Metro AI Suite Software Developer Guide

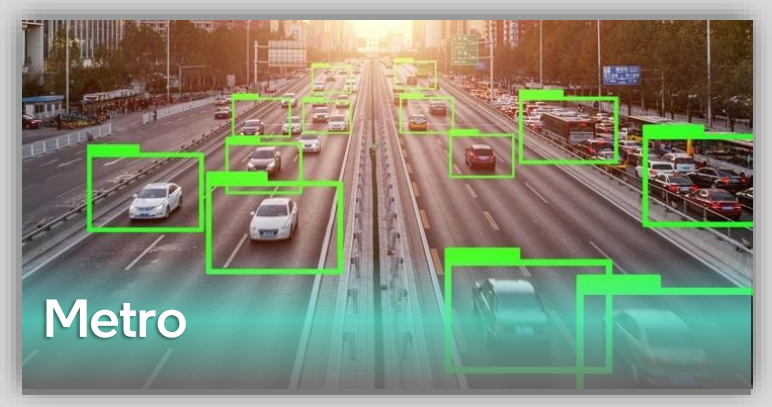
Ver 3.0 – June 2026 Update

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

Contents

Edge AI Suites — Accelerate AI value at the edge

Open, modular suites combine reference apps, pipelines, tools, models and benchmarks to reduce risk and speed deployment on Intel®.



Evaluate reference implementations | Benchmark AI workloads | Demonstrate performance and TCO

Accelerate Edge AI for Critical Use Cases

Rapidly develop Visual AI, Gen AI & Physical AI

- Reference software to simplify & accelerate development
- Deploy Gen AI and Physical AI on any Intel® edge platform — no discrete GPU required

Size Platforms to Fit Diverse AI Needs

- Benchmarking to evaluate hardware before you commit
- Right-size your AI workload across performance, power, and form factor

Optimize AI for Cost Efficiency

- Modular, open-source libraries, tools, and microservices — all free
- Maximize AI and video performance across Intel's full system range

The Philosophy Behind Metro AI Suite

Simplify adoption of Visual AI, Gen AI and Physical AI

Make it easy to bring Visual AI, Gen AI, and Physical AI to any platform

Run AI workloads across Intel's full portfolio — from entry-level to server

Streamline the path from prototype to production-grade AI

Build on open-source foundations with Apache-licensed components

Flex across deployment models — bare metal, containers, services, and Kubernetes

Metro AI Suite:

A Partner-Driven Framework to Accelerate, Optimize, and Scale Metro AI

Metro AI Suite

Vertical Sample AI Applications	Fast-track development of high-value metro AI solutions
Hardware Sizing & Evaluation	Choose the right hardware and measure AI performance with confidence
Metro SDKs	Full optimization toolkit — libraries, tools, and APIs for peak AI performance

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

Metro AI Suite vertical framework for Edge AI: Vision, Physical, & GenAI

Create & Acquire Models	Choose Platform	Build	Add capabilities
Build a model	Evaluate Hardware	Jumpstart your solution with samples	Make your solution Agentic
Pick a model	Search AI-ready qualified systems	Build production-grade AI pipelines, with SDKs	Integrate AI to a VMS
Bring-Your-Own Model	Access pre-benchmarked results		Bring your AI into the physical world

Optimize - maximum performance for edge deployment

Metro AI Suite Developer stack for Edge Vision AI + GenAI + Physical AI

Create & Acquire Models

Build a model with visual prompting :
Geti™

Pick a model from Hugging Face:
Metro Analytics Catalog
&
OpenVINO Model Collections

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

Choose Platform

Evaluate Hardware:

Visual Pipeline and Platform Evaluation Tool (ViPPET)

Search AI-ready qualified systems:
Recommended Hardware Catalog

Access pre-benchmarked results:
Metro AI Suite Benchmarks
&
OpenVINO benchmarks

Build

Jumpstart your solution:
Sample AI Applications & Blueprints


Live Video Caption & Search


Video Search and Summarization


And many more...

Build production-grade AI pipelines, modularly:
Vision + AI SDKs

SDKs that include: microservices, libraries, tools, 3rd party components for video & AI

Deep Learning Streamer

Add capabilities

Make your solution autonomous:
Agentic AI


Live Video Alert Agent


Smart Route Planning Agent


Agentic Predictive Maintenance

Integrate AI to your VMS

VMS integrations

Smart NVR

Bring your AI into the physical world:

Scenescape

Smart Buildings Digital Twin

Optimize

Maximize performance for edge deployment

OpenVINO™

DL Streamer Pipeline Optimization Tool

ViPPET

Performance Blueprints

NEW RELEASE

Sample Applications, Plug-ins and Tools

Video Surveillance as a Service
NEW!

Agentic Augmented RAG
NEW!

AI Cybersecurity Enablement Tool
NEW!

VMS Integrations
NEW!

Metro Analytics Catalog (preview)
NEW!

Critical Infrastructure Multimodal Agentic Predictive Maintenance Blueprint

Deep Learning Streamer

- Vibe coding with AI Coding Agent (preview)
- Seamlessly migrate DeepStream to DL Streamer
- Standardized metadata

Visual Pipeline and Platform Evaluation Tool (ViPPET)

Powerful new VLM pipeline, flexible customization & NPU Support

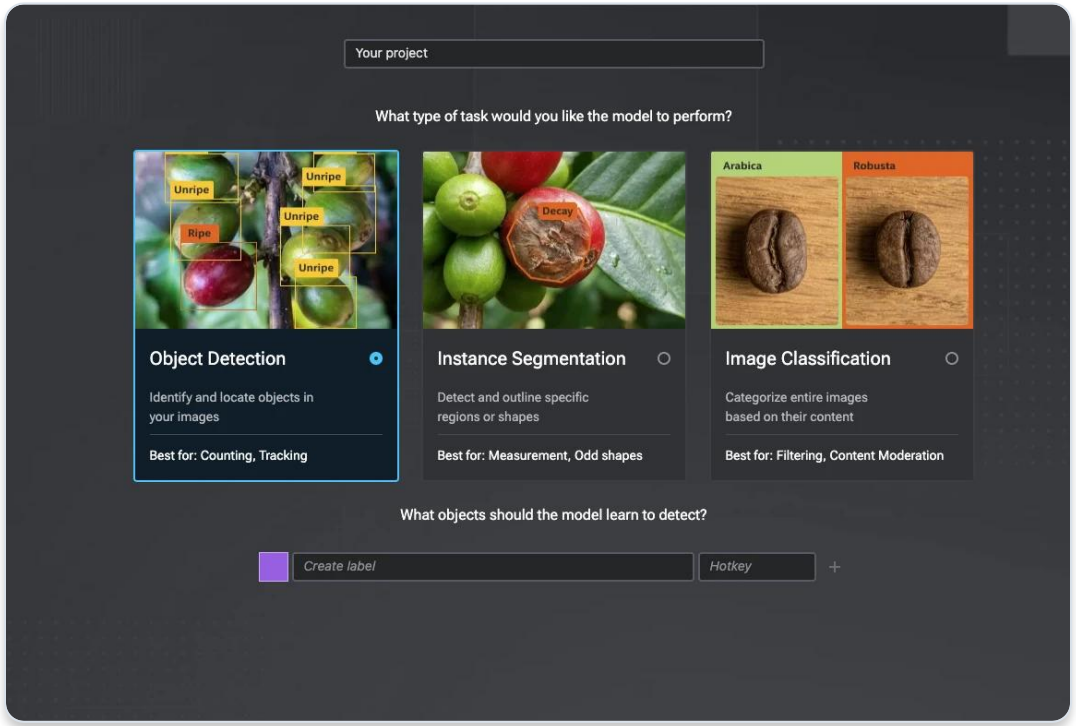
Spatial AI / Physical AI SceneScape

Sharper Physical AI that sees clearly through occluded, dynamic environments

Geti™ : Powerful Vision AI for Everyone

From data labeling to optimization and pipeline deployment, Geti™ creates production-ready vision AI models more efficiently

