Metro AI Suite
Software Developer Guide

Ver 1.9 – December 2025 Update

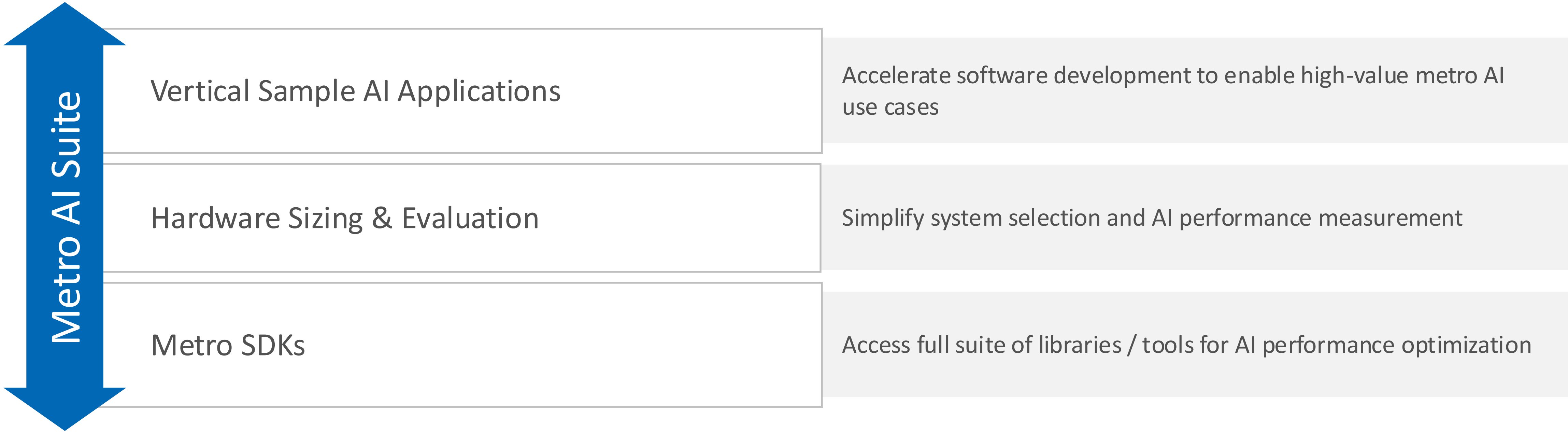
intel

This guide is intended to help software development partners to leverage Metro AI Suite to accelerate their business

Contents

- Introduction to Metro AI Suite
- Metro SDK Overview
- Sample Apps and Blueprints
- Additional Tools
- Recommended Systems
- Partner Support Programs
- Next Steps

Metro AI Suite: An Ecosystem Partner-driven Framework to Accelerate, Optimize and Scale Software Metro AI use Cases.



Vertical Sample AI Applications

Accelerate software development to enable high-value metro AI use cases

Hardware Sizing & Evaluation

Simplify system selection and AI performance measurement

Metro SDKs

Access full suite of libraries / tools for AI performance optimization

Qualified AI-ready systems that reduce risk and accelerate time to AI

Accelerate Edge AI for Critical Use Cases

Rapidly develop GenAI & Visual AI

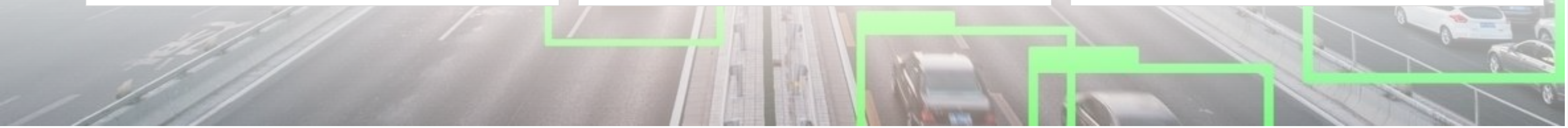
- Reference software to simplify & accelerate development
- Add Gen AI across Intel® edge platforms-even those without discrete GPUs

Size Platforms to Fit Diverse AI Needs

- Hardware evaluation made easy with benchmarking
- Flexible AI sizing options for different performance, power & form factor needs

Optimize AI for Cost Efficiency

- Modular, open and free libraries, tools & microservices
- Maximize AI + Video performance on a range of Intel systems



The Philosophy Behind Metro AI Suite

Simplify adoption of Gen AI and Visual AI

Lower the bar to get Gen AI / Visual AI on your platform

Enable AI on broad range of Intel platforms, from entry to server

Streamline AI performance optimization

Prioritize open-source, Apache License components

Support multiple programming models, including bare metal, containers / services, and Kubernetes

Metro AI Suite for End-to-End Vision AI Solution Enablement

Fine-tune AI Model(s)

Build a model with Geti™

OR

Bring-Your-Own Model
(Custom, Open Source, 3rd party)

Develop & Optimize

Sample AI Applications

Smart Parking

Loitering detection

Smart intersection

...more visual and Gen AI sample apps / tutorials

Vision + AI SDKs

- Microservices
- Video / Media Libraries
- Visual Analytics Libraries

Hardware Selection

Hardware Evaluation

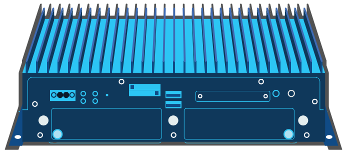
Search Intel® verified AI systems that meet your requirements

Sizing

Use the Visual Pipeline and Platform Evaluation Tool (ViPPET) to quickly evaluate proxy pipelines on potential HW

Go to Market

Deploy and Scale



Create Edge AI systems qualified with Metro AI Suite: Deployable AI Solutions from Partners + Backed by Intel.

NEW RELEASE

Crowd Analytics

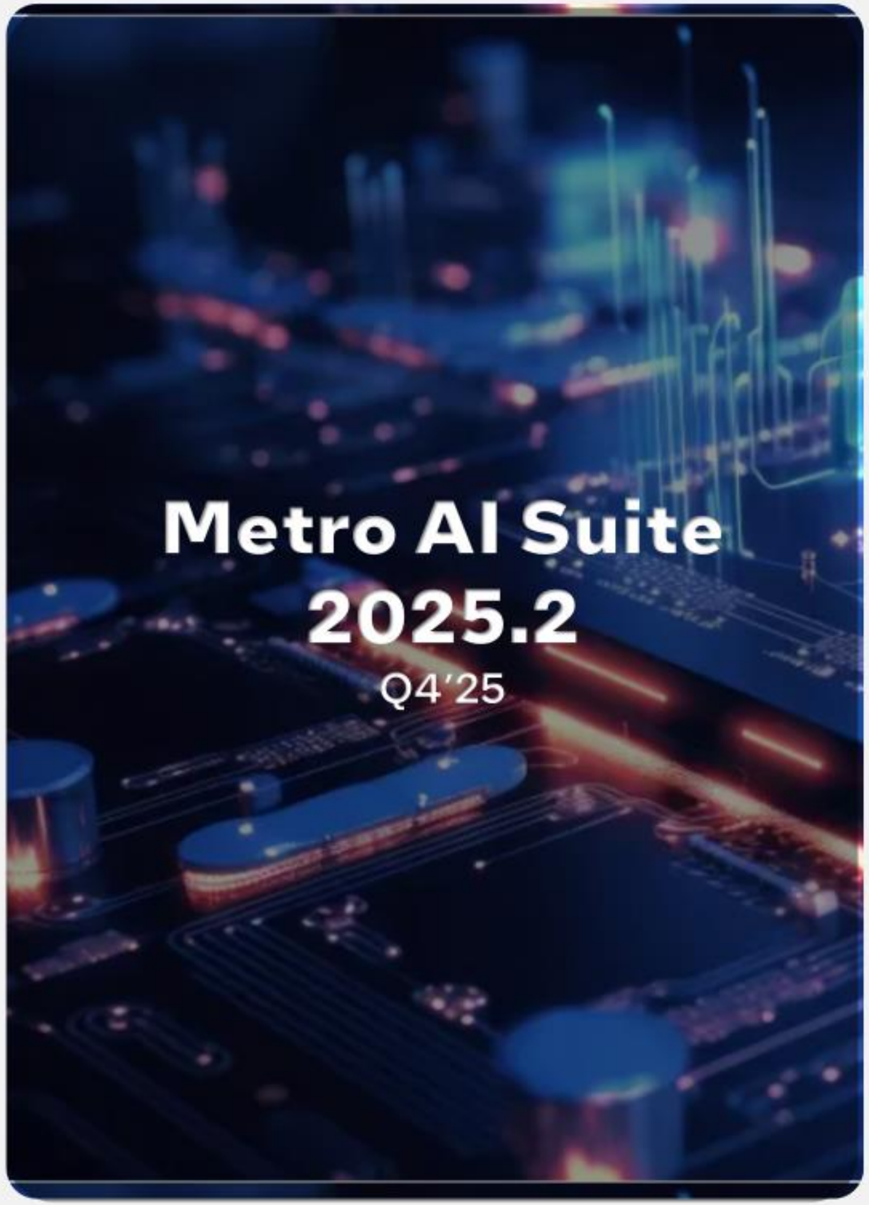
Use Case Tutorial

Smart Intersection: AI Cybersecurity

Use Case Tutorial

Agentic AI Transit Route System

Use Case Tutorial



Metro AI Suite 2025.2

Q4'25

SDK (Software Development Kit) Manager

NEW!

DL Streamer Pipeline Optimization Tool

NEW!

Visual Pipeline and Platform Evaluation Tool (ViPPET)

Improved Ease of Use

Smart NVR (Network Video Recorder)

NEW!

Metro SDK Overview

Metro AI Suite for Solution Development

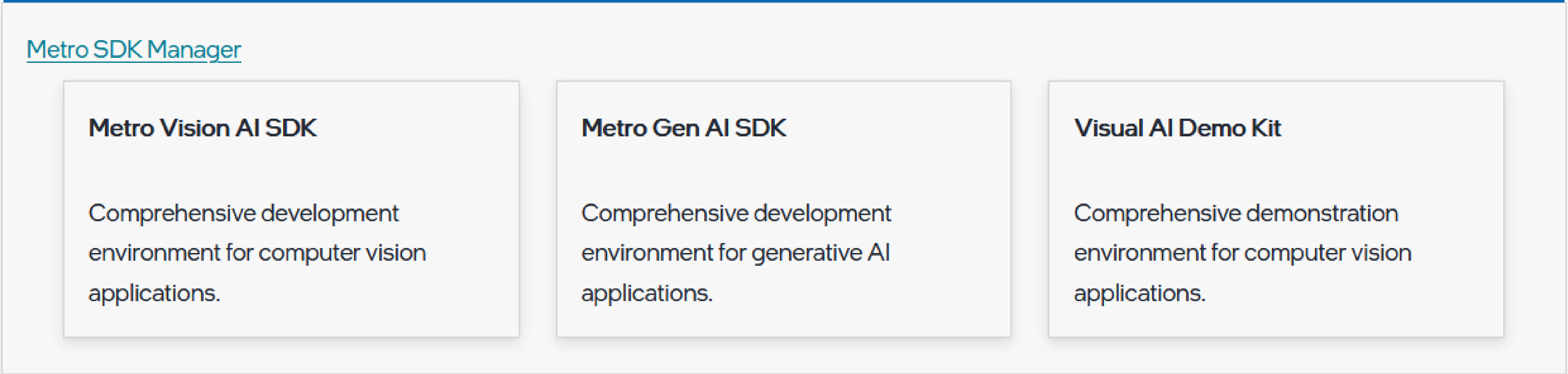
Metro SDKs

OpenVINO

oneAPI

- Microservices and Libraries
- Media, video analytics, and Gen AI
- Tutorials
- Includes selected 3rd party components

- Intel-optimized AI & Media performance
- Modular, Open, and Free libraries
- Internally validated



Metro SDK Manager

Metro Vision AI SDK

Comprehensive development environment for computer vision applications.

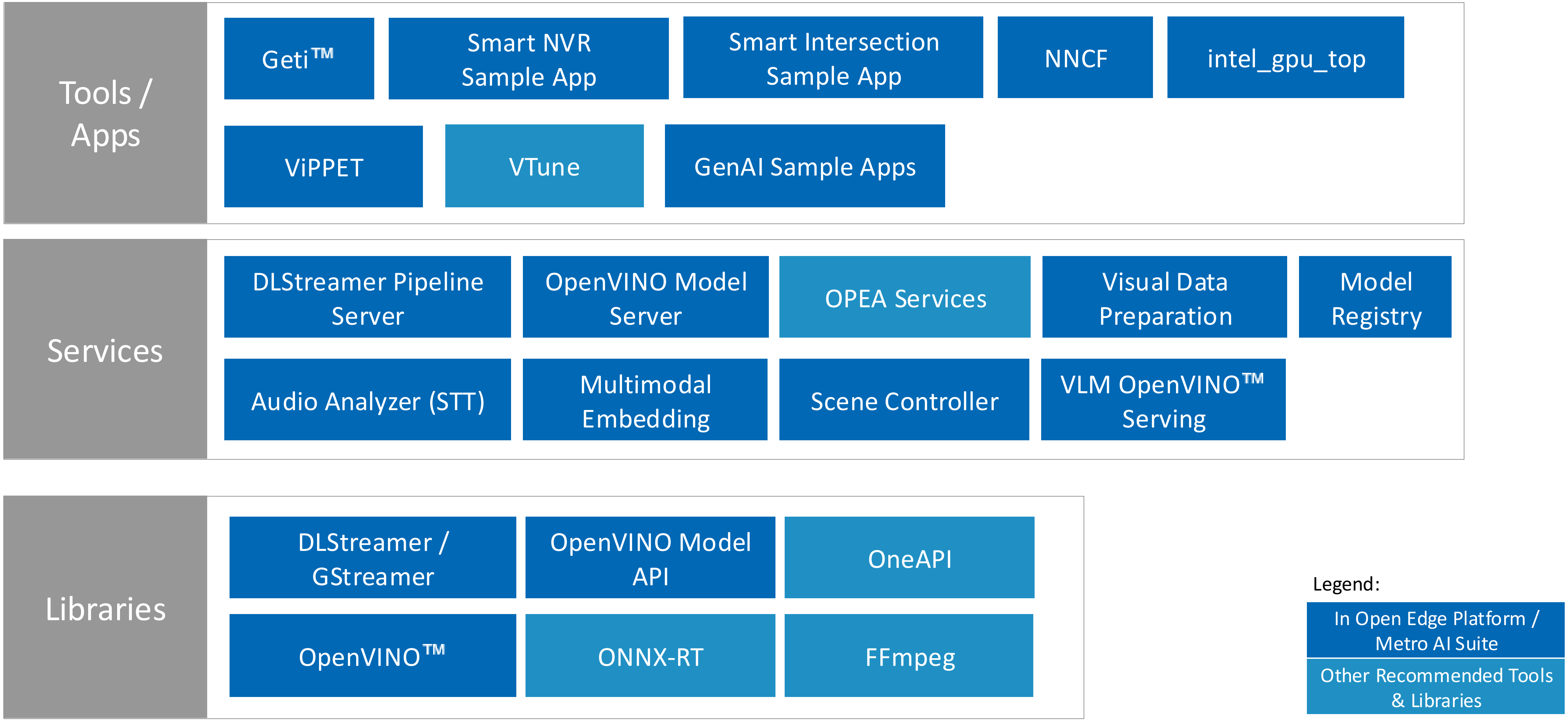
Metro Gen AI SDK

Comprehensive development environment for generative AI applications.

Visual AI Demo Kit

Comprehensive demonstration environment for computer vision applications.

Metro AI Suite Components



Tools / Apps

- Geti™
- Smart NVR Sample App
- Smart Intersection Sample App
- NNCF
- intel_gpu_top
- ViPPET
- VTune
- GenAI Sample Apps

Services

- DLStreamer Pipeline Server
- OpenVINO Model Server
- OPEA Services
- Visual Data Preparation
- Model Registry
- Audio Analyzer (STT)
- Multimodal Embedding
- Scene Controller
- VLM OpenVINO™ Serving

Libraries

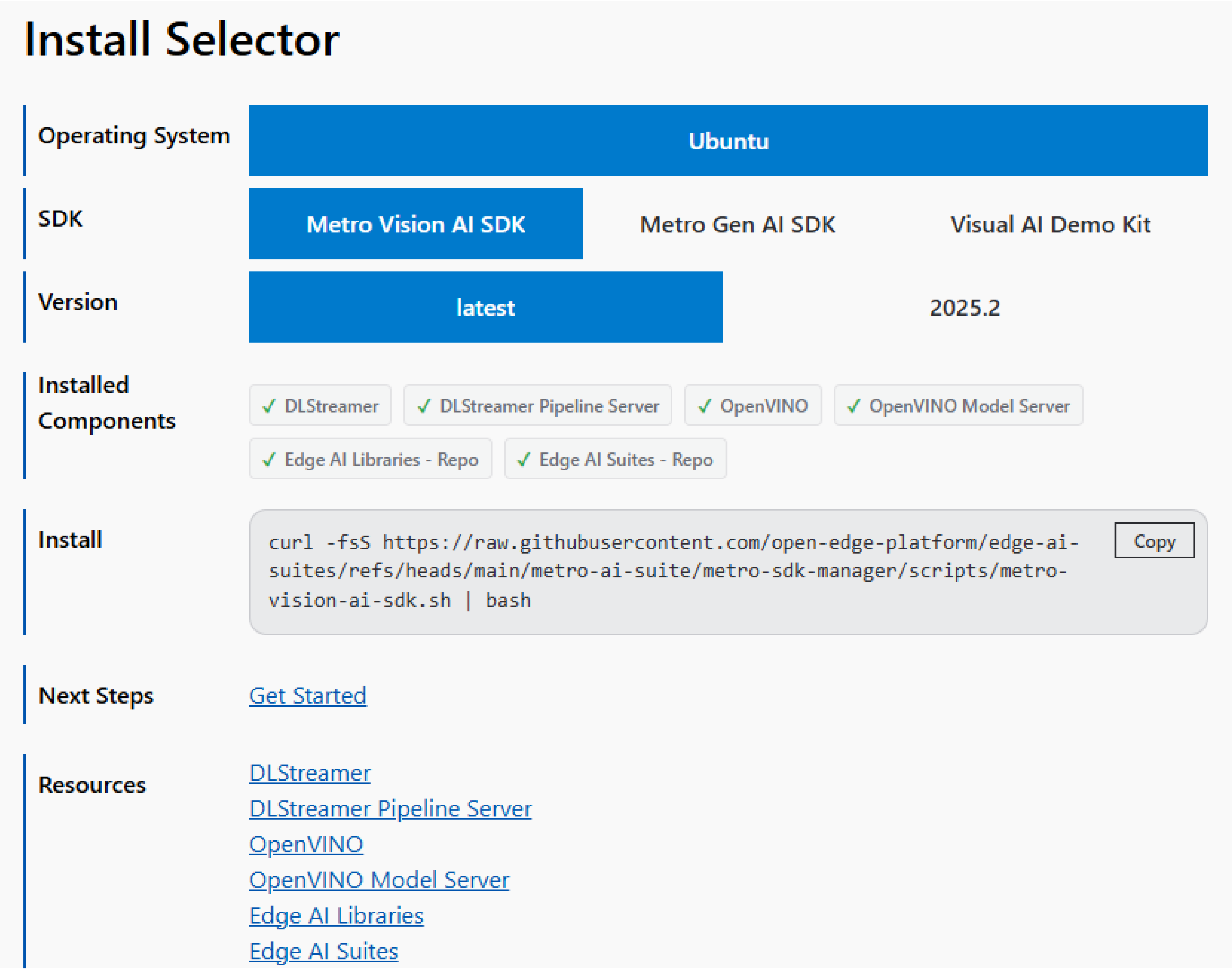
- DLStreamer / GStreamer
- OpenVINO Model API
- OneAPI
- OpenVINO™
- ONNX-RT
- FFmpeg

Legend:

- In Open Edge Platform / Metro AI Suite
- Other Recommended Tools & Libraries

New

Metro SDKs with easier download and installation



Install Selector

Operating System: **Ubuntu**

SDK: **Metro Vision AI SDK** | Metro Gen AI SDK | Visual AI Demo Kit

Version: **latest** | 2025.2

Installed Components: ✓ DLStreamer ✓ DLStreamer Pipeline Server ✓ OpenVINO ✓ OpenVINO Model Server ✓ Edge AI Libraries - Repo ✓ Edge AI Suites - Repo

Install:

```
curl -fsS https://raw.githubusercontent.com/open-edge-platform/edge-ai-suites/refs/heads/main/metro-ai-suite/metro-sdk-manager/scripts/metro-vision-ai-sdk.sh | bash
```

Next Steps: Get Started

Resources:
DLStreamer
DLStreamer Pipeline Server
OpenVINO
OpenVINO Model Server
Edge AI Libraries
Edge AI Suites

Metro SDK Installer

Streamlines discovery, installation, and management of multiple software development kits for edge AI applications with automated toolchain setup.

