred the output of the Metis devices. Confidence can be adjusted on all platforms, however. We believe that all of the test devices could be made more accurate with finely tuned, calibrated custom models. However, out of the box, Axelera found and correctly labeled the most objects within our test video streams.

Final Analysis And Executive Summary

We tested multi-stream video object detection AI with seven popular models across seven AI processing engines, including GPUs and dedicated inference accelerator platforms. The goal was to evaluate these devices and models in a simulated security camera system. We evaluated performance, power consumption and efficiency, the developer startup experience, and the accuracy of detected objects.

Overall, we found that Axelera Metis PCI-Express and M.2 devices had the highest performance compared to accelerators with the same connectivity. Even the lower-power Axelera Metis M.2 module was faster than all competing devices in the majority of our tests. Axelera's accelerators achieved higher energy efficiency in the same inferencing workload as well. The SDK setup process was simple with both Axelera and Hailo, with an edge going to Axelera, since the SDK installer also takes care of drivers. The quality and flexibility of sample code favored Axelera, which led to a lower overall development time to get a proof-of-concept application off the ground.

Considering the accuracy, efficiency, and retail prices of all accelerators in this test group, the Axelera Metis M.2 and PCI Express AI Accelerator cards provide the best total cost of ownership. The comparison of SDK features and example applications also makes Axelera the easiest, most straight-forward platform currently available to get started with multi-stream AI video inferencing.

About Hot Tech Vision and Analysis

Industry Research: With decades of experience in the computing, communications, and semiconductor markets, both at the executive level and as media, HTVA has direct insight into industry trends, forecasts, product execution, and market impact. From whitepaper research data, event coverage, or live speaking engagements on TV, Radio, and Internet channels, our team provides specific, targeted analysis on the hottest technologies that shape the digital landscape. We cover emerging and mature markets within Computing and Semiconductor technologies, but always maintain a pulse on the cutting-edge.

Product and Market Analysis: Excellence in product development can't happen in a vacuum. Who and what are your competitors? And what does your product or product's relative SWOT matrix really look like? If you're competing in the enterprise or client computing, datacenter, storage, VR/AR, AI, PC gaming, mobile/handset, or the IOT markets, contact us. We can help with our depth and breadth of technical knowledge. We can help with decades of experience in product testing, technical benchmarking, use- case/experiential hands-on analysis, and easy-to-digest feedback. And we can help with insight from hundreds of major technology brands and over three decades of tenure in the industry.

Consulting Services: As trusted advisers to dozens of major tech brands, we already live and breathe in the landscape you're trying to navigate. Whether you require specific product guidance, market feedback, competitive analysis, or Marketing and PR strategic planning, we've seen the best and worst of it. More importantly, we know what works and what doesn't. We'll help you achieve your goals with the critical, clear vision and relevant knowledge to become a respected industry leader.

HOT TECH | VISION AND ANALYSIS

Hot Tech Vision and Analysis
213 Lake Drive
Chepachet, RI 02814
(508) 377-7575
www.hottech.com