Label Data → Train & Optimize → Configure Pipeline → Deploy at Scale



Develop vision models with Geti™ :

- **Smart Annotations** – Expedite and simplify data labeling
- **Iterative Training** – Annotate, predict, and retrain to 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** – Optimize and quantize trained models with software that auto-detects available compute across Intel CPUs, GPUs, and NPUs, then export to OpenVINO IR

New

Metro Analytics Catalog

Ready-to-deploy AI models and analytics workflows optimized for Intel hardware and built for Metro use cases.

- **Curated, tested, converted, quantized models** for Intel hardware
- **DL Streamer pipelines** and sample code for rapid deployment
- **End-to-end documentation** for setup and usage

Loitering Detection
Flags when a person lingers in an area beyond a set time.

License Plate Recognition
Reads and logs vehicle plate numbers from video.

Crowd Detection
Estimates crowd density and flags overcrowding.

Person Detection
Locates people in each frame.

Motion Tracking
Follows people or objects as they move across the scene.

Worker Safety Detection
Detects safety gear like vests and hard hats to verify PPE compliance.

Vehicle Detection
Identifies and counts vehicles.

Object Detection
Locates and classifies general objects.

Now available on 🤗 **Hugging Face**

Platform Sizing

ViPPET (Visual Pipeline and Platform Evaluation Tool)

Choose the right platform with real pipeline evidence

1
Choose base pipeline

Select a video workflow

2
Customize pipeline

Choose models and other parameters to match your workflow

3
Optimize & Benchmark

ASSIGN WORKLOADS TO

CPU | iGPU | NPU

Measure throughput, latency, utilization & TCO

Flexible stage-to-hardware mapping reflects real deployment choices.

- See how many camera channels run on CPU, iGPU, and NPU.
- Reduce rollout risk with evidence-based hardware choices.
- Benchmark end-to-end media and AI workloads on real pipelines.
- Auto-optimize for the best throughput, power, and TCO.

Explore details →

Easy End-to-End Visual AI Benchmarking with ViPPET

Why ViPPET?

Create a new pipeline or select pre-defined pipelines

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.

Explore details →

Easy End-to-End Visual AI Benchmarking with ViPPET

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.

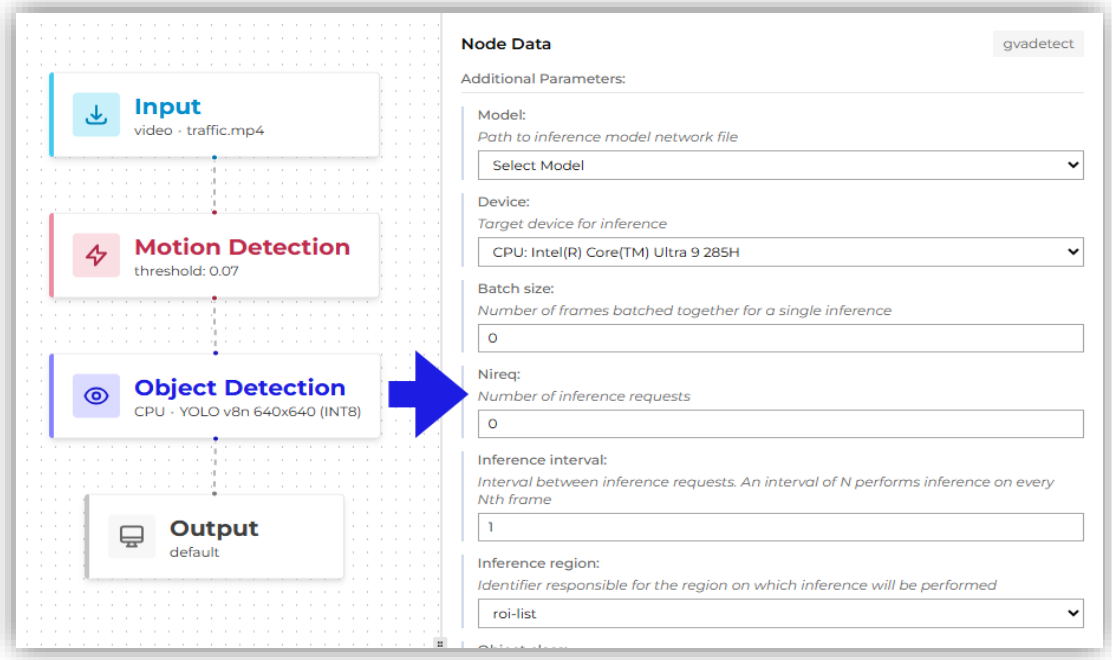
Advanced Usage

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.



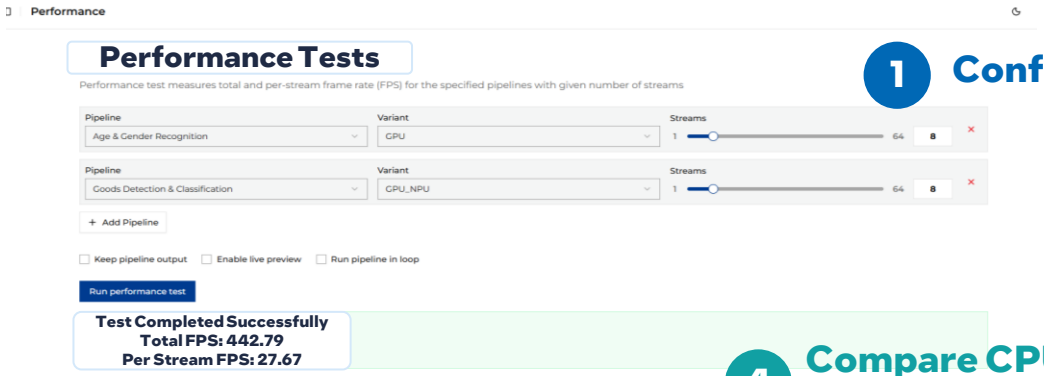
Pipeline Customization

Explore details →

Easy End-to-End Visual AI Benchmarking with ViPPET

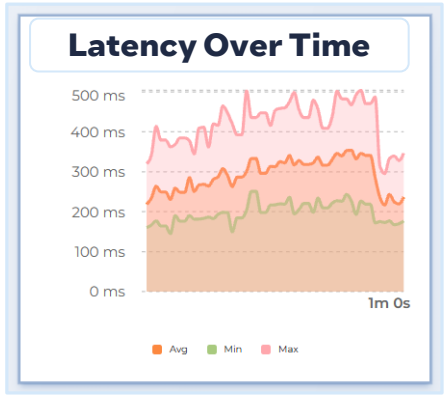
Benchmark end-to-end media and AI workloads on real pipelines.