Key Features:

- Automated dependency resolution
  Eliminate manual SDK configuration complexity
- Version compatibility checking:
  Real-time validation across SDK versions
- Interactive wizard selection
  Guided process for optimal development setup

Download

New

Metro AI Suite SDK

Metro Vision AI SDK

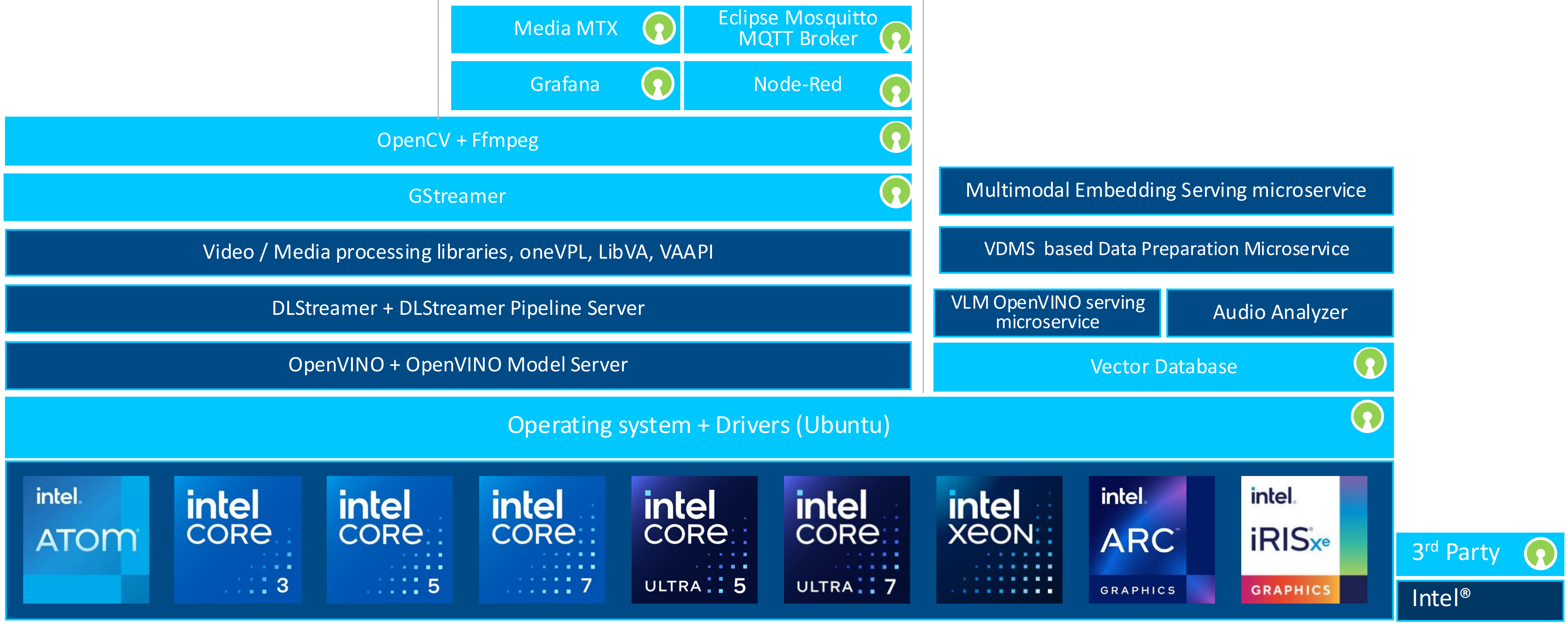
comprehensive development environment for computer vision

Visual AI Demo Kit

comprehensive demonstration environment for computer vision applications

Metro Gen AI SDK

comprehensive development environment for generative AI applications



Media MTX

Eclipse Mosquitto MQTT Broker

Grafana

Node-Red

OpenCV + Ffmpeg

GStreamer

Video / Media processing libraries, oneVPL, LibVA, VAAPI

DLStreamer + DLStreamer Pipeline Server

OpenVINO + OpenVINO Model Server

Multimodal Embedding Serving microservice

VDMS based Data Preparation Microservice

VLM OpenVINO serving microservice

Audio Analyzer

Vector Database

Operating system + Drivers (Ubuntu)

3rd Party

Intel®

Intel® AI – Optimized for the Edge

Conversion Guide

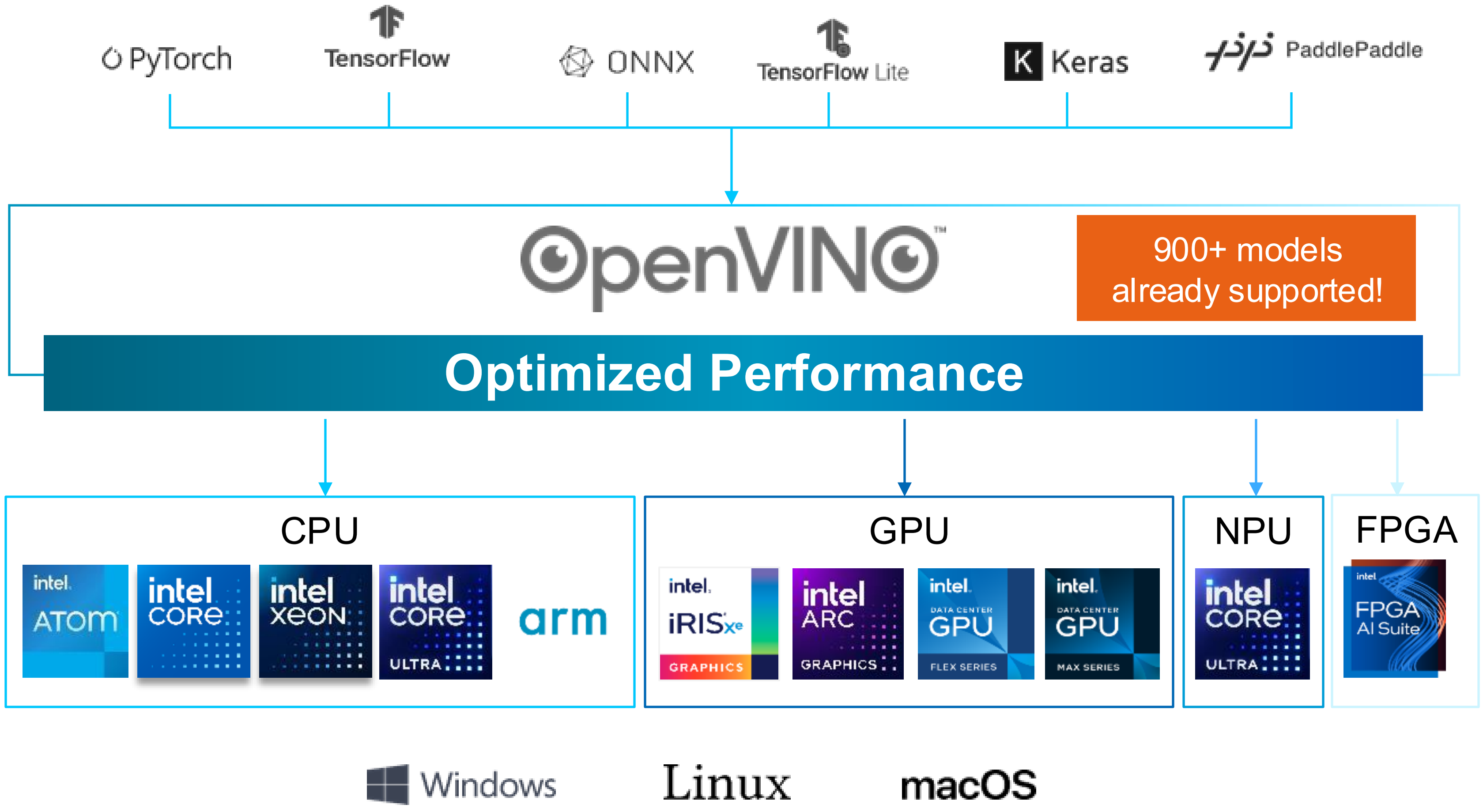
<table>
<tr><th></th><th colspan="2">INTEL
Optimized Performance on Intel Hardware</th><th colspan="2">NVIDIA
Optimized Performance on NVIDIA Hardware</th><th></th></tr>
<tr><td>Model Optimization
Model optimized for AI inference</td><td colspan="2">OpenVINO™</td><td colspan="2">TensorRT</td><td rowspan="5">ISV Applications/Solutions</td></tr>
<tr><td>GenAI
Simplified API</td><td colspan="2">OpenVINO™ GenAI</td><td colspan="2">TensorRT-LLM</td></tr>
<tr><td>Application Building
Using popular AI pipeline framework</td><td>Edge centric reference apps & microservices</td><td colspan="2">Create own application and/or use popular AI pipeline frameworks like LangChain or GStreamer</td><td>Pre-packaged pipelines via NIMs optimized for datacenter</td></tr>
<tr><td>Model Serving
Scalable inference serve</td><td colspan="2">OpenVINO™ Model Server</td><td colspan="2">Triton Inference Server</td></tr>
<tr><td>Model Streaming
Streaming media analytics framework</td><td colspan="2">DL Streamer™</td><td colspan="2">DeepStream</td></tr>
<tr><td>Low-level Programming</td><td colspan="2">SYCL/DPC++</td><td colspan="2">CUDA</td><td></td></tr>
</table>

Optimized AI Performance on Intel Hardware

1 Model

2 Optimize

3 Deploy



Benefits of Developing with OpenVINO™

Optimized for Intel

Unlocks key Intel hardware features for peak AI Inference performance on CPU, GPU, and NPU

Mature

- Tuned for various AI tasks like Computer Vision, GenAI, Agentic AI, ”AI Next”
- Widely adopted by hundreds of ISVs across AI PC, Edge and Cloud

Cost-Efficient

- High performance with accuracy and compact model size
- Flexible deployment via smart hardware optimization

Ease of Use

- CUDA-free adoption
- Simple APIs for inferencing
- “Write once deploy everywhere” flexibility

Ecosystem

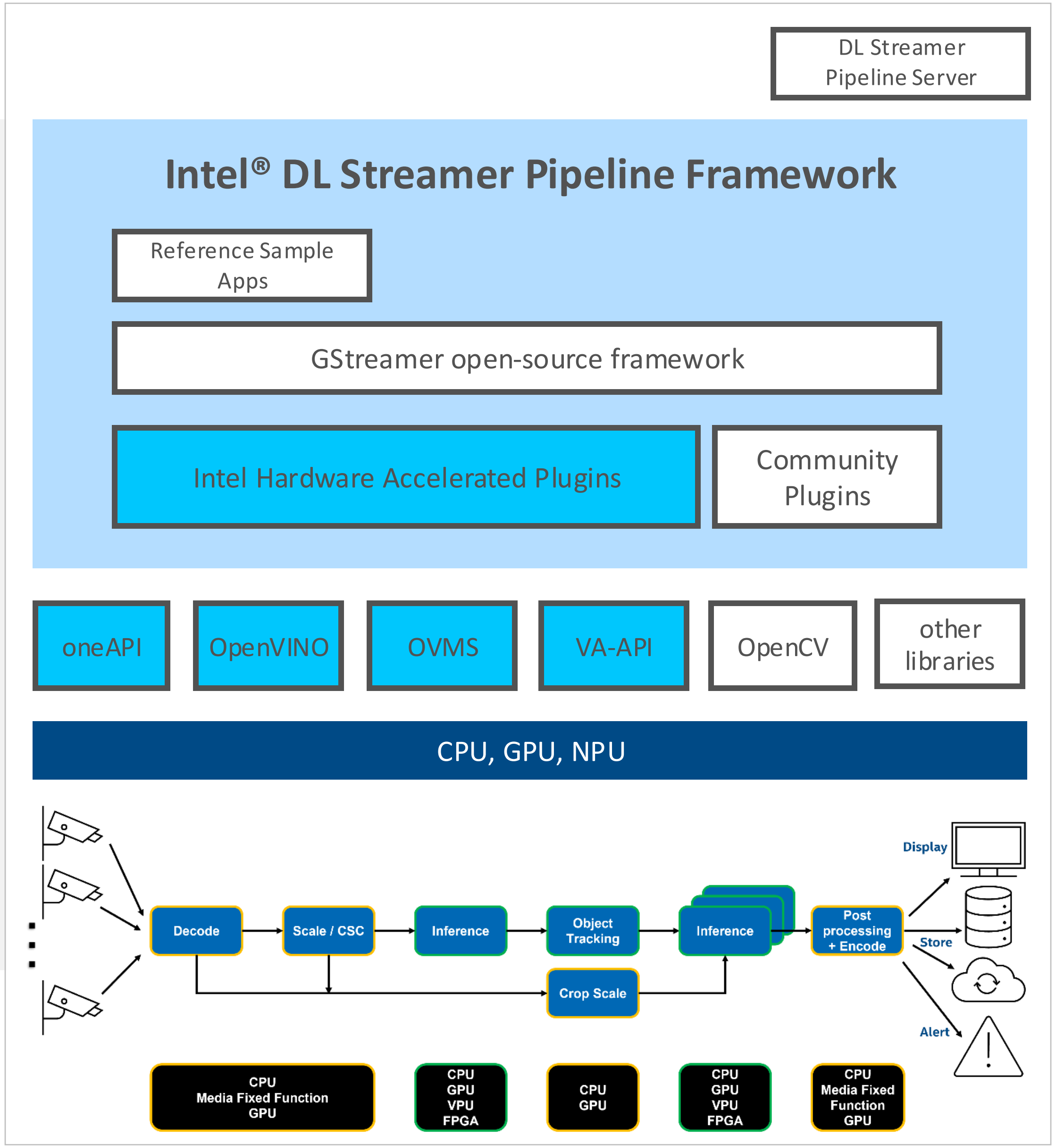
- Broad model coverage for 1,000+ options
- Ecosystem Integrations: ONNX Runtime, vLLM, LangChain, Hugging Face, Triton, and more

Deep Learning Streamer

What is DL Streamer?

- Deep Learning Streamer (DL Streamer) is an open-source streaming media analytics framework, based on the GStreamer multimedia framework.
- **Easily create complex media analytics pipelines** by linking pre-built media & analytic building blocks (Gstreamer Plugins) together
- **DL Streamer adds inline AI inference elements** (including metadata processing) to Gstreamer
- **DeepSORT tracks objects temporarily blocked from view** or out-of-frame for consistent detection

Write once, deploy on any Intel platforms, from Edge to Cloud



DL Streamer Pipeline Server

Intel® DL Streamer Pipeline Framework

Reference Sample Apps

GStreamer open-source framework

Intel Hardware Accelerated Plugins

Community Plugins

oneAPI

OpenVINO

OVMS

VA-API

OpenCV

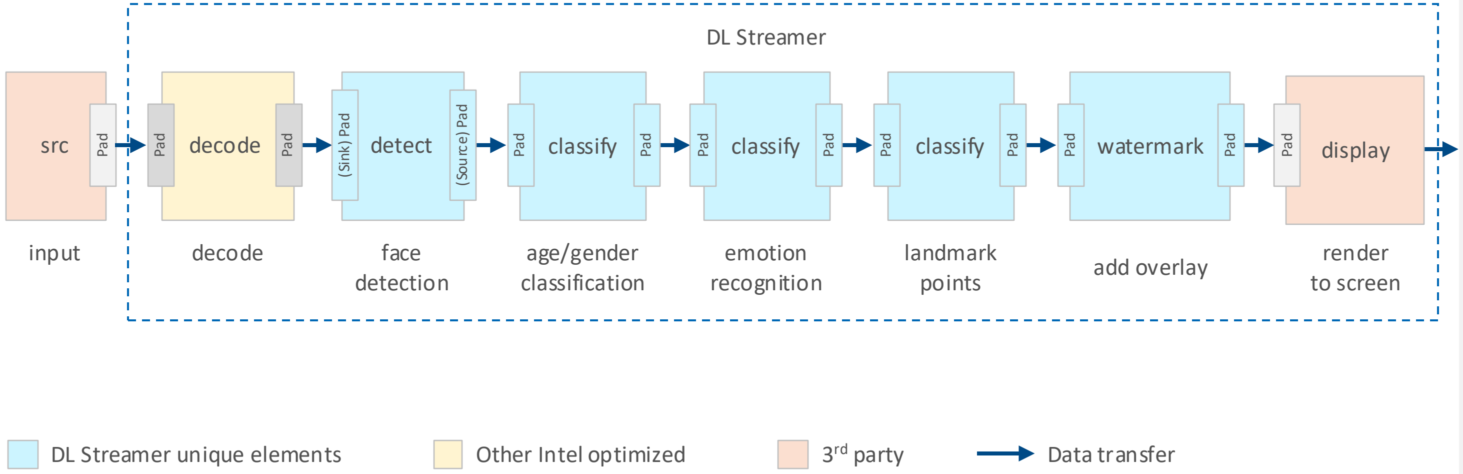
other libraries

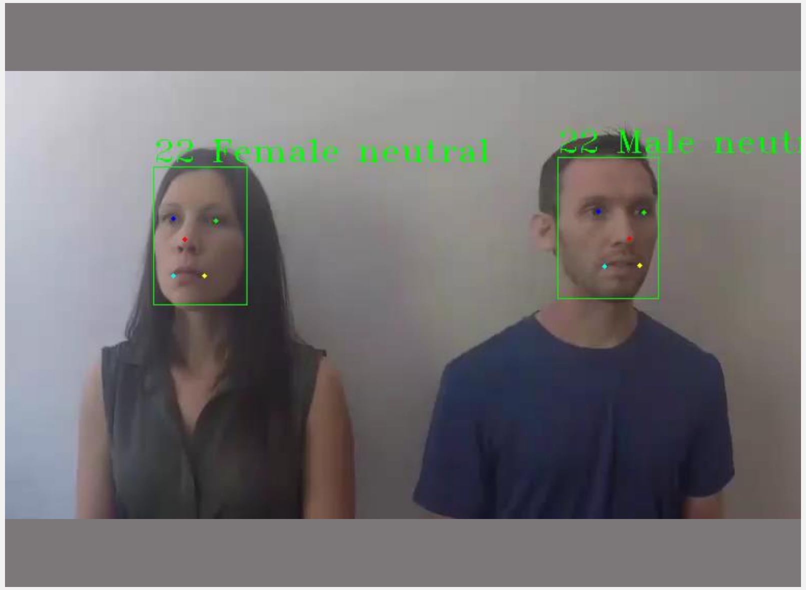
CPU, GPU, NPU

Deep Learning Streamer makes Media Analytics Easy

Example: Face detection, age, gender, emotion classification, overlay application