1 Configure a real pipeline test

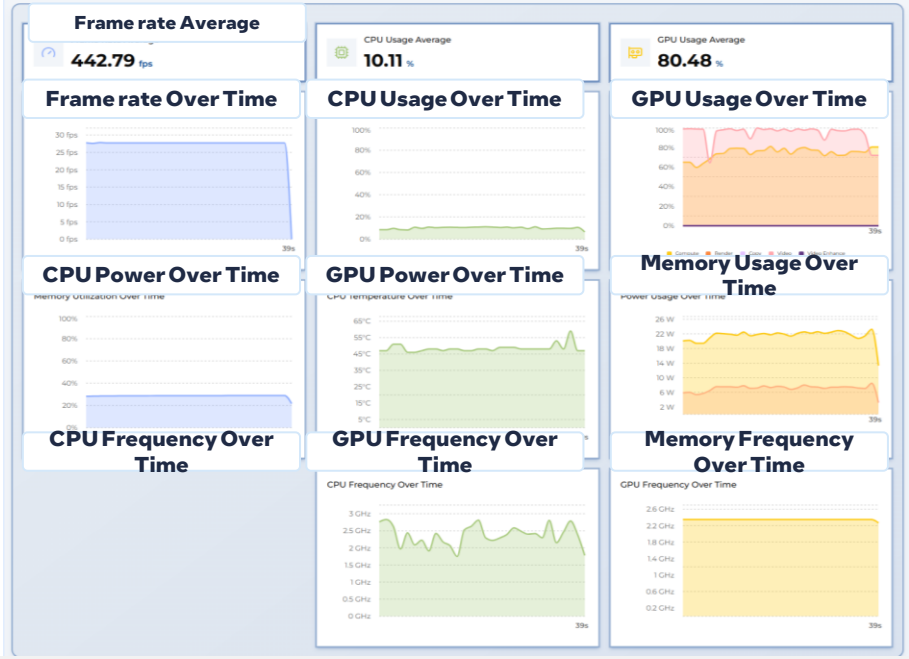


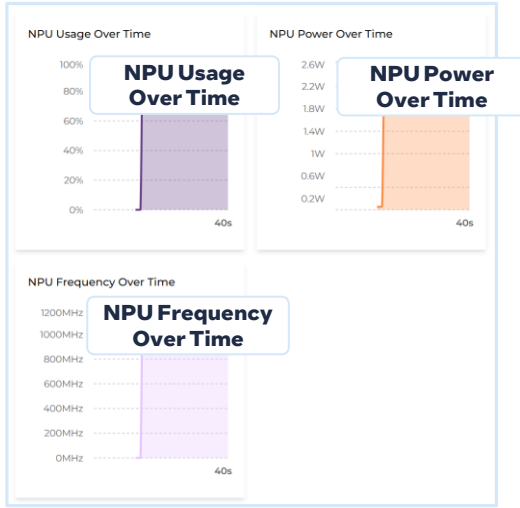
2 Confirm the workload has target FPS.

3 Monitor live latency and spot latency spikes



4 Compare CPU, iGPU and now NPU telemetry before deployment.



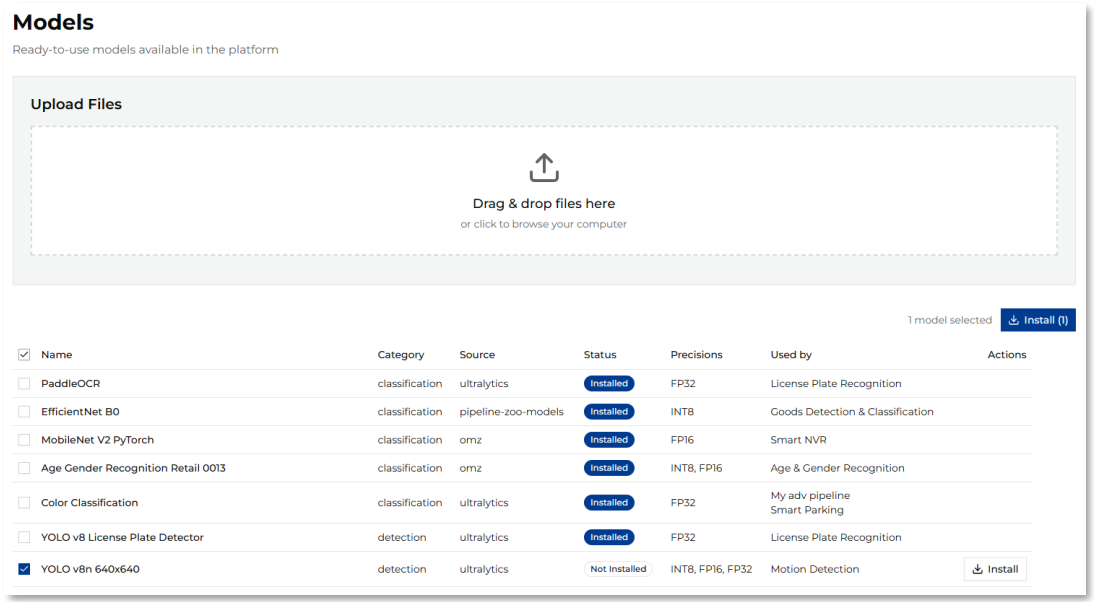


Explore details →

New

Easy End-to-End Visual AI Benchmarking with ViPPET

Flexible Pipelines, Deeper Performance Visibility



New Predefined Pipelines supporting VLMs for more vertical business use cases.

Custom OpenVINO and Intel Geti model uploads for easy bring-your-own-model testing and reuse.

Camera input, video and image-set uploads for flexible, real-world pipeline inputs.

Real-time pipeline output for improving experience.

NPU utilization reporting alongside CPU and GPU for full visibility into AI pipeline performance on modern Intel hardware.

End-to-end latency metrics alongside throughput for spotting bottlenecks and verifying real-time pipeline performance.