Reference End-to-End Pipeline Composition via DL Streamer





With only 8 lines of bash command line syntax:

```
$ gst-launch-1.0 filesrc location=/path/to/video.mp4 ! \
decodebin ! \
gvadetect model=face-detection-adas-0001.xml model-proc=face-detection-adas-0001.json ! queue ! \
gvaclassify model=age-gender-recognition.xml model-proc=age-gender-recognition.json ! queue ! \
gvaclassify model=emotions-recognition.xml model-proc=emotions-recognition.json ! queue ! \
gvaclassify model=landmarks-regression.xml model-proc=landmarks-regression.json ! queue ! \
gvawatermark ! \
ximagesink
```

Compared with ~1800 C++ lines of code for ‘interactive_face_detection_demo’ at https://github.com/openvinotoolkit/open_model_zoo

Porting Example

https://dlstreamer.github.io/dev_guide/converting_deepstream_to_dlstreamer.html

DeepStream

```
filesrc location=input_file.mp4 ! decodebin3 !

nvstreammux batch-size=1 width=1920 height=1080 ! queue !
nvinfer config-file-path=./config.txt !
nvvideoconvert ! “video/x-raw(memory:NVMM), format=RGBA” !

nvdsosd ! queue !

nvvideoconvert ! “video/x-raw, format=I420” !
videoconvert ! avenc_mpeg4 bitrate=8000000 !

qtmux ! filesink location=output_file.mp4
```

DeepStream Element	DLStreamer Element
nvinfer	gvadetect, gvaclassify, gvainference
nvdsosd	gvawatermark
nvtracker	gvatrack
nvmsgconv	gvametaconvert
nvmsgbroker	gvametapublish

DL Streamer

```
filesrc location=input_file.mp4 ! decodebin3 !

gvadetect model=./model.xml model-proc=./model_proc.json batch-size=1
! queue !

gvawatermark ! queue !

videoconvert ! avenc_mpeg4 bitrate=8000000 !

qtmux ! filesink location=output_file.mp4
```

DeepStream Element	GStreamer Element
nvvideoconvert	videoconvert
nvv412decoder	decodebin3
nvv412h264dec	vah264dec
nvv412h265dec	vah265dec
nvv412h264enc	va264enc
nvv412h265enc	va265enc

New

DL Streamer enables Vision Language Models

DL Streamer enables Vision Language Models (VLMs) through the gvagenai GStreamer element

Real-time video understanding using OpenVINO GenAI to generate text descriptions from video and text prompts

- Converts frames to embeddings (CLIP, VLM)
- Integrates with GenAI LLMs
- Transforms video processing into intelligent descriptions

Qwen/Qwen2.5-VL-3B-Instruct model (3B parameters)

New

DL Pipeline Optimizer

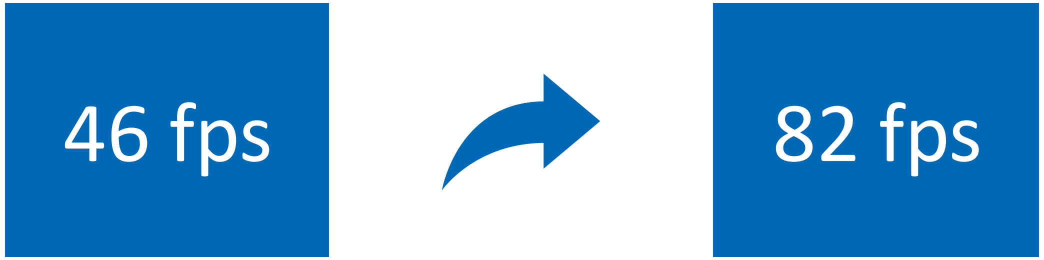
Automatically finds the optimal pipeline configuration to maximize FPS performance by intelligently testing parameters

The tool automatically tests many configuration combinations

Uses intelligent algorithms to find the optimal settings for maximum camera capacity on your existing hardware

Maximizes Frames per Seconds for the pipeline

```
FpsCounter(last 1.00sec): total=46.87 fps, number-streams=1, per-stream=46.87 fps
FpsCounter(average 1.00sec): total=46.87 fps, number-streams=1, per-stream=46.87 fps
FpsCounter(last 1.01sec): total=43.70 fps, number-streams=1, per-stream=43.70 fps
FpsCounter(average 2.01sec): total=45.28 fps, number-streams=1, per-stream=45.28 fps

...

FpsCounter(last 1.09sec): total=73.45 fps, number-streams=1, per-stream=73.45 fps
FpsCounter(average 8.70sec): total=73.65 fps, number-streams=1, per-stream=73.65 fps

[__main__] [    INFO] - Best found pipeline: urisourcebin buffer-size=4096
uri=https://videos.pexels.com/video-files/1192116/1192116-sd_640_360_30fps.mp4
!decodebin3!gvadetect model=/home/optimizer/models/public/yolo11s/INT8/yolo11s.xml device=GPU
pre-process-backend=va-surface-sharing batch-size=2 nireq=2  queue ! gvawatermark ! vah264enc !
h264parse ! mp4mux ! fakesink with fps: 81.987923.2
```



- replacing the decodebin with the decodebin3 element.
- configuring the gvadetect element to use GPU for processing
- setting the batch-size parameter to 2
- setting the nireq parameter to 2

DL Streamer™

Drives E2E Gen AI Performance on Intel® Processors

- AI inference optimization (OpenVINO™)
- Handles memory spaces efficiently, minimizing copies
- Inference batching across video streams
- Easily assigns pipeline operations to compute resources
- Easy setting of performance parameters
- Simple integration of optimized decode, encode, preprocessing

Metro AI Suite for Solution Development

Metro SDK

GeTi
OpenVINO
1 oneAPI

- Microservices
- Video / Media Libraries
- Visual Analytics Libraries

- Intel-optimized AI & Media performance
- Modular, Open, and Free libraries
- Internally validated

Sample Apps & Platform Blueprints

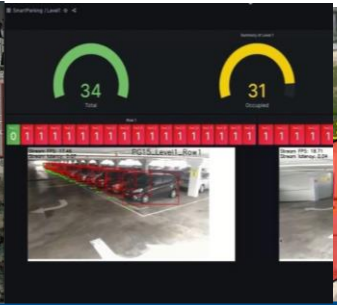
Smart Search	Visual Q&A	Loitering Detection
Search by Image	Smart Intersection	Platform Blueprints

- Quick and easy way to develop Visual and Gen AI features
- Use sample app code to apply Intel-optimized best practices
- Review documentation for scalable & heterogenous compute techniques

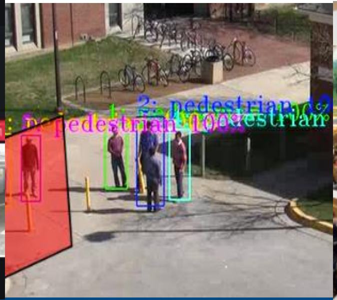
Reference Visual and Gen AI Sample Apps and Blueprints

Smart Intersection

Smart Parking

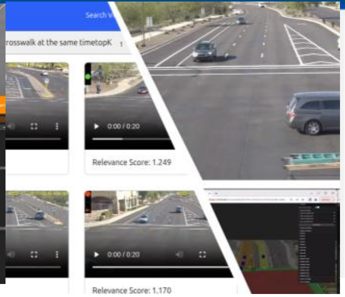


Loitering Detection

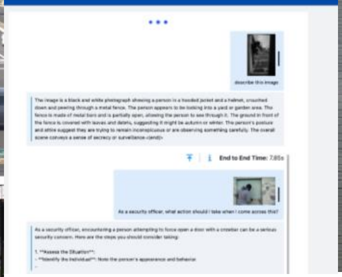


Agentic AI Tutorials

Reidentification (Image Based Video Search)



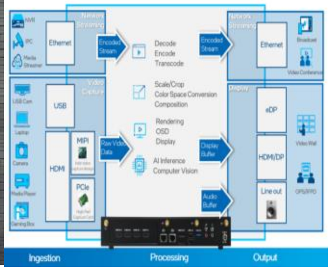
Video Search and Summarization



```
describe this image

The image is a black and white photograph showing a person in a hooded jacket and a helmet, crouched down and peering through a metal fence. The person appears to be looking into a yard or garden area. The fence is made of metal bars and is partially open, allowing the person to see through it. The ground in front of the fence is covered with leaves and debris, suggesting it might be autumn or winter. The person's posture and attire suggest they are trying to remain inconspicuous or are observing something carefully. The overall scene conveys a sense of secrecy or surveillance.<end>

End to End Time: 7.85s

As a security officer, what action should I take when I come across this?

As a security officer, encountering a person attempting to force open a door with a crowbar can be a serious security concern. Here are the steps you should consider taking:

1. **Assess the Situation**:
- **Identify the Individual**: Note the person's appearance and behavior.
```

Visual Q&A

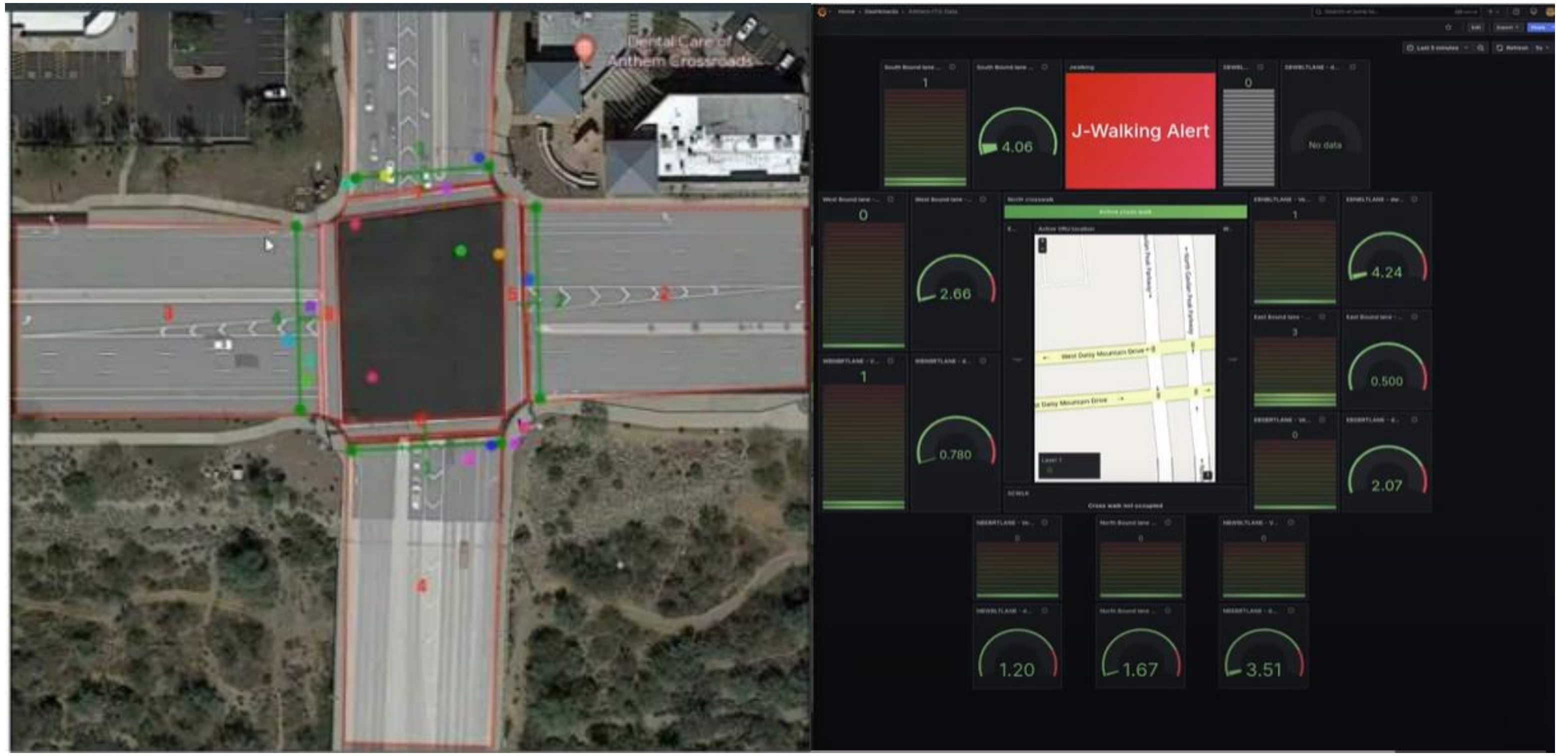
Smart NVR



Blueprints

Why Use AI Sample Apps?

- Understand & evaluate Intel platforms for Computer Vision & Gen AI use cases
- Helps developers streamline code or jumpstart development



Smart Intersection

Advanced traffic management via Edge AI, scene-based analytics.

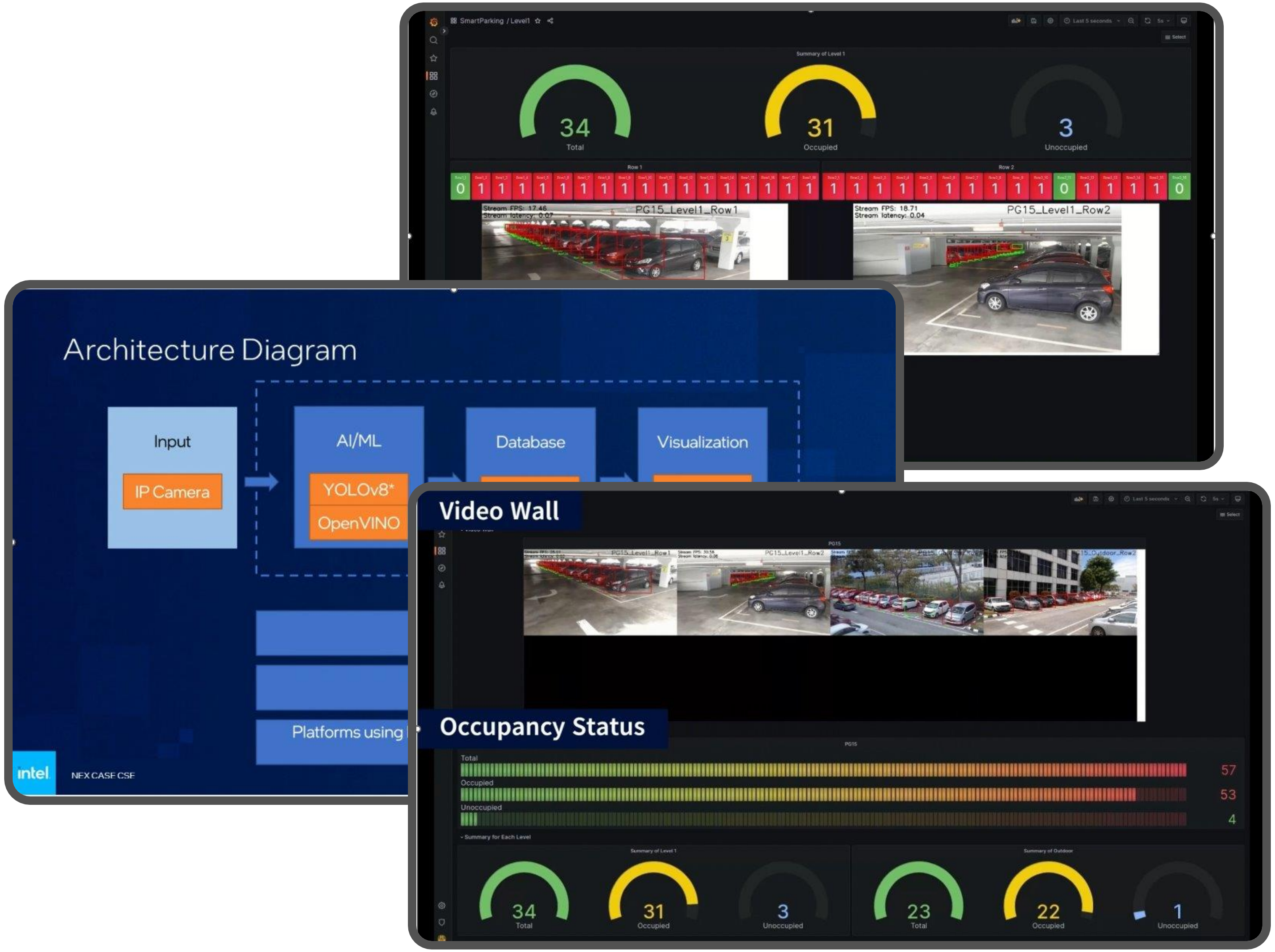
Key Features:

- Multi-sensor integration, including cameras, lidar, and radar help serve use cases like pedestrian safety and traffic analytics
- Scene-based / Unified analytics: Define regions of interest via an independent map view, simplifying multi-object tracking, motion vector analysis, and business logic across sensors
- Integration with MQTT, InfluxDB, Node-Red, and Grafana: Facilitates efficient message handling, near real-time monitoring, and insightful data visualization.
- Modular, microservice-based architecture (including DLStreamer) enables composability and reconfiguration

Supported Intel Platforms:

- Intel® Core^{TM}, Intel® Core^{TM} Ultra, Intel® Xeon® platforms

Download



Smart Parking

Effortlessly manage parking spaces with AI-driven video analytics for real-time insights and enhanced efficiency.

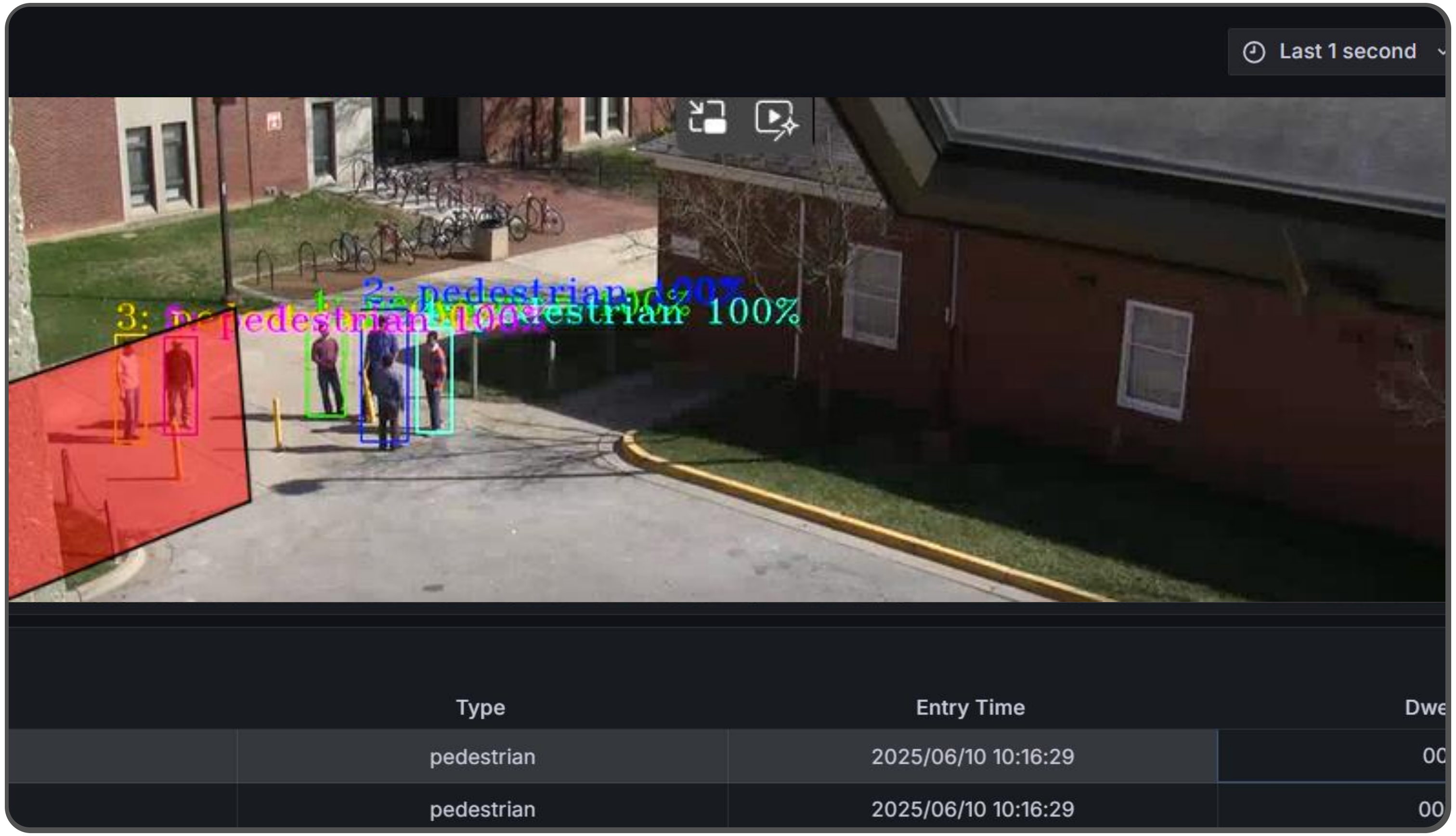
Key Features:

- Modular, microservice-based architecture (including Intel® DL Streamer)
- Vision Analytics Pipeline: Detect and classify objects using pre-configured AI models. Customize parameters (thresholds and object types) without requiring additional coding.
- Integration with MQTT, Node-RED, and Grafana: Facilitates efficient message handling, real-time monitoring, and insightful data visualization.
- User-Friendly: Simplifies configuration and operation through prebuilt scripts and configuration files.
- Made with Visual AI Demo Kit low-code app framework

Supported Intel Platforms:

- Intel® Core^{TM}, Intel® Core^{TM} Ultra, Intel® Xeon® platforms

Download



Type	Entry Time	Dwe
pedestrian	2025/06/10 10:16:29	00
pedestrian	2025/06/10 10:16:29	00

Loitering Detection

Loitering Detection leverages advanced AI algorithms to monitor and analyze real-time video feeds, identifying individuals lingering in designated areas.

Key Features:

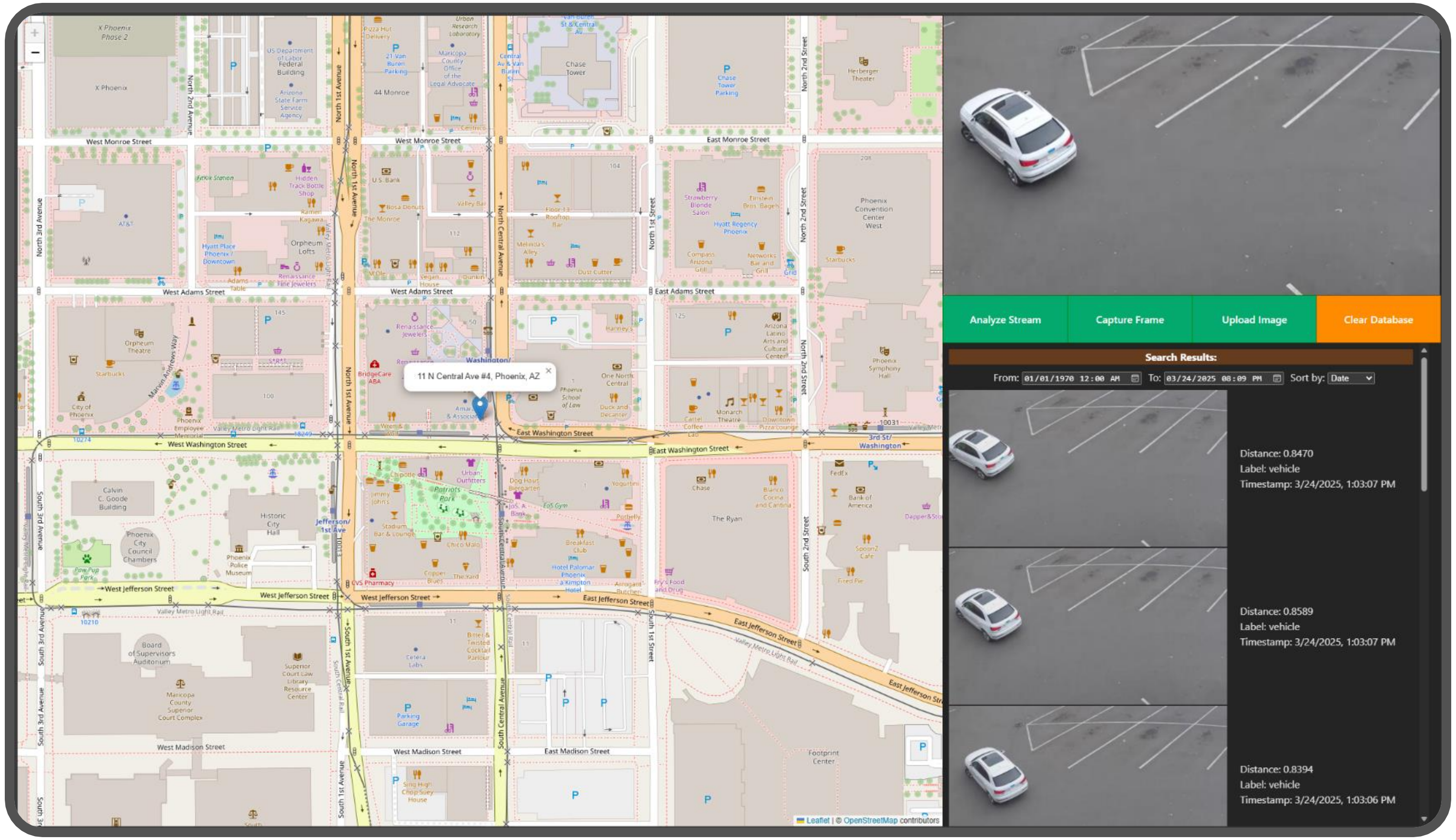
- Vision Analytics Pipeline: Detect and classify objects using pre-configured AI models. Customize parameters such as thresholds and object types without requiring additional coding.
- Integration with WebRTC Server, MQTT, Node-RED, and Grafana: Facilitates efficient message handling, real-time monitoring, and insightful data visualization.
- User-Friendly: Simplifies configuration and operation through prebuilt scripts and configuration files.
- Made with Visual AI Demo Kit low-code app framework

Supported Intel Platforms:

- Intel® Core™, Intel® Core™ Ultra, Intel® Xeon® platforms

Download

Reidentification (Image based Video Search)



Search Results:

From: 01/01/1970 12:00 AM To: 03/24/2025 08:09 PM Sort by: Date

Distance: 0.8470
Label: vehicle
Timestamp: 3/24/2025, 1:03:07 PM

Distance: 0.8589
Label: vehicle
Timestamp: 3/24/2025, 1:03:07 PM

Distance: 0.8394
Label: vehicle
Timestamp: 3/24/2025, 1:03:06 PM

The "Image-Based Video Search" sample application lets users search live or recorded camera feeds by providing an image and view matching objects with location, timestamp, and confidence score details

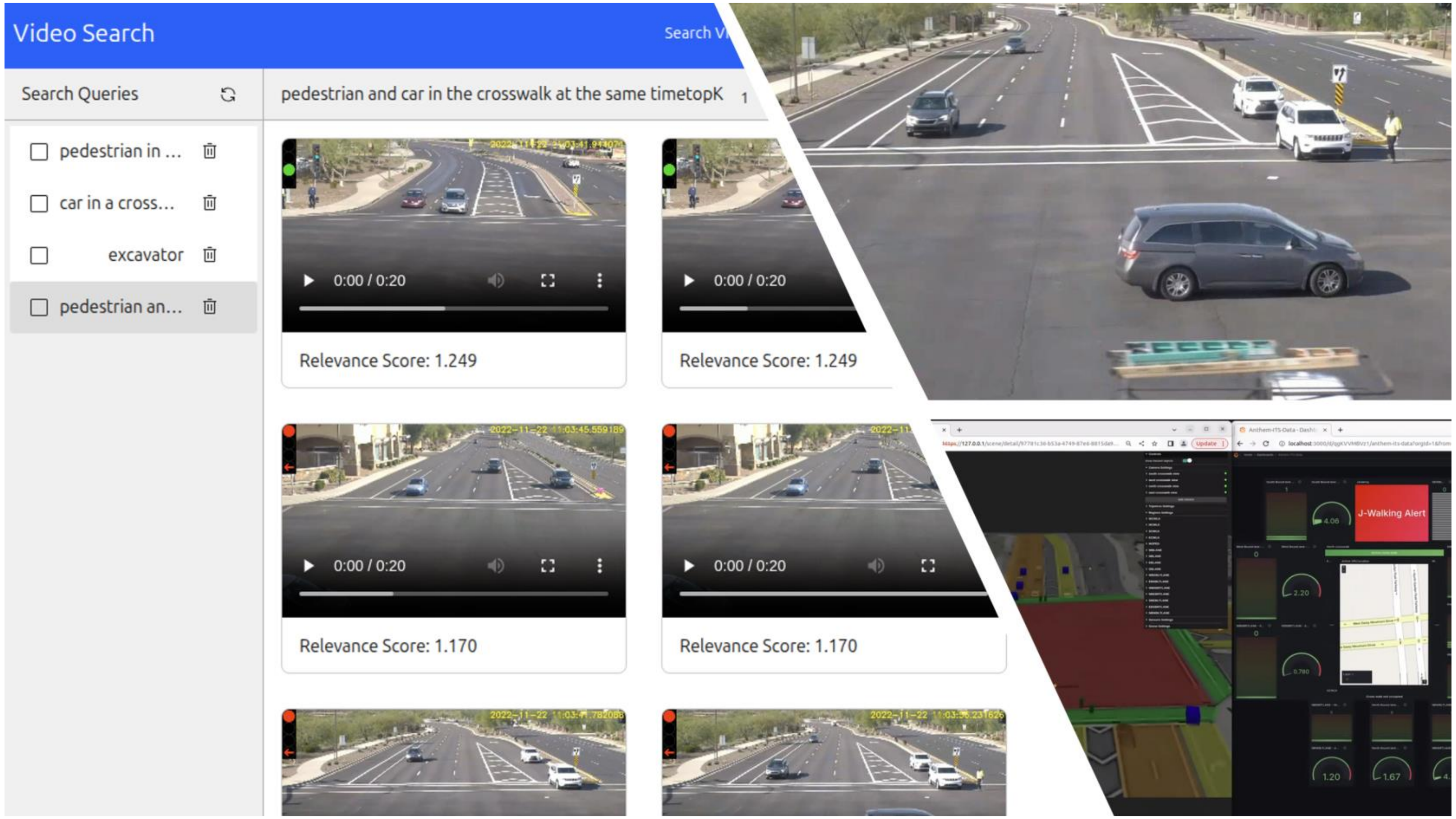
Key Features:

- Enables cross-camera tracking / re-identification – useful in both real-time and forensic investigations
- Shows how to combine edge AI microservices for video ingestion, object detection, feature extraction, and vector-based search.
- Integration with DLStreamer Pipeline Server, MediaMTX, MQTT, MilvusDB, ImageIngestor

Supported Intel Platforms:

- Intel® Core^{TM} and Intel® Core^{TM} Ultra platforms

Download



pedestrian and car in the crosswalk at the same timetopK 1

Relevance Score: 1.249

Relevance Score: 1.249

Relevance Score: 1.170

Relevance Score: 1.170

Video Search and Summarization

Video Search application leverages Generative AI tools to conduct comprehensive searches across vast video datasets, ensuring the extraction of key data points and making essential insights readily accessible. This technology enables identifying and highlighting sought-after information within the immense volume of video data in today's digital era.

Key Features:

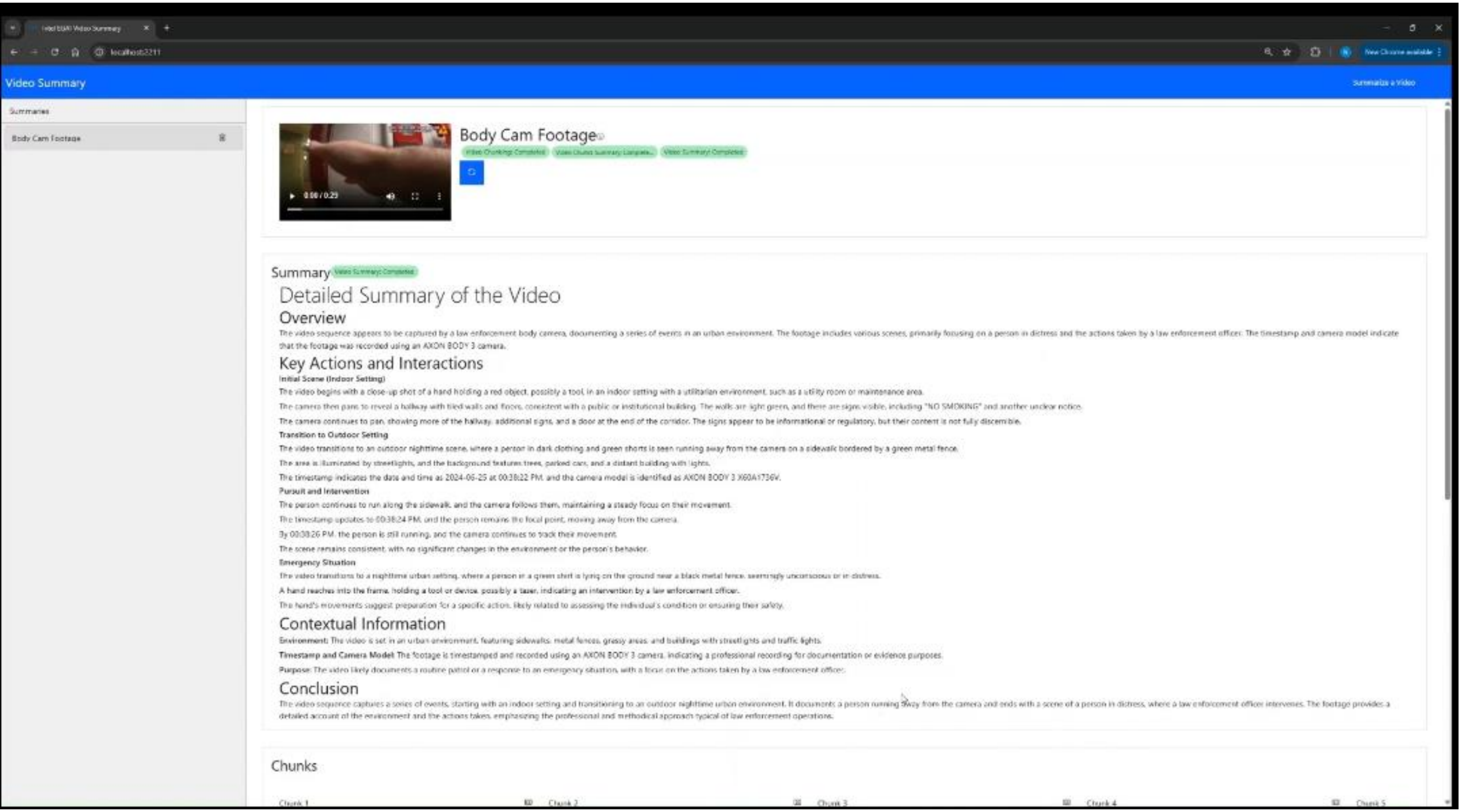
- Video Search: This functionality leverages LangChain, multimodal embedding models, and agentic reasoning to enable efficient and intelligent search over video content directly at the edge.

Supported Intel Platforms:

- Intel® Core^{TM}, Intel® Core^{TM} Ultra, Intel® Xeon® platforms, Intel® Arc^{TM} Graphics

Continued next page…

Download



Body Cam Footage

Summary

Detailed Summary of the Video

Overview

The video sequence appears to be captured by a law enforcement body camera, documenting a series of events in an urban environment. The footage includes various scenes, primarily focusing on a person in distress and the actions taken by a law enforcement officer. The timestamp and camera model indicate that the footage was recorded using an AXON BODY 3 camera.

Key Actions and Interactions

Initial Scene (Indoor Setting)

The video begins with a close-up shot of a hand holding a red object, possibly a tool, in an indoor setting with a utilitarian environment, such as a utility room or maintenance area.

The camera then pans to reveal a hallway with tiled walls and floors, consistent with a public or institutional building. The walls are light green, and there are signs visible, including "NO SMOKING" and another unclear notice.

The camera continues to pan, showing more of the hallway, additional signs, and a door at the end of the corridor. The signs appear to be informational or regulatory, but their content is not fully discernible.

Transition to Outdoor Setting

The video transitions to an outdoor nighttime scene, where a person in dark clothing and green shorts is seen running away from the camera on a sidewalk bordered by a green metal fence.

The area is illuminated by streetlights, and the background features trees, parked cars, and a distant building with lights.

The timestamp indicates the date and time as 2024-06-25 at 00:38:22 PM, and the camera model is identified as AXON BODY 3 X60A1736V.

Pursuit and Intervention

The person continues to run along the sidewalk, and the camera follows them, maintaining a steady focus on their movement.

The timestamp updates to 00:38:24 PM, and the person remains the focal point, moving away from the camera.

By 00:38:26 PM, the person is still running, and the camera continues to track their movement.

The scene remains consistent, with no significant changes in the environment or the person's behavior.