Explore details →

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

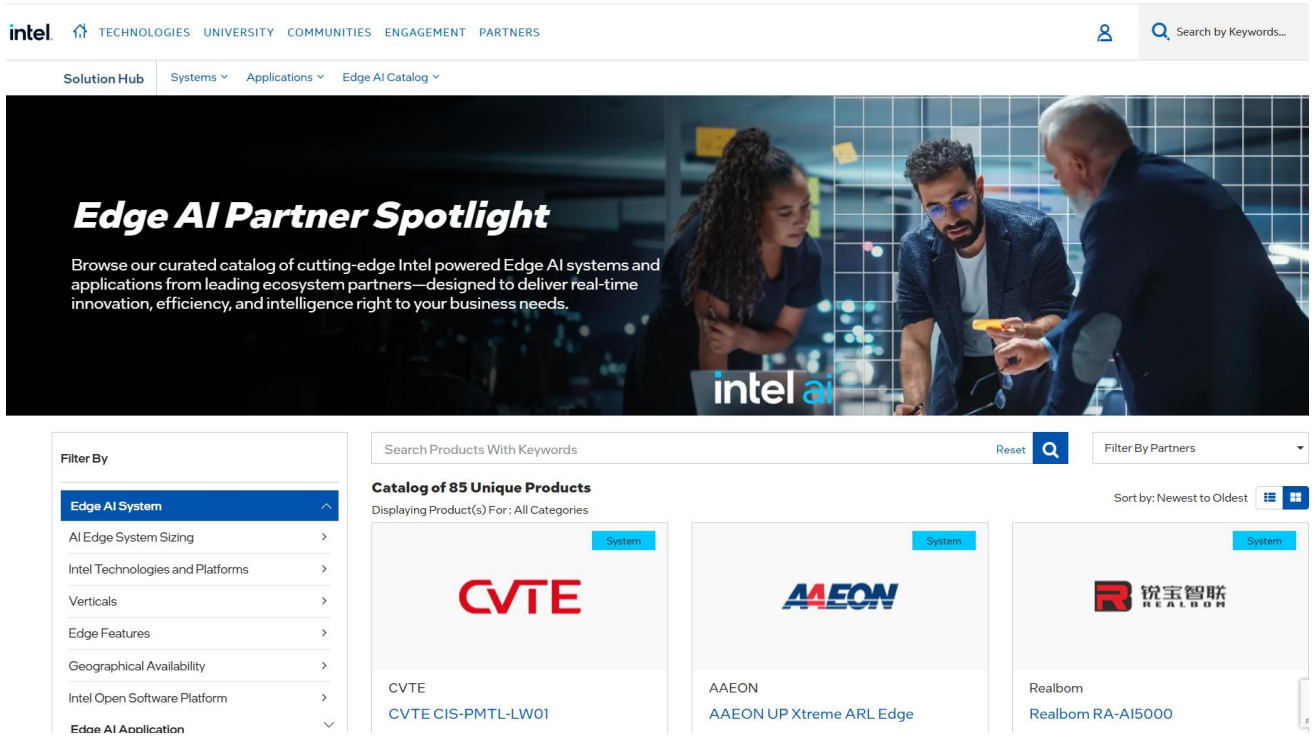
Leadership for Edge:
Support
Supply Life
Ecosystem

- Scalable Performance — intel Xeon
- Scalable AI, Graphics, and Media — intel ARC GRAPHICS
- Efficiency Optimized AI — intel CORE ULTRA
- Mainstream — intel CORE ULTRA; intel CORE; intel processor
- Entry — intel CORE; intel processor; intel. ATOM

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



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!

Reference Visual and Gen AI Sample Apps and Blueprints

Why Use AI Sample Apps?

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

Sample Applications, Tutorials, Blueprints

- ✓ Smart Parking
- ✓ Smart Intersection
- ✓ Loitering Detection
- ✓ Smart Tolling
- ✓ Reidentification (Image based Video Search)
- ✓ Interactive Digital Avatar
- ✓ Gen AI Based Video Q&A (VQ&A)
- ✓ Sensor Fusion for Traffic Management
- ✓ Video Processing for NVR
- ✓ Video Search and Summarization
- ✓ Live Video Applications
- ✓ NVIDIA → Intel Migration Guide
- ✓ Agentic AI Tutorials

Smart Parking

AI-driven parking management application that analyzes live video with pre-trained deep learning models to detect, count, and track open spaces and report parking availability in real time.

Real-time parking space detection and occupancy monitoring

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:

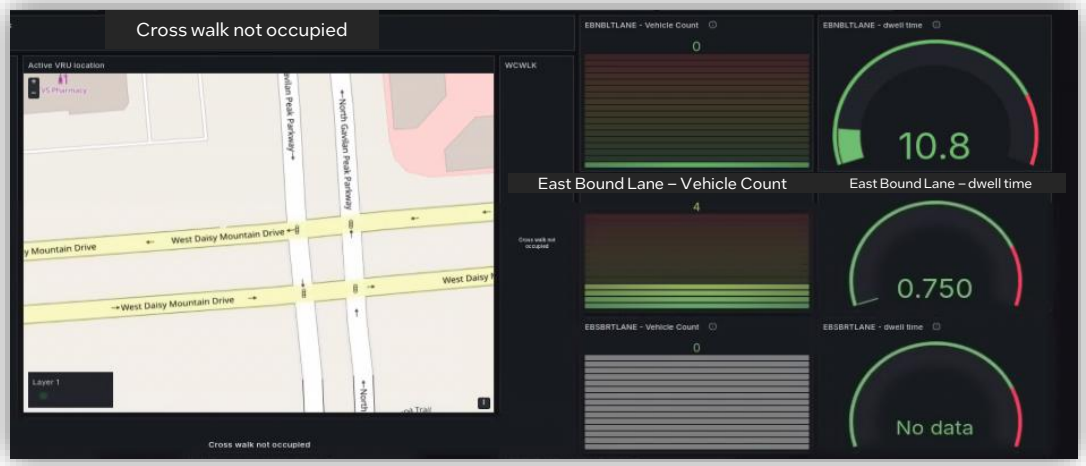
- Intel® Core™, Intel® Core™ Ultra, Intel® Xeon® platforms

Explore details →

Smart Intersection

Edge-AI traffic management application that fuses multi-camera, scene-based analytics — including lidar and radar — to track vehicles and pedestrians across an intersection and deliver real-time, coordinated traffic insights.

Real-time, multi-sensor traffic insights across a complex intersection



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

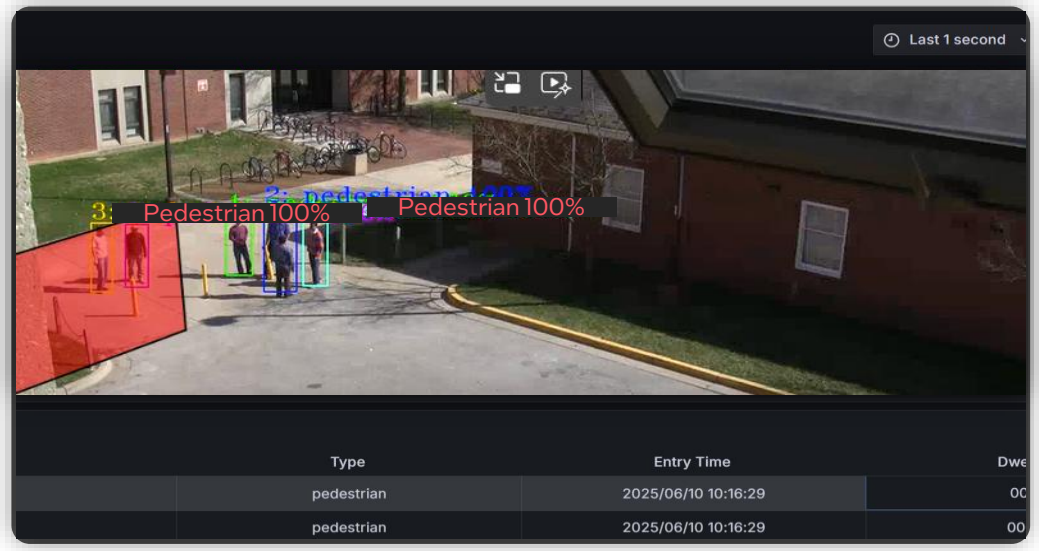
Supported Intel Platforms:

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

Explore details →

Loitering Detection

AI vision application that tracks people in live video and applies dwell-time thresholds to flag lingering individuals and raise real-time alerts.



Loitering detection in action

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

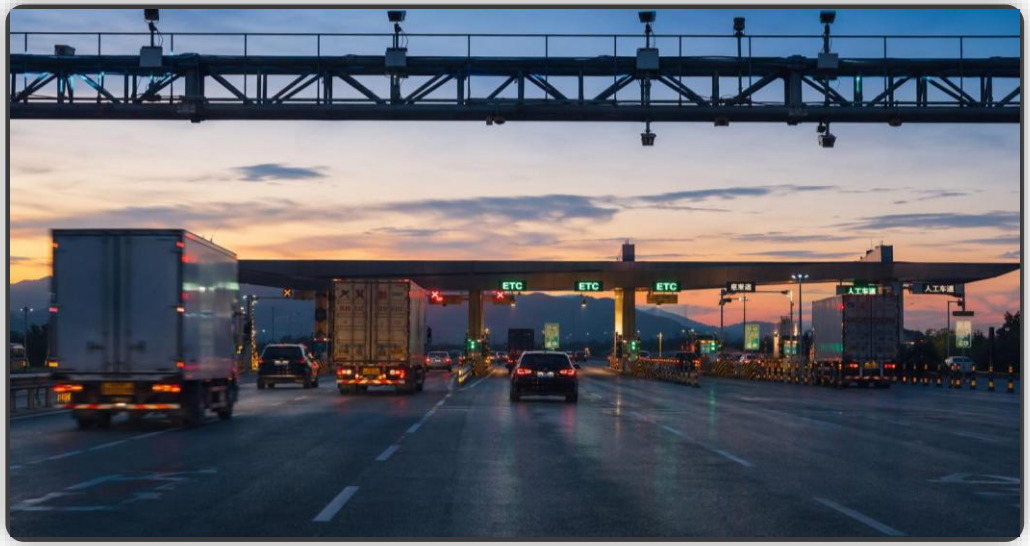
Supported Intel Platforms:

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

Explore details →

Smart Tolling

Edge AI application that automates vehicle identification, classification, and accurate toll calculation across multi-lane toll points.



Key Features:

- **Multi-Camera Vehicle Recognition:** Reads plates and identifies vehicle type and color in real time.
- **Sensor Fusion with Intel® SceneScape**: Fuses camera views to track each vehicle through the toll zone.
- **Accurate Axle Counting**: Detects lifted axles for fair taxation
- **Operator Dashboard:** Grafana dashboard for traffic, flow, and audits on InfluxDB time-series data.

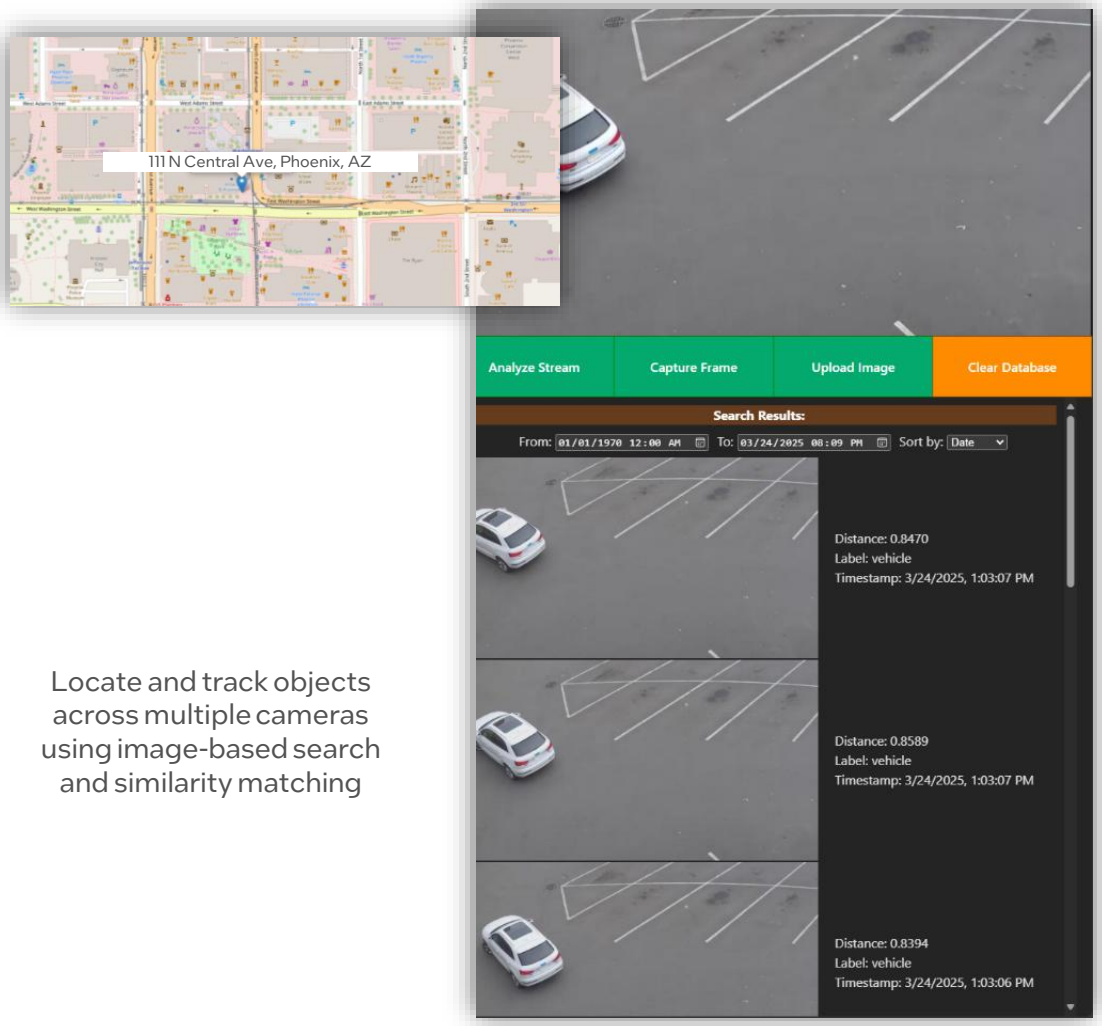
Supported Intel Platforms:

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

Explore details →

Reidentification (Image based Video Search)

Image-based video search application that combines object detection, feature extraction, and vector search to locate a query object across live or recorded camera feeds in near real time.