Emergency Situation

The video transitions to a nighttime urban setting, where a person in a green shirt is lyniq on the ground near a black metal fence, seemingly unconscious or in distress.

A hand reaches into the frame, holding a tool or device, possibly a taser, indicating an intervention by a law enforcement officer.

The hand's movements suggest preparation for a specific action, likely related to assessing the individual's condition or ensuring their safety.

Contextual Information

Environment: The video is set in an urban environment, featuring sidewalks, metal fences, grassy areas, and buildings with streetlights and traffic lights.

Timestamp and Camera Model: The footage is timestamped and recorded using an AXON BODY 3 camera, indicating a professional recording for documentation or evidence purposes.

Purpose: The video likely documents a routine patrol or a response to an emergency situation, with a focus on the actions taken by a law enforcement officer.

Conclusion

The video sequence captures a series of events, starting with an indoor setting and transitioning to an outdoor nighttime urban environment. It documents a person running away from the camera and ends with a scene of a person in distress, where a law enforcement officer intervenes. The footage provides a detailed account of the environment and the actions takes, emphasizing the professional and methodical approach typical of law enforcement operations.

Video Search and Summarization (Continued)

A developmental sample application that demonstrates summarization of video streams.

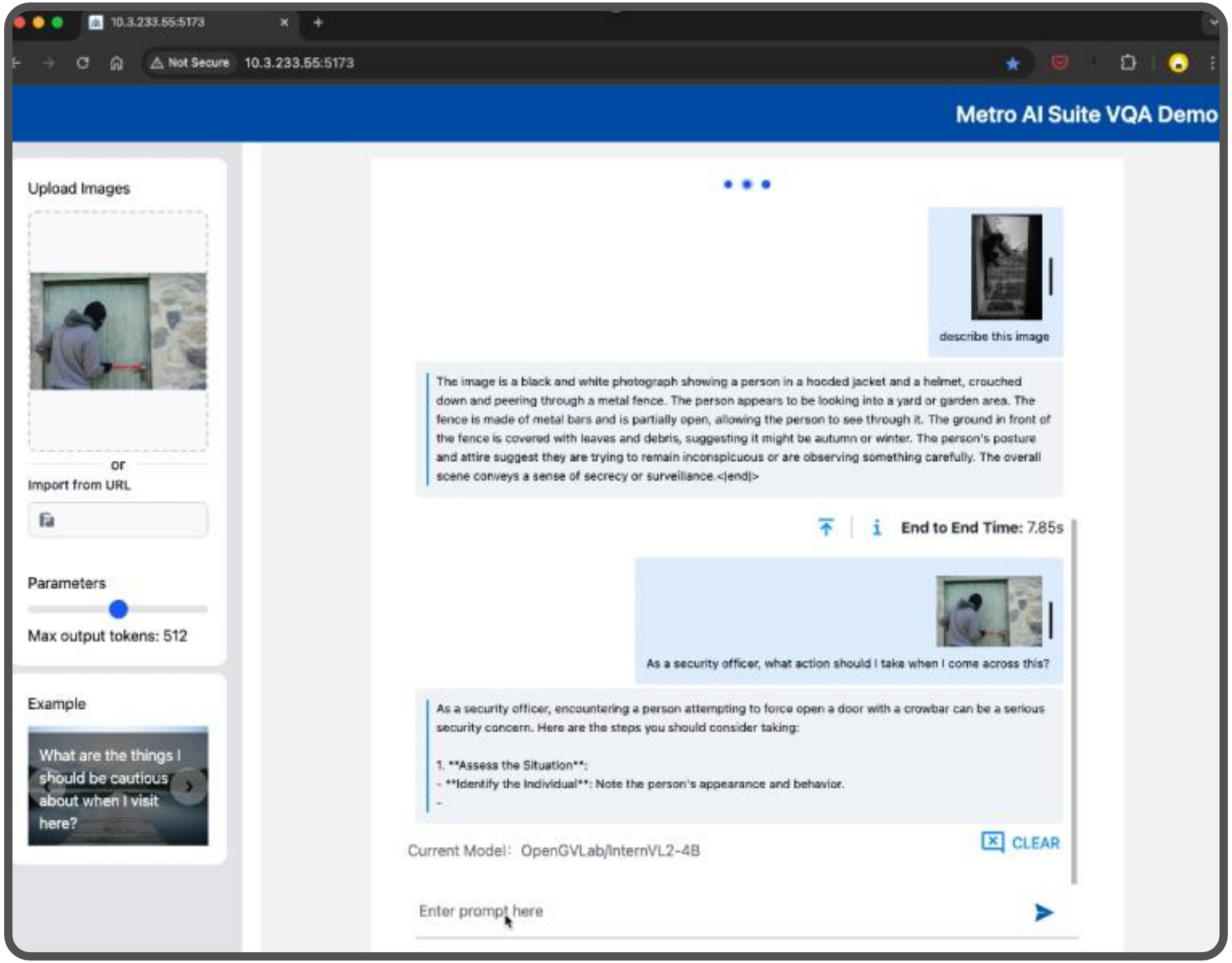
Key Features:

- Video Summarization: Using Vision Language Models (VLMs), Computer Vision, and Audio Analysis, the application distills key information into brief synopses from large volumes of data within long-form videos.

Supported Intel Platforms:

- Intel® Core^{TM}, Intel® Core^{TM} Ultra, Intel® Xeon® platforms, Intel® Arc^{TM} Graphics

Download



describe this image

The image is a black and white photograph showing a person in a hooded jacket and a helmet, crouched down and peering through a metal fence. The person appears to be looking into a yard or garden area. The fence is made of metal bars and is partially open, allowing the person to see through it. The ground in front of the fence is covered with leaves and debris, suggesting it might be autumn or winter. The person's posture and attire suggest they are trying to remain inconspicuous or are observing something carefully. The overall scene conveys a sense of secrecy or surveillance.<|end|>

End to End Time: 7.85s

As a security officer, what action should I take when I come across this?

As a security officer, encountering a person attempting to force open a door with a crowbar can be a serious security concern. Here are the steps you should consider taking:

1. **Assess the Situation**:
- **Identify the Individual**: Note the person's appearance and behavior.
-

Current Model: OpenGVLab/InternVL2-4B

Gen AI Based Video Q&A (VQ&A)

VQA (Video Question Answering) is the task of answering open-ended questions based on images. The input to models supporting this task is typically a combination of an image and a question, and the output is an answer expressed in natural language.

Key Features:

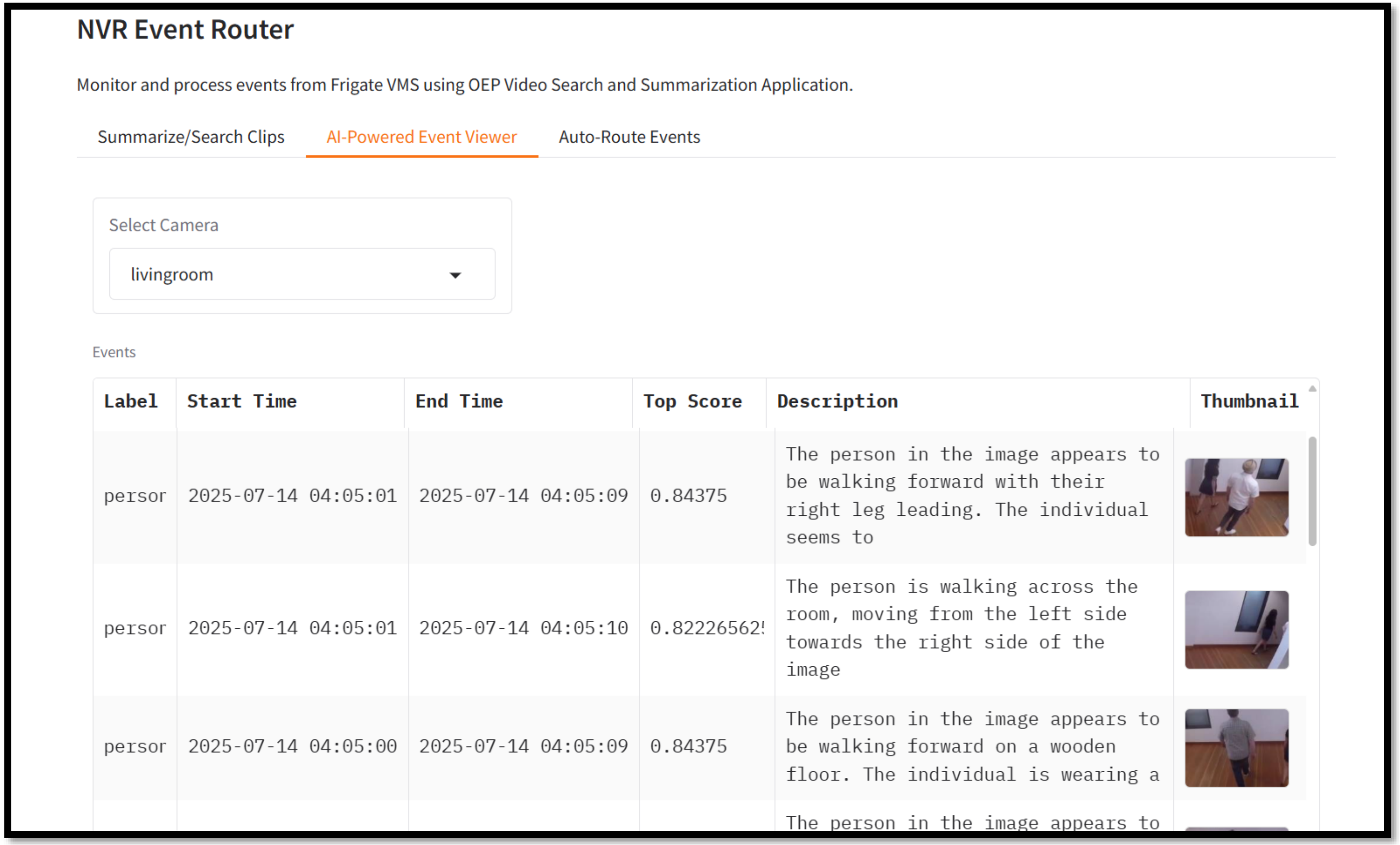
- VLM and LLM GenAI, including video summarization
- Integrates LVM Server, IPEX-LLM Server, & Web UI with chat
- Supported Models: Llava-1.5-7b, Qwen2-VL-7B-Instruct, InternVL2-4B

Supported Intel Platforms:

- Intel® Core^{TM} and Intel® Arc^{TM} A-series Graphics, such as A770 GPU

Download

New



NVR Event Router

Monitor and process events from Frigate VMS using OEP Video Search and Summarization Application.

Select Camera: livingroom

Events

Label	Start Time	End Time	Top Score	Description	Thumbnail
persor	2025-07-14 04:05:01	2025-07-14 04:05:09	0.84375	The person in the image appears to be walking forward with their right leg leading. The individual seems to	
persor	2025-07-14 04:05:01	2025-07-14 04:05:10	0.82226562!	The person is walking across the room, moving from the left side towards the right side of the image	
persor	2025-07-14 04:05:00	2025-07-14 04:05:09	0.84375	The person in the image appears to be walking forward on a wooden floor. The individual is wearing a	
				The person in the image appears to	

Smart NVR

Transforms traditional NVRs into intelligent, context-aware systems using GenAI-powered vision analytics to process video streams at the edge for real-time insights and automation.

Key Features:

- Edge-optimized analytics, Process video locally, reducing bandwidth and enabling faster response
- AI-powered summaries, Generate context-aware event descriptions using vision-language models
- Advanced video search, Find and access key video moments with smart embedding-based search.
- Real-time automated routing, Automatically process and route clips using custom rules
- Supported Models: Llava-1.5-7b, Qwen2-VL-7B-Instruct, InternVL2-4B

Supported Intel Platforms:

- Intel® Core™, Intel® Xeon® and Intel® Arc™ A-series Graphics, such as A770 GPU

Download

New

Agentic AI Tutorials : Learn how to design and deploy production-grade Agentic Visual AI on Intel edge hardware with step-by-step guidance and ready-to-run app recipes.

AI Cybersecurity
in Smart Intersection

Ensure data privacy

- Enable built-in security features with strong encryption protocols and blockchain.
- Safeguard infrastructure with advanced threat detection
- mitigate vulnerabilities across distributed edge devices.

https://docs.openedgeplatform.intel.com/dev/edge-ai-suites/smart-intersection/application-security-enablement.html

AI Route Planner

Find the best route in real time

- Fetch and analyze live feeds from traffic intersections
- Leverage scenes across multiple intersections simultaneously
- Use Gen AI, spatial intelligence, and SLAM libraries to map routes quickly.

https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite/smart-route-planning-agent

AI Crowd Analytics

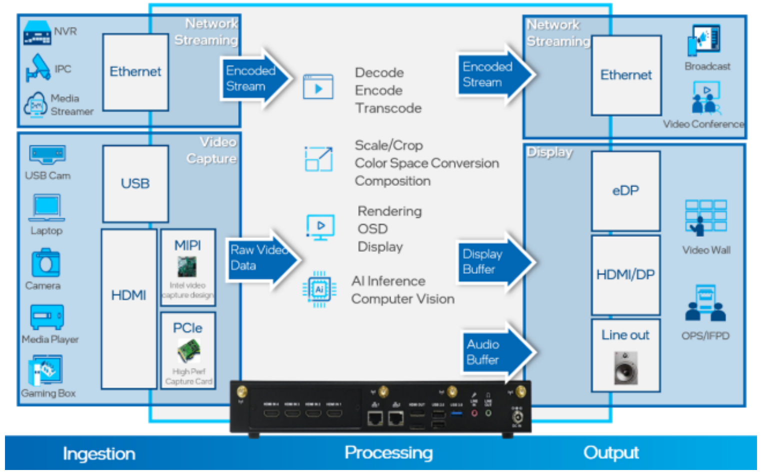
Customize for metro use cases

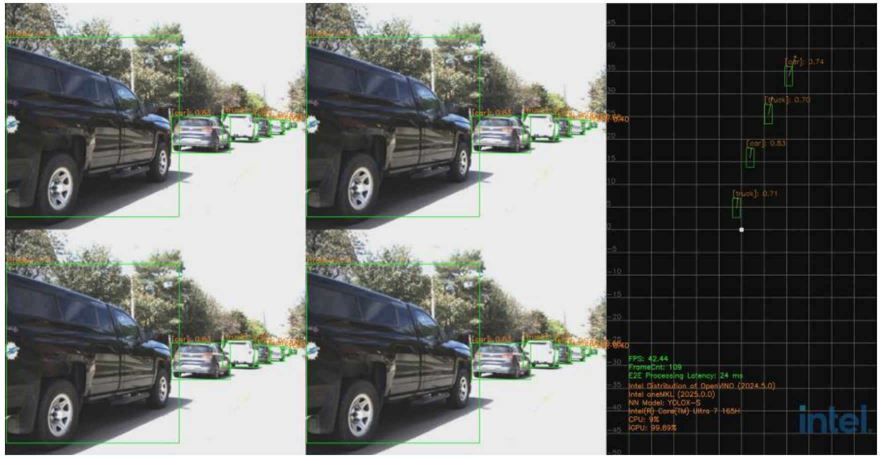
- Use vibe coding with Metro Vision AI App Recipe
- Use pre-configured AI models to detect and classify objects
- Develop customized crowd analytics with customized parameters

https://github.com/open-edge-platform/edge-ai-suites/blob/main/metro-ai-suite/metro-vision-ai-app-recipe/docs/user-guide/tutorial-4.md

Platform Blueprints

Intel® Video Processing Platform





```
FPS: 42.44
FrameCnt: 109
E2E Processing Latency: 24 ms
Intel Distribution of OpenVINO (2024.5.0)
Intel oneMKL (2025.0.0)
NN Model: YOLOX-S
Intel(R) Core(TM) Ultra 7 165H
CPU: 9%
iGPU: 99.69%
```

Platform Blueprints:

System software, middleware, and applications bundled to provide a starting point for building complete appliances / solutions:

- Video Processing Platform (VPP)
- Sensor Fusion for Traffic Management

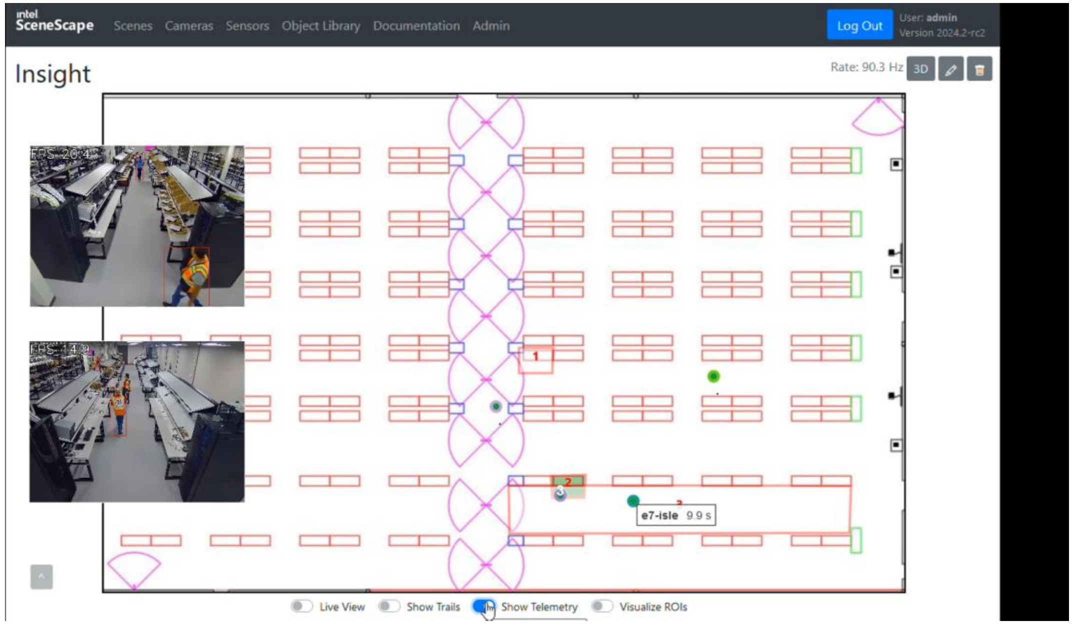
Spatial AI and Training

What is SceneScape?