Locate and track objects across multiple cameras using image-based search and similarity matching

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**

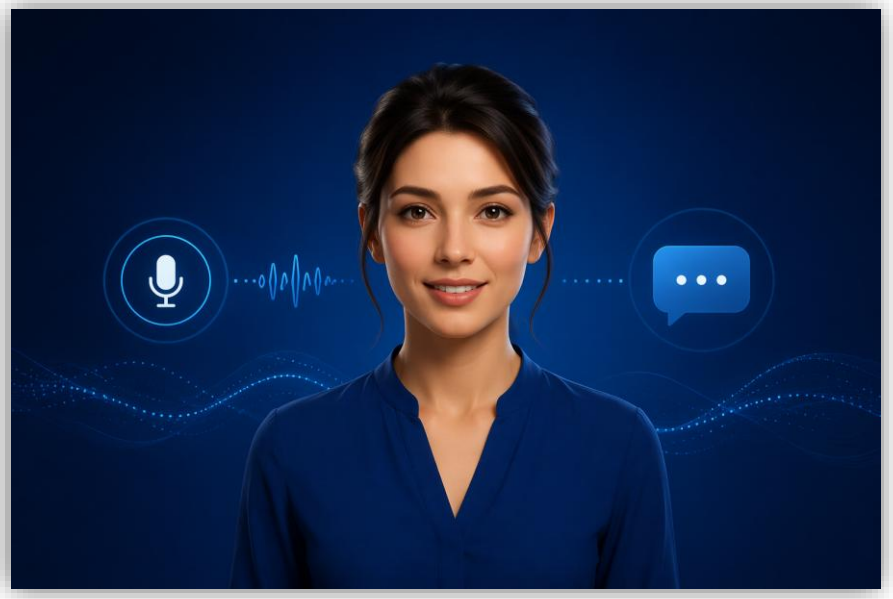
Supported Intel Platforms:

- Intel® Core™ and Intel® Core™ Ultra platforms

Explore details →

Interactive Digital Avatar

An AI-powered avatar that understands speech, analyzes video context, and responds in real time—enabling natural, interactive experiences for users.



Key Features:

- **Multimodal Interaction** Combines voice, vision, and language models to understand user input and respond naturally.
- **Real-Time Conversational AI** Delivers low-latency, context-aware responses using live video, captions, and RAG pipelines.

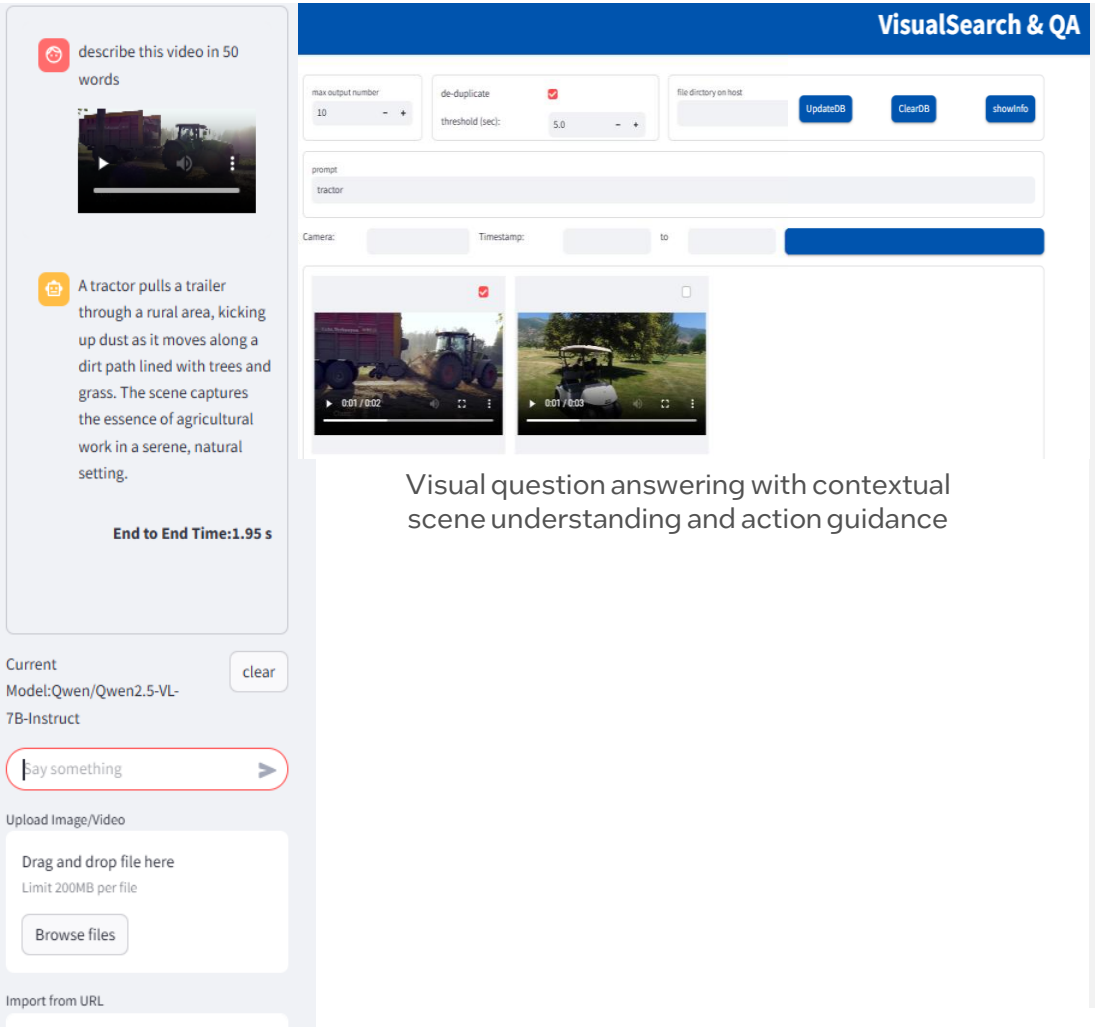
Supported Intel Platforms:

- Intel® Arc™ A770 graphics

Explore details →

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

Generative-AI assistant that uses a large vision language model to answer open-ended, natural-language questions about image and video content in real time.