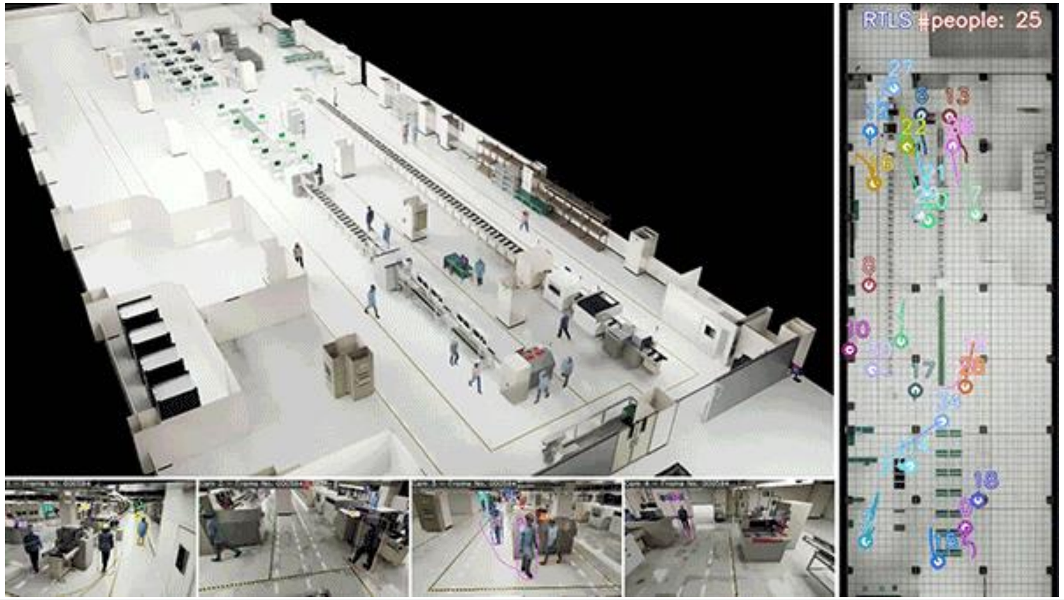
- SceneScape is a **software framework** that reaches beyond vision-based AI to realize spatial awareness from sensor data.
- It transforms data from many sensors to create and provide live updates to a 4-dimensional digital twin of a physical space.
- This spatial data can be applied to use cases including past analytics, tracking what is happening in the present, and making predictive decisions for the future.
- SceneScape brings physical AI to spaces like intersections, warehouses, stores, and campuses.

Track up to
100 concurrent objects at 15fps
on Intel® Core™ Ultra Processor Series 3

Intel® Scenescape:



The competition:



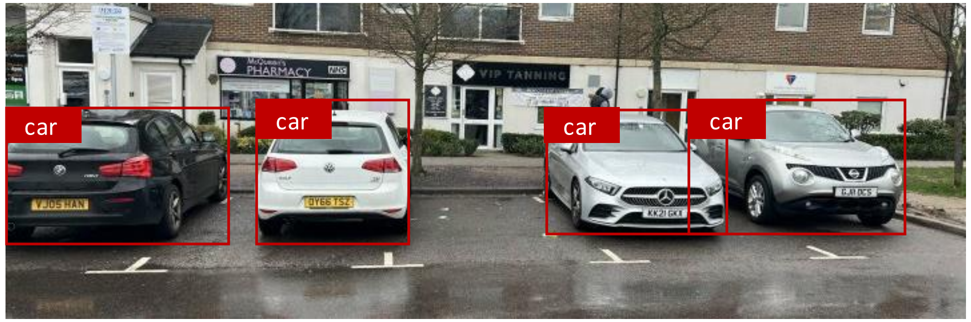
https://developer.nvidia.com/blog/optimize-processes-for-large-spaces-with-the-multi-camera-tracking-workflow/

Expanding single-camera use cases

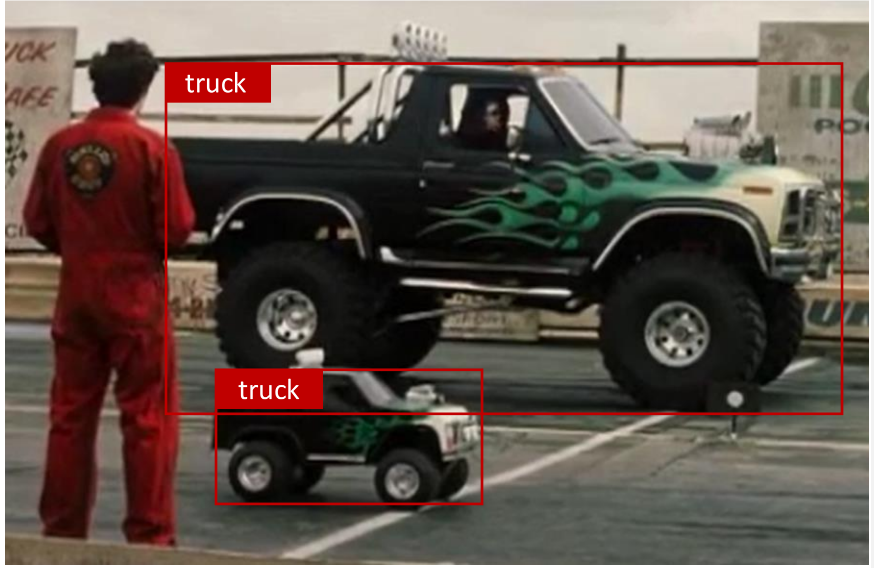
- Many use cases today are well served by analytics in a single camera using pixel-based bounding boxes
- However, fundamental limitations of pixel-based detections prevent critical use cases
  - Measuring size (in meters)
  - Determining speed (in meters per second)
  - Determining orientation and position

SceneScape™ solves these real-world problems, *even on single cameras*, by utilizing world coordinates and camera calibration

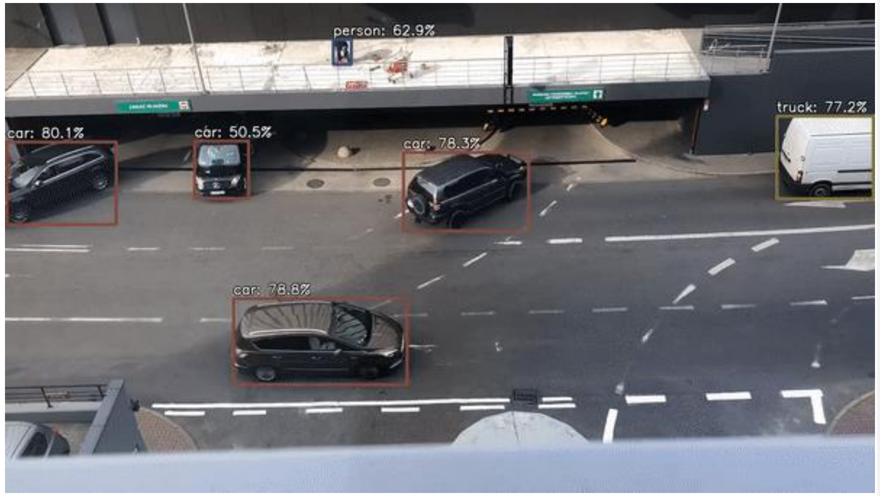
Which cars are parked properly?



Which truck is the toy?



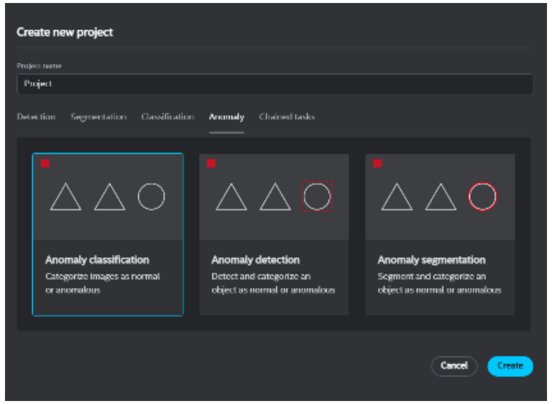
How fast is each car going?



Get Started with SceneScape

Geti™ : Powerful Vision AI for Everyone

From data input to optimization and model export, Geti™ creates vision AI models more efficiently



Create new project

Project name

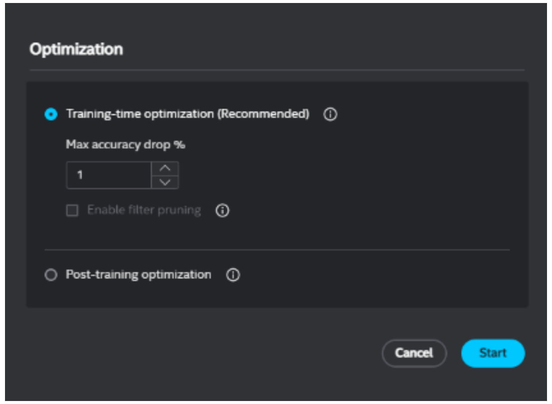
Project

Detection Segmentation Classification **Anomaly** Chained tasks

Anomaly classification
Categorize images as normal or anomalous

Anomaly detection
Detect and categorize an object as normal or anomalous

Anomaly segmentation
Segment and categorize an object as normal or anomalous



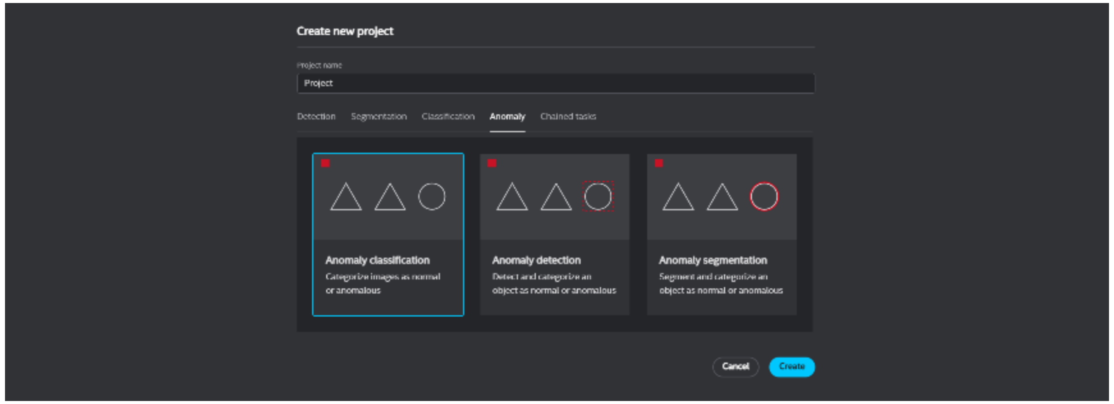
Optimization

● Training-time optimization (Recommended)

Max accuracy drop %: 1

□ Enable filter pruning

○ Post-training optimization



Create new project

Project name

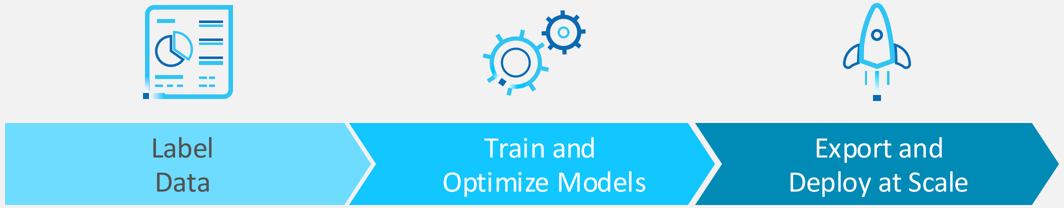
Project

Detection Segmentation Classification **Anomaly** Chained tasks

Anomaly classification
Categorize images as normal or anomalous

Anomaly detection
Detect and categorize an object as normal or anomalous

Anomaly segmentation
Segment and categorize an object as normal or anomalous



Label Data → Train and Optimize Models → Export and Deploy at Scale

Develop vision models with Geti™ :

- Smart Annotations – Expedite and simplify data labeling
- Active Learning – Build effective models with less data
- SDK Support for REST API – Simplify and automate the development pipeline

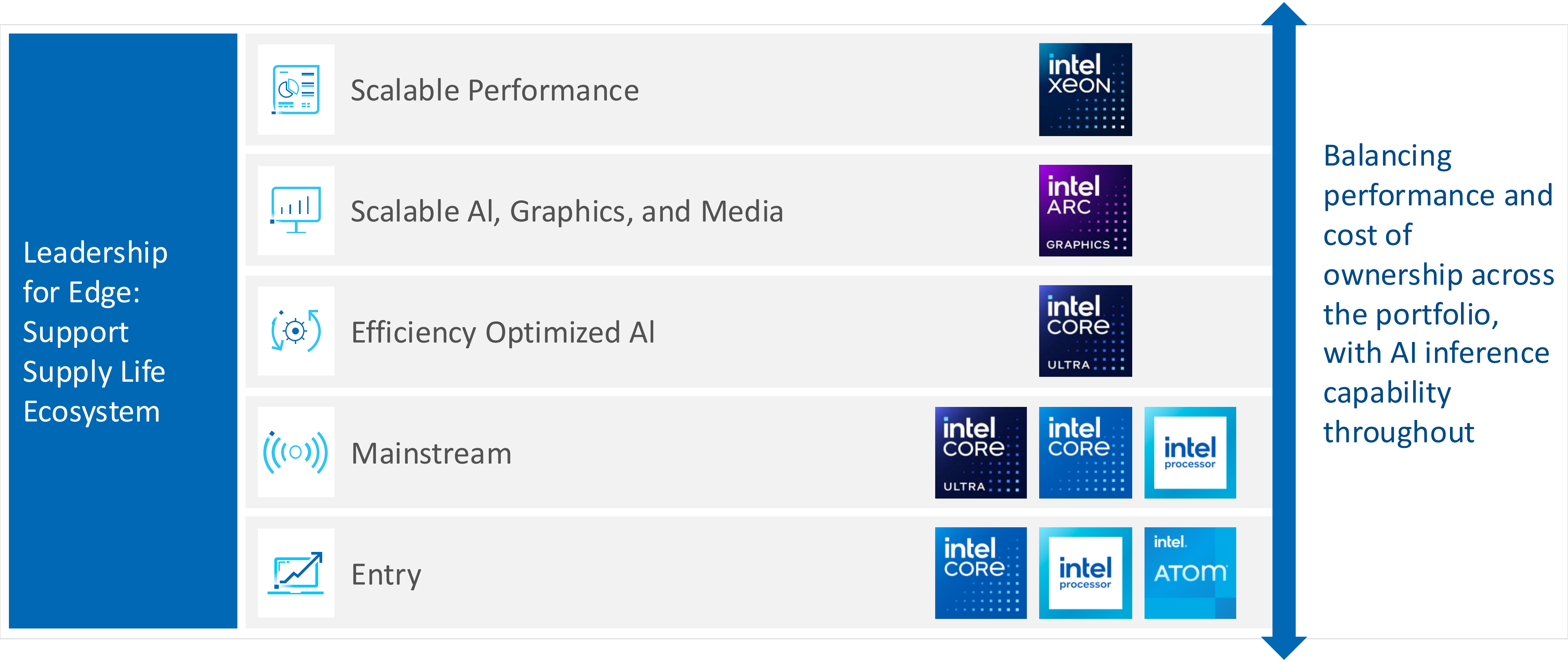
Scale AI solution deployment with OpenVINO™:

- Built-in OpenVINO™ optimizations – Easily optimize and quantize trained models with inferencing software that automatically detects available compute across Intel CPUs and GPUs

Recommended Systems

Intel Edge Product Portfolio

Intel’s Edge Roadmap Promise: We will deliver a comprehensive portfolio with best-in-class support, supply, ecosystem enablement, and performance per TCO across compute & AI



Leadership for Edge: Support Supply Life Ecosystem

Scalable Performance

Scalable AI, Graphics, and Media

Efficiency Optimized AI

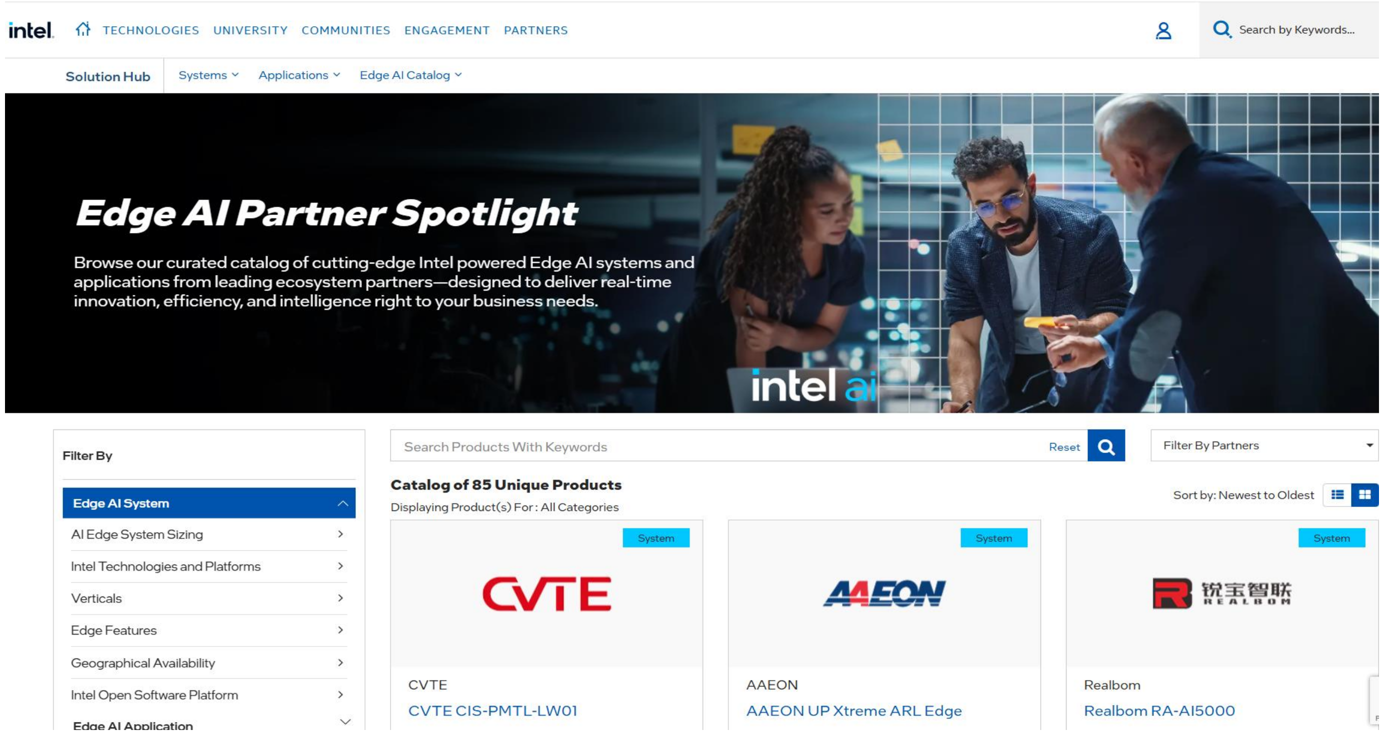
Mainstream

Entry

Balancing performance and cost of ownership across the portfolio, with AI inference capability throughout

Find Qualified Systems in the Recommended Hardware Catalog

Metro AI Suite **Recommended Hardware Catalog**



Edge AI Partner Spotlight

Browse our curated catalog of cutting-edge Intel powered Edge AI systems and applications from leading ecosystem partners—designed to deliver real-time innovation, efficiency, and intelligence right to your business needs.