Visual question answering with contextual scene understanding and action guidance

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**

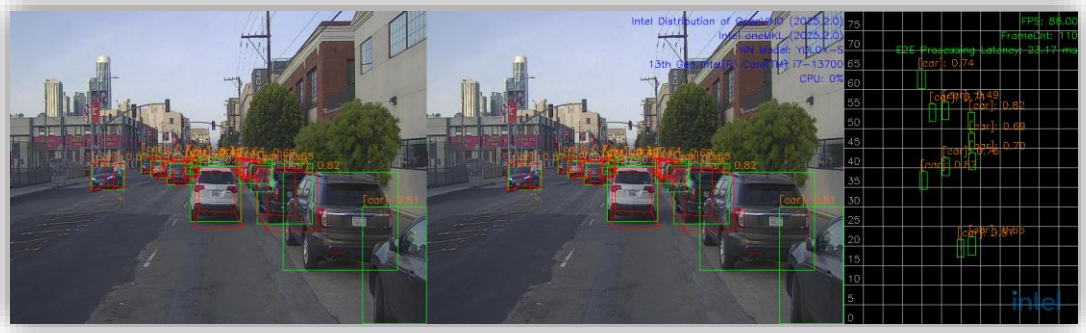
Supported Intel Platforms:

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

Explore details →

Sensor Fusion for Traffic Management

A multi-modal sample application to accurately monitor traffic conditions by fusing camera and sensor inputs



2 camera streams and 1 LiDAR sensor configuration

The sample application combines camera + LiDAR outputs to improve accuracy

Key Features:

- **Multi-modal sensor fusion:** Combine camera, radar, and lidar inputs for robust detection in fog, rain, or darkness
- **Accurate measurement:** Deliver reliable speed and distance estimation for traffic monitoring
- **End-to-end pipeline:** Lidar signal processing, video analytics, camera-lidar fusion, and visualization on a single Intel SoC
- **Scalable configurations:** Support multiple camera + lidar setups, from 1 camera +1 lidar sensor up to 12 cameras+4 Lidar censors

Supported Intel Platforms:

- Intel® Core™ Ultra processors with integrated GPU or Intel® Core™ processors with Intel® Arc™ B-Series Graphics

Explore details →

Video Processing for NVR

A sample application to evaluate and optimize NVR video processing workflows.

Key Features:

- **Concurrent multi-channel processing:** Decode, process, and display many H.264/H.265 streams at once on Intel integrated GPU
- **Video analytics & transcoding:** Run analytics and transcoding with ready-made example apps
- **Configuration-driven workloads:** Set channels, scaling, and layout via text config—no code changes
- **Multiview monitoring:** Check performance and debug pipelines across multiple 4K displays
- **Built for NVR evaluation:** Benchmark and debug core workloads in a real NVR setup

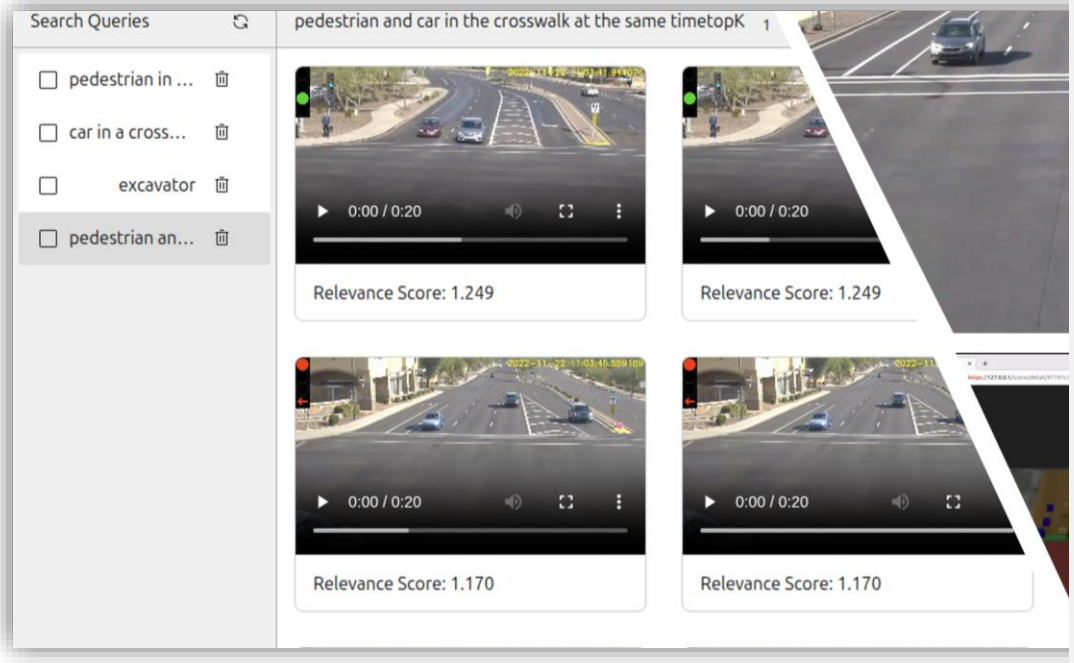
Supported Intel Platforms:

- Intel® Core™ processors with integrated GPU or Intel® Arc™ Graphics

Explore details →

Video Search and Summarization

Generative-AI video search and summarization application that uses multimodal embeddings, vector search, vision language models, and audio analysis to retrieve relevant clips from natural-language queries and condense long video streams into concise summaries.



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.
- **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.

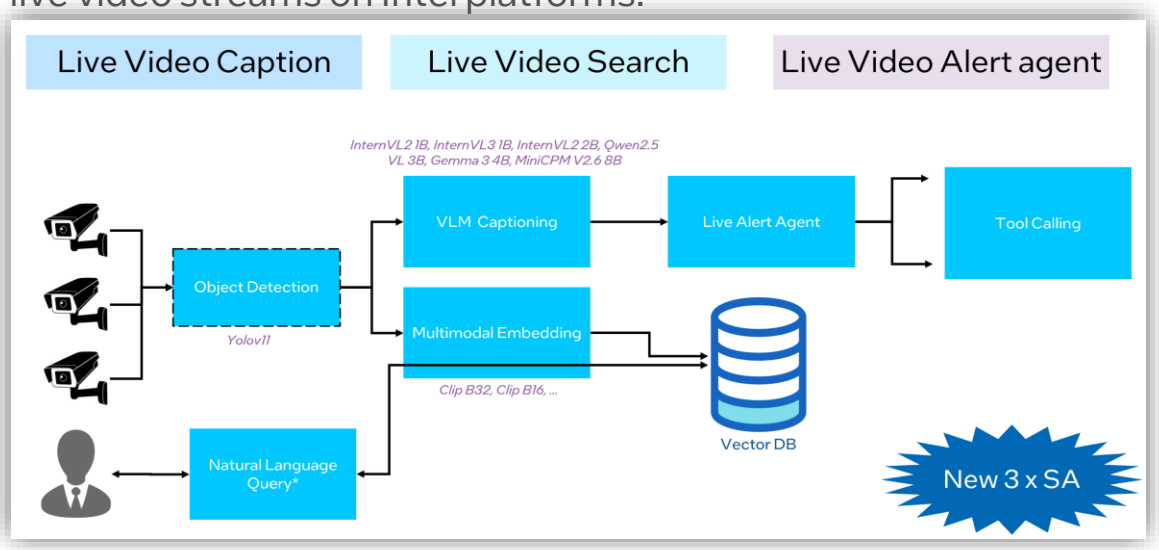
Supported Intel Platforms:

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

Explore details →

Live Video Applications

Real-time, AI-powered analytics for live video streams on Intel platforms.