Catalog of 85 Unique Products

Displaying Product(s) For : All Categories

CVTE
CVTE CIS-PMTL-LW01

AAEON
AAEON UP Xtreme ARL Edge

Realbom
Realbom RA-AI5000

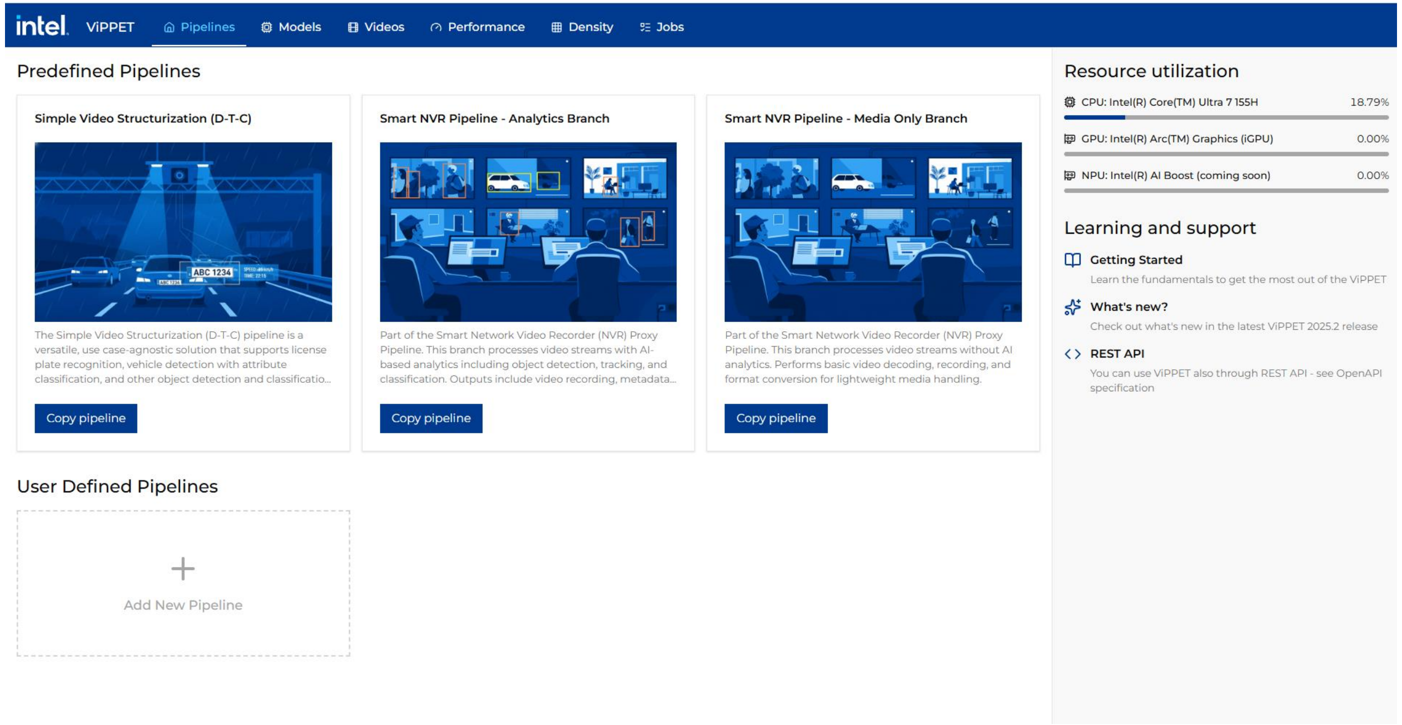
Catalog Benefits for Software Partners:

- Fast and easy way to see Intel AI-ready hardware options
- Only Metro AI Suite-qualified hardware listed
- Filter systems by processor type, industry, product application, geography, and more

Don’t see what you need in the Catalog? Contact us and we’ll help!

Easy End-to-End Visual AI Benchmarking with ViPPET

Visual Pipeline and Platform Evaluation Tool



Predefined Pipelines

Simple Video Structurization (D-T-C)

The Simple Video Structurization (D-T-C) pipeline is a versatile, use case-agnostic solution that supports license plate recognition, vehicle detection with attribute classification, and other object detection and classificatio...

Copy pipeline

Smart NVR Pipeline - Analytics Branch

Part of the Smart Network Video Recorder (NVR) Proxy Pipeline. This branch processes video streams with AI-based analytics including object detection, tracking, and classification. Outputs include video recording, metadata...

Copy pipeline

Smart NVR Pipeline - Media Only Branch

Part of the Smart Network Video Recorder (NVR) Proxy Pipeline. This branch processes video streams without AI analytics. Performs basic video decoding, recording, and format conversion for lightweight media handling.

Copy pipeline

User Defined Pipelines

+ Add New Pipeline

Resource utilization

CPU: Intel(R) Core(TM) Ultra 7 155H 18.79%

GPU: Intel(R) Arc(TM) Graphics (iGPU) 0.00%

NPU: Intel(R) AI Boost (coming soon) 0.00%

Learning and support

Getting Started
Learn the fundamentals to get the most out of the ViPPET

What's new?
Check out what's new in the latest ViPPET 2025.2 release

REST API
You can use ViPPET also through REST API - see OpenAPI specification

Why ViPPET?

Decision-Ready Insights

- Evaluate real-world Visual AI performance on Intel hardware
- Measure FPS & device utilization across full video + AI pipelines

Faster System Evaluation

- Quickly see how your workload scales across CPU / GPU / NPU.
- Compare platforms to choose the right hardware with confidence.

Performance Optimization

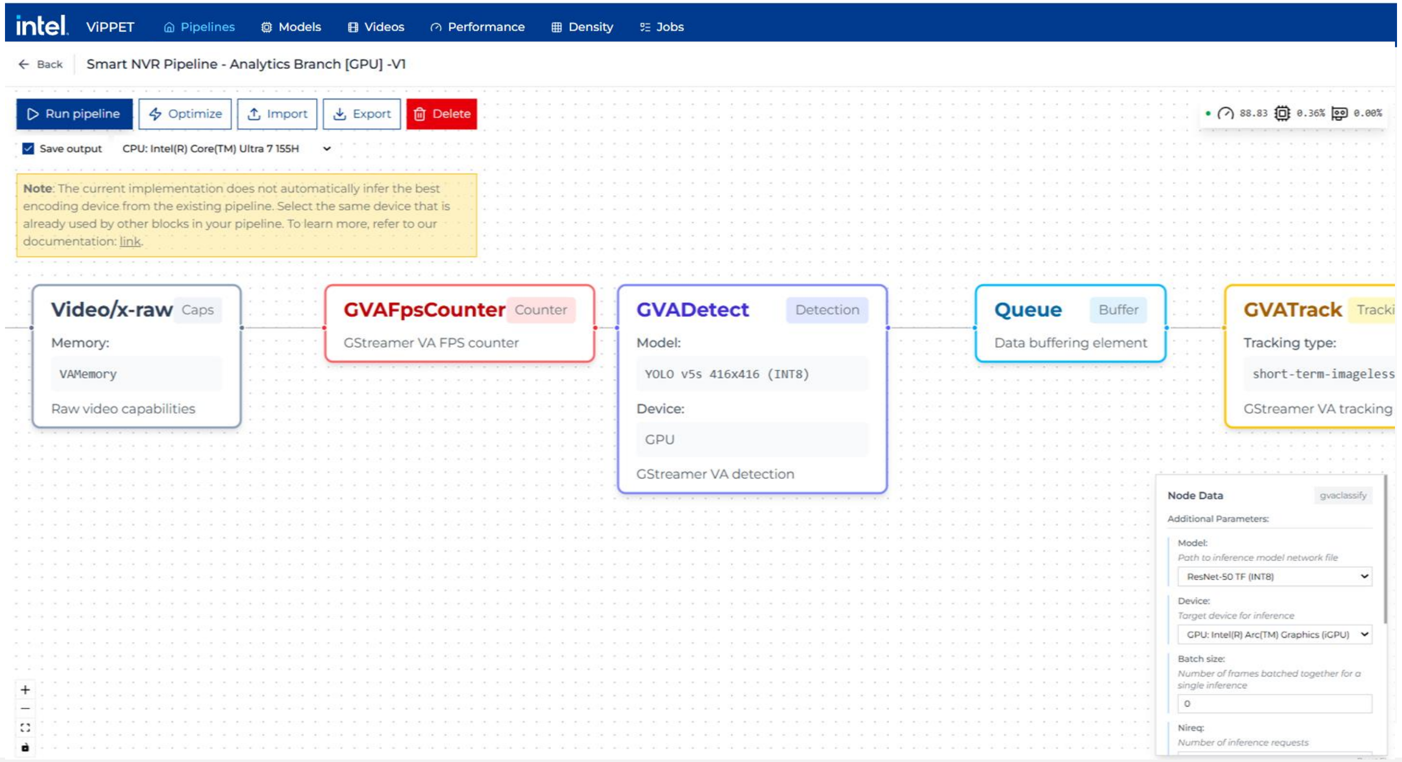
- Experiment with models and parameters
- Map individual pipeline stages — or the whole chain — to CPU, GPU, or NPU
- Identify the highest-throughput configuration for your solution with the built-in optimizer.

Predictable Deployment

- Use proxy pipelines that mirror production workloads.
- Reduce guesswork and optimize TCO early in development.

Easy End-to-End Visual AI Benchmarking with ViPPET

Visual Pipeline and Platform Evaluation Tool



Smart NVR Pipeline - Analytics Branch [GPU] -V1

☑ Save output CPU: Intel(R) Core(TM) Ultra 7 155H

Note: The current implementation does not automatically infer the best encoding device from the existing pipeline. Select the same device that is already used by other blocks in your pipeline. To learn more, refer to our documentation: link.

Video/x-raw (Caps) — Memory: VAMemory — Raw video capabilities

GVAFpsCounter (Counter) — GStreamer VA FPS counter

GVADetect (Detection) — Model: YOLO v5s 416x416 (INT8) — Device: GPU — GStreamer VA detection

Queue (Buffer) — Data buffering element

GVATrack (Tracki) — Tracking type: short-term-imageless — GStreamer VA tracking

Node Data (gvaclassify)

Additional Parameters:

Model: *Path to inference model network file* — ResNet-50 TF (INT8)

Device: *Target device for inference* — GPU: Intel(R) Arc(TM) Graphics (iGPU)

Batch size: *Number of frames batched together for a single inference* — 0

Nireq: *Number of inference requests*

Getting Started 🚀

Select a ready-made pipeline and run it on your device in just a few clicks to assess your Intel hardware's end-to-end Visual AI capability.

Advanced Usage

Select a ready-made pipeline or bring your own using a DLStreamer pipeline string

Configure the pipeline by choosing or swapping proxy models, videos, devices, and adjusting step-specific parameters such as batch size, parallel inference requests, inference interval for different stages.

Rerun and compare pipelines to benchmark detailed end-to-end (E2E) performance across different configurations.

Track and analyze results in the Jobs section to understand system capacity for Visual AI workloads.

Partner Support Programs

CASE Enabling & Support Model

Enabling & Support Level	Intel Support Scope	Support Tools & Channels	Support Examples
Intel® Premium Enabling Services	▪ Per SLA/ TSoW – Custom, Customer Specific ▪ Custom POC/ Demo/ RI ▪ Onsite Support ▪ Priority on Early Access ▪ Pre-launch Preview	Intel Premier Support Portal (IPS) & Email & Phone & Meetings & As defined in TSOW	“Help me build traffic monitor AI solution on Intel SW & HW”
Standard Ticket Support	▪ Dedicated AE, NDA collateral ▪ Ticket based case ▪ Reference Implementation ▪ Training/ Workshop, Remote Access ▪ Time bound, SLA based on ticket priority	Intel Premier Support Portal (IPS) & Email	“Intel SW returns error while my SW is calling Intel stack”
Public Community Support	▪ Post Launch Support ▪ Community support ▪ Self Enabling Tool ▪ Public & limited NDA Collateral ▪ How-to Video ▪ Webinar	Community Forum/ Service Cloud/ GitHub/ CSDN/ 51Openlab/ Stack Overflow.	“How to install recent released Intel software?”

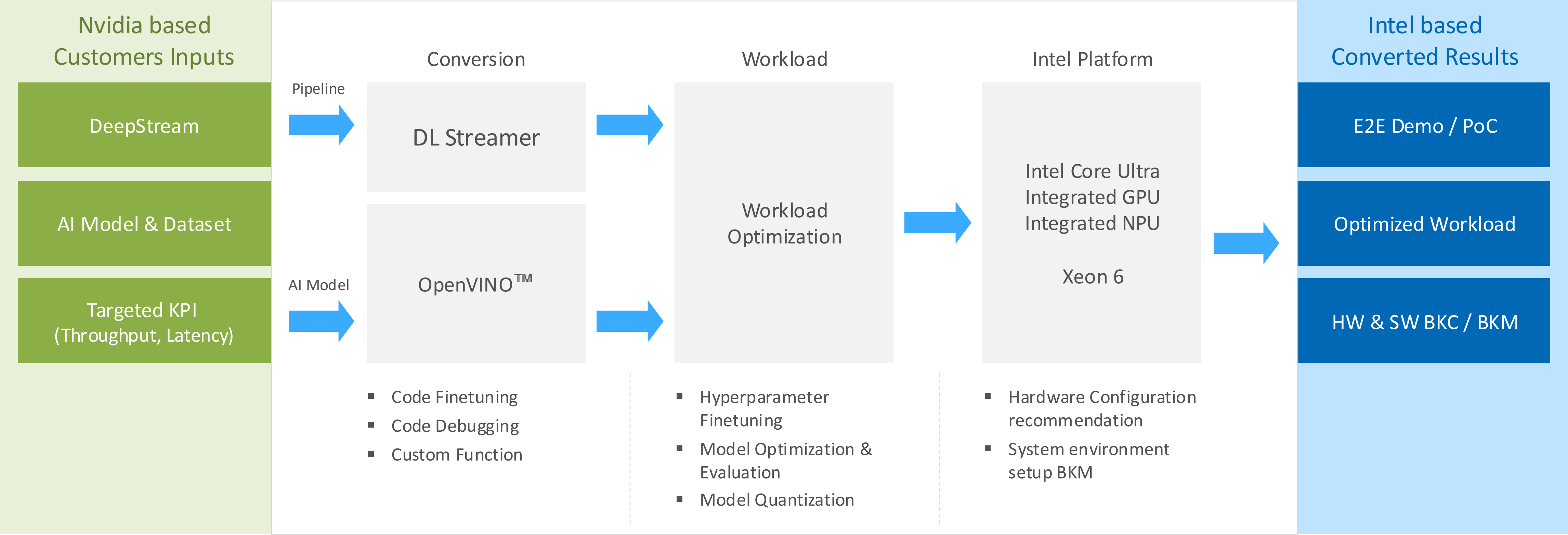
Get Intel Support

Intel Edge AI Workload Optimization Services

AI Application	Optimization Services	Customer Provides	Intel Service	Outcome
Object Detection Segmentation People/Face Detection Image Classification Gen AI/LLM Big Data Analytics Optical Character Recognition Secured AI Speech to Text Recognition Sound Classification	Convert to Intel	▪ CUDA source codes ▪ AI model ▪ KPI (Throughput, Latency)	▪ CUDA > SYCL migration ▪ Throughput & Latency Optimization ▪ Platform recommendation ▪ Evaluation on Intel Hardware	▪ Demo/POC ▪ Optimized workload ▪ HW/SW BKC & BKM
	LLM/ ML on Intel	▪ LLM Model Requirement ▪ Current/Target Platform ▪ KPI (Throughput, Latency)	▪ LLM Model Recommendation ▪ Model Quantization ▪ Workload Optimization ▪ Platform recommendation ▪ Sample application	▪ Demo/POC ▪ Optimized workload ▪ HW/SW BKC & BKM
	Securing AI Model & Dataset	▪ KPI (Throughput, Latency) ▪ Use Case ▪ AI Model & Sample Dataset ▪ Security Requirement	▪ End-to-end encrypted training & inferencing ▪ Protecting the NN model in Secure Enclave ▪ Model copyright protection through Watermark/fingerprint technology	▪ Demo/POC ▪ Encrypted Model & Dataset ▪ HW/SW BKC & BKM
	Model Training & Inference Optimization	▪ KPI (Throughput, Latency) ▪ Use Case ▪ Sample Dataset ▪ Trained Model	▪ Model Optimization & Quantization ▪ Hyper Parameter Tuning ▪ Performance Optimization & Demonstration ▪ Platform recommendation ▪ Reference Configuration	▪ Demo/POC ▪ Optimized workload ▪ HW/SW BKC & BKM

Conversion Journey and Intel Support*

*Details varies based on project needs



Nvidia based Customers Inputs

- DeepStream
- AI Model & Dataset
- Targeted KPI (Throughput, Latency)

Pipeline

AI Model

Conversion

- DL Streamer
- OpenVINO™

- Code Finetuning
- Code Debugging
- Custom Function

Workload

Workload Optimization

- Hyperparameter Finetuning
- Model Optimization & Evaluation
- Model Quantization

Intel Platform

Intel Core Ultra
Integrated GPU
Integrated NPU

Xeon 6

- Hardware Configuration recommendation
- System environment setup BKM

Intel based Converted Results

- E2E Demo / PoC
- Optimized Workload
- HW & SW BKC / BKM

Public Community Support

"How to install recent released Intel software?"

Standard Ticket Support

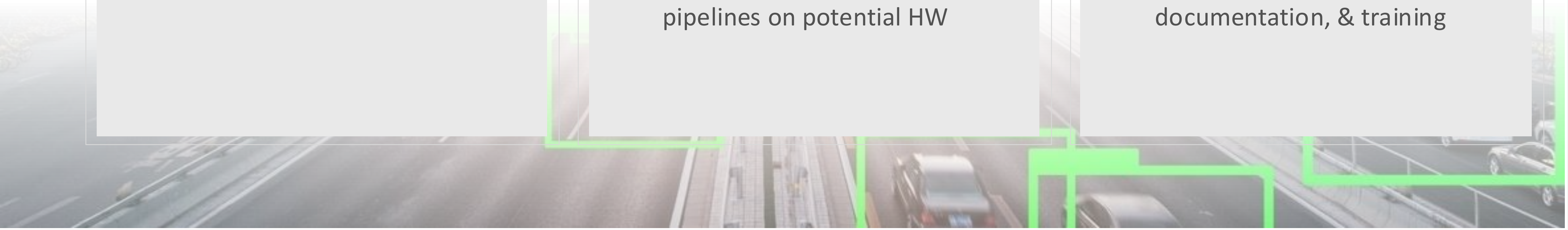
"Intel SW returns error while my API calls Intel SDK"

Premium Enabling Services

"Help me build traffic monitor AI solution on Intel SW & HW"

Next Steps

Metro AI Suite: Your gateway to optimized Visual AI & Gen AI solution development



Rapidly develop GenAI & Visual AI

- **Jumpstart AI feature development** by reviewing Proxy Pipelines, Sample Apps, and Blueprints

Size Platforms to Fit Diverse AI Needs

- **Search HW Catalog** for AI systems that meet your requirements
- **Use ViPPET** to quickly evaluate proxy pipelines on potential HW

Optimize AI for Cost Efficiency

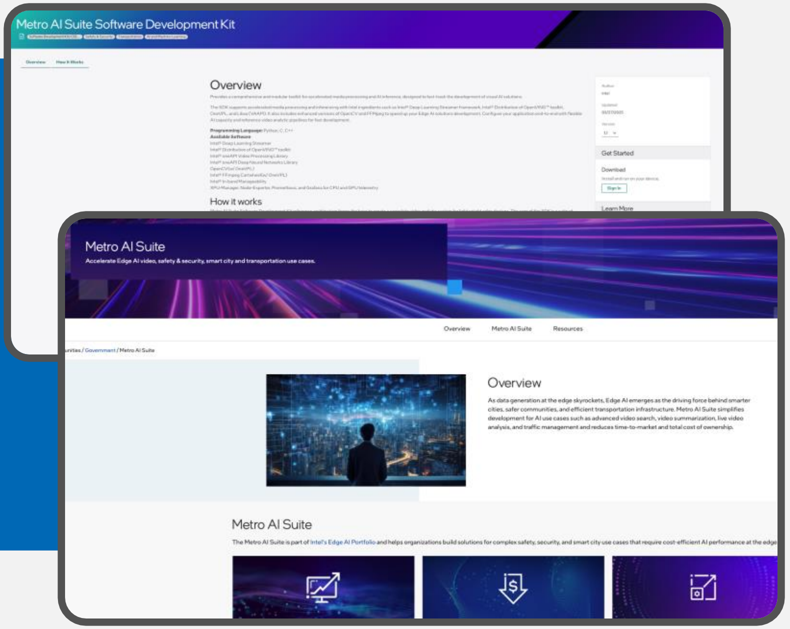
- **Gain performance improvements** via SDK libraries
- **Refer to resources,** such as optimized code samples, documentation, & training

Take the Next Step

- Software Collection: starting point for Video Analytics & AI solution development
- Enables Software Developers to quickly add Visual & GenAI to existing solutions
- Helps maximize performance of market leading edge AI silicon from Intel

Metro AI Suite

- intel.com/metroai
- Github page



Metro AI Suite Software Development Kit

Overview

How it Works

Overview

Provides a comprehensive and modular toolkit for accelerated media processing and AI inference, designed to fast-track the development of visual AI solutions.

The SDK supports accelerated media processing and inferencing with Intel ingredients such as Intel® Deep Learning Streamer framework, Intel® Distribution of OpenVINO™ toolkit, OneVPL, and Libva (VAAPI). It also includes enhanced versions of OpenCV and FFMpeg to speed up your Edge AI solutions development. Configure your application end-to-end with flexible AI capacity and reference video analytic pipelines for fast development.

Programming Language: Python, C, C++

Available Software

Intel® Deep Learning Streamer
Intel® Distribution of OpenVINO™ toolkit
Intel® oneAPI Video Processing Library
Intel® oneAPI Deep Neural Networks Library
OpenCV(w/ OneVPL)
Intel® FFmpeg Cartwheel(w/ OneVPL)
Intel® In-band Manageability
XPU-Manager, Node-Exporter, Prometheus, and Grafana for CPU and GPU telemetry

How it works

Get Started

Download

Install and run on your device.

Sign In

Learn More

Metro AI Suite

Accelerate Edge AI video, safety & security, smart city and transportation use cases.

Overview | Metro AI Suite | Resources

...unities / Government / Metro AI Suite

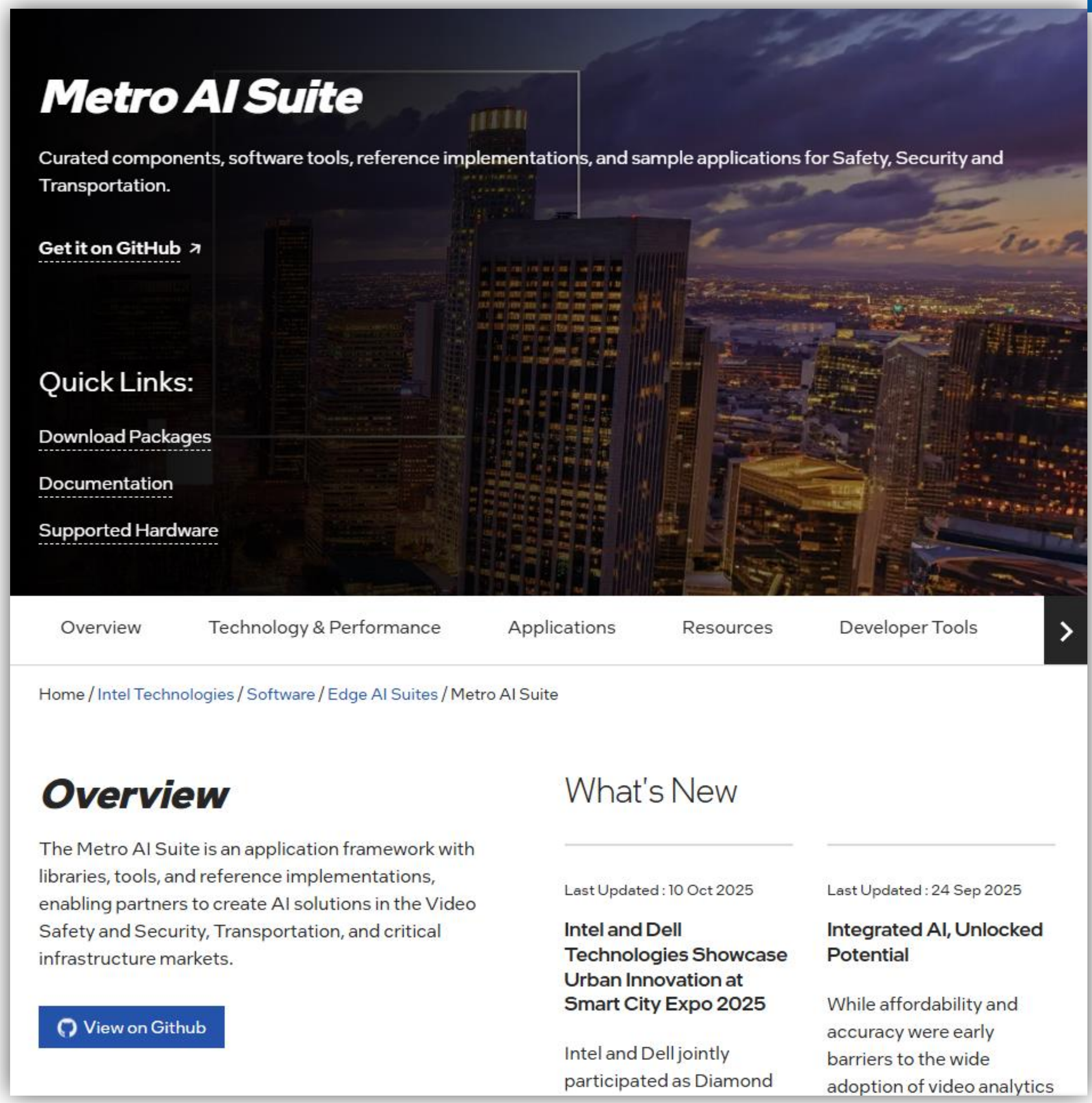
Overview

As data generation at the edge skyrockets, Edge AI emerges as the driving force behind smarter cities, safer communities, and efficient transportation infrastructure. Metro AI Suite simplifies development for AI use cases such as advanced video search, video summarization, live video analysis, and traffic management and reduces time-to-market and total cost of ownership.

Metro AI Suite

The Metro AI Suite is part of Intel's Edge AI Portfolio and helps organizations build solutions for complex safety, security, and smart city use cases that require cost-efficient AI performance at the edge

- Use Metro SDK for optimized software development with OpenVINO™, DL Streamer, and more
- Reference Sample Apps, pipelines, blueprints, and documentation for fast & easy solution development
- Speed up system evaluation by reviewing prequalified AI systems in the Hardware Catalog
- Leverage support programs for technical enablement, including architecture transition



Metro AI Suite

Curated components, software tools, reference implementations, and sample applications for Safety, Security and Transportation.

Quick Links:

Download Packages

Documentation

Supported Hardware

Home / Intel Technologies / Software / Edge AI Suites / Metro AI Suite

Overview

The Metro AI Suite is an application framework with libraries, tools, and reference implementations, enabling partners to create AI solutions in the Video Safety and Security, Transportation, and critical infrastructure markets.

What's New

Last Updated : 10 Oct 2025

Intel and Dell Technologies Showcase Urban Innovation at Smart City Expo 2025

Intel and Dell jointly participated as Diamond

Last Updated : 24 Sep 2025

Integrated AI, Unlocked Potential

While affordability and accuracy were early barriers to the wide adoption of video analytics

Resources

Getting started with Metro AI Suite

Featuring…

- Technologies & Tools
- Tutorials
- Recipes and Blueprints
- Code Samples
- Developer Guides
- Training Videos

https://builders.intel.com/intel-technologies/software/edge-ai-suites/metro-ai-suite

Appendix

Claim # & Statement	Slide & Title/Details
Slide : Intel Visual AI Competitive Advantage	
1. Intel® Core™ Ultra 7 Processor 155H is 55% lower in System Cost vs NVIDIA Jetson Orin AGX 64GB	As on February 06, 2025: Intel Cost as per ark.intel.com, NVIDIA Cost as per arrow.com
2. Intel® Core™ Ultra 7 Processor 265H delivers 3.5x higher Media decode performance than NVIDIA Jetson Orin AGX 64GB	Results are based on Intel internal measurements/estimation/calculations as of January 2025 . Intel® Core™ Ultra 7 Processor 265H has up to 130 streams, Intel® Core™ Ultra 7 Processor 155H has up to 126 streams vs NVIDIA Jetson Orin AGX 64GB has 37 streams.
3. Intel® Core™ Ultra 7 Processor 265H delivers 2.3x higher E2E Pipeline performance than NVIDIA Jetson Orin AGX 64GB	Results are based on Intel internal measurements/estimation/calculations as of January 2025 . Intel® Core™ Ultra 7 Processor 265H has up to 71 streams, Intel® Core™ Ultra 7 Processor 155H has up to 55 streams vs NVIDIA Jetson Orin AGX 64GB has 32 streams.
4. Intel® Core™ Ultra 7 Processor 265H and Intel® Core™ Ultra 7 Processor 155H delivers 3.9x & 5.0x higher Performance/$ respectively vs NVIDIA Jetson Orin AGX 64GB	Results are based on Intel internal measurements/estimation/calculations as of January 2025 : Ratio of End to end number of streams performance to system cost of Intel® Core™ Ultra 7 Processor 265H and Intel® Core™ Ultra 7 Processor 155H normalized to NVIDIA Jetson Orin AGX 64GB
5. With Intel® Core™ Ultra 7 Processor you can achieve upto Double E2E pipeline Performance at almost half the system Cost price vs NVIDIA Jetson Orin AGX 64GB	As calculated by Intel: end to end pipelines and cost comparison
Configurations for claims 1-5	Media Pipeline Workload : HEVC 1080p30 decode E2E Pipeline Workload : 1080p30 HEVC decode + pre-processing + detection using Yolov5s_640 @ 5fps 1 object per frame + classification using Mobilenet-V2 @ 5 inf/s/str + Resnet50 @ 5 inf/s/str Example iGPU + NPU pipeline: taskset -c 12-19 gst-launch-1.0 filesrc location="./Videos/svetclip_1080p_30p_2M_loop100.h265" ! h265parse ! vaapih265dec ! capsfilter caps="video/x-raw(memory:VASurface)" ! queue ! gvadetect model="./Networks/yolo-v5s/INT8/yolov5s-v6-1.xml" model-proc="./Networks/yolo-v5s/INT8/yolo-v5.json" device=GPU nireq=2 pre-process-backend=vaapi-surface-sharing batch-size=8 ie-config=NUM_STREAMS=2 inference-interval=6 threshold=0.5 model-instance-id=yolov5 ! queue ! vaapipostproc crop-right=1696 crop-bottom=856 ! queue ! gvaclassify model="./Networks/rn50-optimized/resnet-50.xml" device=NPU nireq=2 model-proc=./intel/dlstreamer_gst/samples/gstreamer/model_proc/intel/resnet50-binary-0001.json pre-process-backend=opencv batch-size=1 inference-interval=6 inference-region=0 model-instance-id=resnet50 ! queue ! gvaclassify model="/home/arl/Networks/mv2_optimized/mobilenet-v2.xml" device=NPU nireq=2 model-proc=./intel/dlstreamer_gst/samples/gstreamer/model_proc/onnx/mobilenetv2-7.json pre-process-backend=opencv batch-size=1 inference-interval=6 inference-region=0 model-instance-id=mobilenetv2 ! gvafpscounter starting-frame=4000 ! fakesink sync=false async=false Performance results are based on testing as of Nov 08, 2024 : Processor: Intel® Core Ultra 7 processor 265H; tested on an Intel Internal development system; Memory 16GB (2x8GB DDR5 6400 MT/s [6400 MT/s]); Storage: 1x 465.8G WD_BLACK SN770 500GB, 1x 465.8G Samsung SSD 980 500GB; OS: Ubuntu 2024.04.1 LTS; Intel Arc Graphics 128 EUs @ 2.30GHz OpenCL Compute Driver 24.17.29377.6; Intel AI Boost @1.4GHz NPU, NPU Driver v1.8.0; BIOS: MTLPEMI1.R00.4341.D43.2409200738; Scaling Governor set to Performance. June 03, 2024 : Processor: Intel® Core Ultra 7 processor 155H; tested on publicly obtained system; Memory 32GB (2x16GB DDR5 5600 MT/s [4800 MT/s]); Storage: 1x 953.9G KINGSTON OM8SEP41024Q-A0; OS: Ubuntu 22.04.4 LTS; Intel Arc Graphics 128 EUs @ 2.25GHz OpenCL Compute Driver 24.17.29377.6; Intel AI Boost @1.4GHz NPU, NPU Driver v1.5.0; BIOS: 5.3; Scaling Governor set to Performance. June 03, 2024 : NVIDIA Jetson Orin AGX 64GB tested on publicly obtained system; CPU : 12-core Arm Cortex-A78AE v8.2 64-bit CPU 3MB L2 + 6MB L3 @ 2.2Ghz, GPU: NVIDIA Ampere architecture with 2048 NVIDIA® CUDA® cores and 64 Tensor cores @ 1.3Ghz, Memory: 64GB 256-bit LPDDR5 @ 204.8 GB/s, Storage: 64GB eMMC 5.1, Accelerator: 2x NVDLA v2.0, PVA v2.0 Accelerator APE , NVENC , NVDEC , NVJPG , NVJPG1, VIC , SE , OS: Ubuntu 22.04.4 LTS, Kernel: 5.15.136-tegra, Jetpack: 6.0 (Rev 2), Deepstream: 7.0, CUDA: 12.2, TensorRT: 8.6.2.3-1+cuda12.2, Jetson clocks: Enable, NVP modes: 0, /etc/sysctl.conf: vm.overcommit_memory=0 Learn more at intel.com/performanceindex.

Appendix

Claim # & Statement	Slide # & Title/Details
Slide : Intel Visual AI Competitive Advantage	
1. Intel® Arc™ Graphics can deliver upto 14.0x higher Performance/$ than NVIDIA L4 Tensor Core GPU	Results are based on Intel internal measurements/estimation/calculations as of January 2025 : Ratio of End to end number of streams performance to system cost of Intel® Arc™ GPUs vs NVIDIA L4 Tensor Core GPU. Cost as on February 06, 2025: Intel Arc Graphics - ark.intel.com, Intel Arc A580 Graphics Available Worldwide, NVIDIA: Dell System Build/Shop: Dell PowerEdge R760 Rack Server
Configurations for claim	Media Pipeline Workload : HEVC 1080p30 decode E2E Pipeline Workload : 1080p30 HEVC decode + pre-processing + detection using Yolov5M_640 @ 10fps 1 object per frame + classification using Mobilenet-V2 @ 10 inf/s/str + Resnet50 @ 10 inf/s/str Example dGPU pipeline: taskset -c 12-19 gst-launch-1.0 filesrc location="./Videos/svetclip_1080p_30p_2M_loop100.h265" ! h265parse ! vaapih265dec ! capsfilter caps="video/x-raw(memory:VASurface)" ! queue ! gvadetect model="./Networks/yolo-v5m/INT8/yolov5m-v6-1.xml" model-proc="./Networks/yolo-v5s/INT8/yolo-v5.json" device=GPU nireq=2 pre-process-backend=vaapi-surface-sharing batch-size=8 ie-config=NUM_STREAMS=2 inference-interval=6 threshold=0.5 model-instance-id=yolov5 ! queue ! vaapipostproc crop-right=1696 crop-bottom=856 ! queue ! gvaclassify model="./Networks/rn50-optimized/resnet-50.xml" device=GPU ie-config=NUM_STREAMS=2 nireq=2 model-proc=./intel/dlstreamer_gst/samples/gstreamer/model_proc/intel/resnet50-binary-0001.json pre-process-backend=vaapi-surface-sharing batch-size=8 inference-interval=6 inference-region=0 model-instance-id=resnet50 ! queue ! gvaclassify model="/home/arl/Networks/mv2_optimized/mobilenet-v2.xml" device=GPU ie-config=NUM_STREAMS=2 nireq=2 model-proc=./intel/dlstreamer_gst/samples/gstreamer/model_proc/onnx/mobilenetv2-7.json pre-process-backend=vaapi-surface-sharing batch-size=8 inference-interval=6 inference-region=0 model-instance-id=mobilenetv2 ! gvafpscounter starting-frame=4000 ! fakesink sync=false async=false Performance measured on Intel® Arc™ A750 GPU can also be used as proxy for Intel ®Arc ™A750E GPU System & Software Configuration: Host setup: Intel(R) Xeon(R) Platinum 8468V, 2.4GHz, Turbo & Hyperthreading Enabled, 512GB RAM, 465.8 GB WDC WDS500G2B0B-00YS70. BIOS 05.08.01, OS Ubuntu 22.04.4 LTS NVIDIA L4, 7424 cores, 24GB GDDR6, 72W TBP. NVIDIA driver 535.216.01. Container for Deepstream 6.4 (source), Container for TensorRT 24.01 (source) 1 Decode Throughput Pipeline Configuration: NVIDIA: The decode density pipeline script runs at near maximum or over maximum capability in stream count. Maximum capability is how many streams we can run at 30FPS. The input was a H265 1080p 2Mbps single object video, the same one that is used in the E2E Pipeline. The calculated density is what is used to show the stream count on NVIDIA products. Intel: The decode density pipeline script runs at near maximum or over maximum capability in stream count. Maximum capability is how many streams we can run at 30FPS. The input was a H265 1080p intersection video, the same one that is used in the E2E Pipeline. 2 AI Benchmark Target Networks: NVIDIA: YoloV5-M (source): INT8 Model, YoloV5-S (source): INT8 Model Intel: NN Model: Yolo-V5S, Yolo-V5M, (Quantized models follow standard guide) 3 Video Analytics Pipelines Both host systems and all the accelerators were run with default settings, no frequency setting or core pinning took place NVDIA: The chosen workload uses a configuration script and calculates the density based on several different runs. The details of this configuration are shown in the next two pages. Example workload: python3 deepstream_bench.py pipeline object-detection-2-classification --detection-model yolov5m-640 --classification-model resnet-50-tf --classification-model-2 mobilenet-v2-pytorch --gpu-name L4 --mode calculate-density --measurements-dir measurements --batch-size 8 --detection-interval 2 --video Uniform_video_FHD_HEVC_loop9.mp4 *As of January 2025; Intel® Arc™ B580 graphics Performance is projected based on Intel analysis of current product definition and available information. Projection range of accuracy is +/- 10% Learn more at intel.com/performanceindex.