Live Video Captioning

Generates real-time, AI-powered captions from live video streams to improve accessibility and situational awareness.

Supported Intel Platforms:

- Intel® CoreTM, Intel® CoreTM Ultra processors

Explore details →

Live Video Search

Enables natural-language search across live and recorded video streams using AI-generated metadata and embeddings to quickly find relevant moments.

Supported Intel Platforms:

- Intel® Xeon® processors, Intel® Arc™ Graphics

Explore details →

Live Video Alert Agent

New

Monitors live video streams in real time to detect events and trigger configurable AI-driven alerts for rapid response.

Supported Intel Platforms:

- Intel® Core™, Intel® Xeon® processors

Explore details →

Live Video Captioning RAG

Converts live video captions into a RAG-powered knowledge base for natural-language search and contextual insights.

Supported Intel Platforms:

- Intel® CoreTM, Intel® CoreTM Ultra processors

Explore details →

NVIDIA → Intel: Extend or Migrate

Bring your video analytics to Intel

Extend to Intel

Run your existing stack, now on Intel too

- **Media decode:** Intel GPU decode runs alongside NVDEC
- **Pipelines:** DL Streamer runs side by side with DeepStream
- **Inference:** your models run on Intel via OpenVINO, unconverted
- **Frameworks:** OpenVINO runs next to your CUDA stack
- **Serving:** Triton + OpenVINO backend
- **Hardware:** GPU + Intel iGPU, NPU & CPU

Explore details →

OR

Migrate

Full switch to the Intel stack

- **Media decode:** NVDEC → Intel GPU decode
- **Pipelines:** DeepStream → DL Streamer
- **Inference:** TensorRT → OpenVINO IR
- **Frameworks:** CUDA stack → OpenVINO
- **Serving:** Triton → OpenVINO Model Server
- **Hardware:** GPU-only → CPU + iGPU + NPU

Explore details →

Standardize on Intel — broader reach, lower TCO

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.

Application Security Enablement 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.

Explore details →

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

Explore details →

Metro SDK Overview

Metro AI Suite for Solution Development

Metro SDKs

- 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

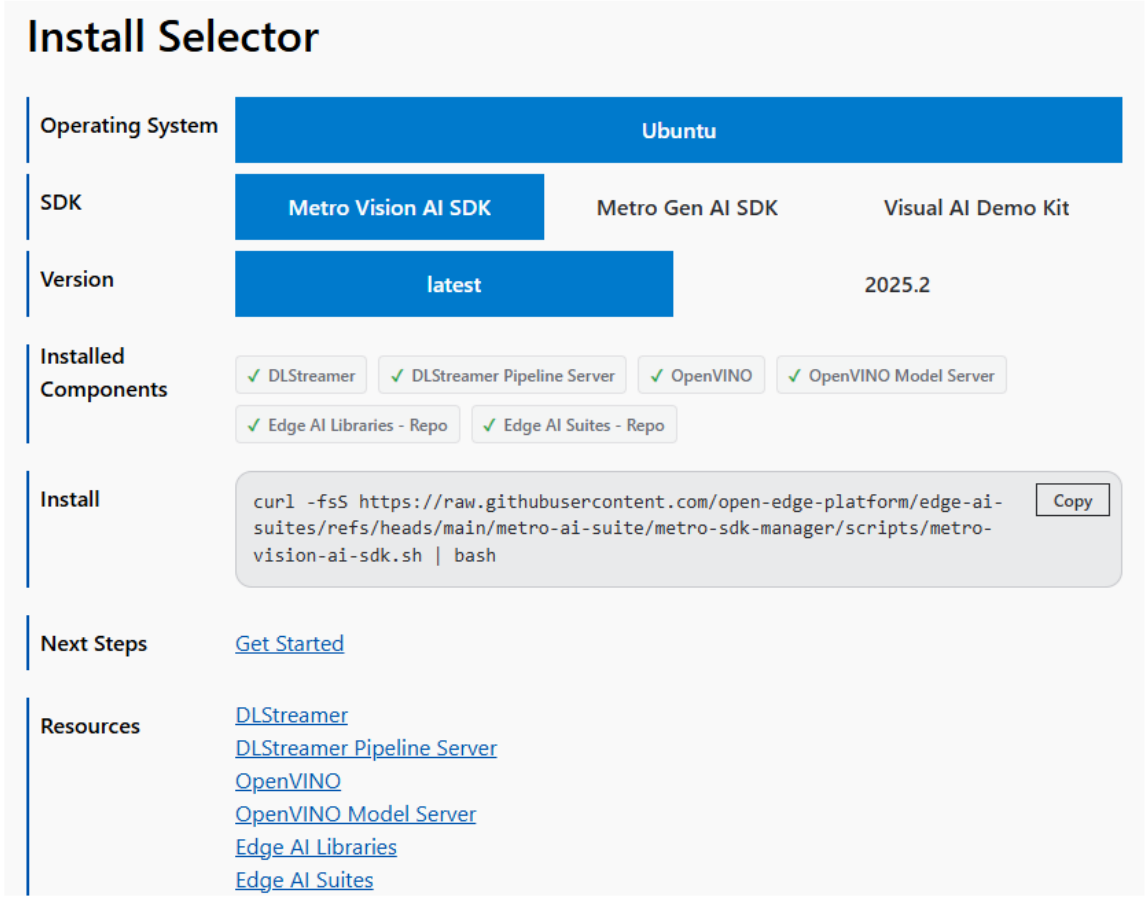
Comprehensive development environment for generative AI applications.

Visual AI Demo Kit

Comprehensive demonstration environment for computer vision applications.

Metro 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

Explore details →

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

- Media MTX
- Eclipse Mosquitto MQTT Broker
- Grafana
- Node-Red

Metro Gen AI SDK

comprehensive development environment for generative AI applications

- Multimodal Embedding Serving microservice
- VDMS based Data Preparation Microservice
- VLM OpenVINO serving microservice
- Audio Analyzer
- Vector Database

OpenCV + Ffmpeg

GStreamer

Video / Media processing libraries, oneVPL, LibVA, VAAPI

DLStreamer + DLStreamer Pipeline Server

OpenVINO + OpenVINO Model Server

Operating system + Drivers (Ubuntu)

intel. ATOM | intel CORE 3 | intel CORE 5 | intel CORE 7 | intel CORE ULTRA 5 | intel CORE ULTRA 7 | intel XEON | intel. ARC GRAPHICS | intel. iRISXe GRAPHICS

3rd Party

Intel®

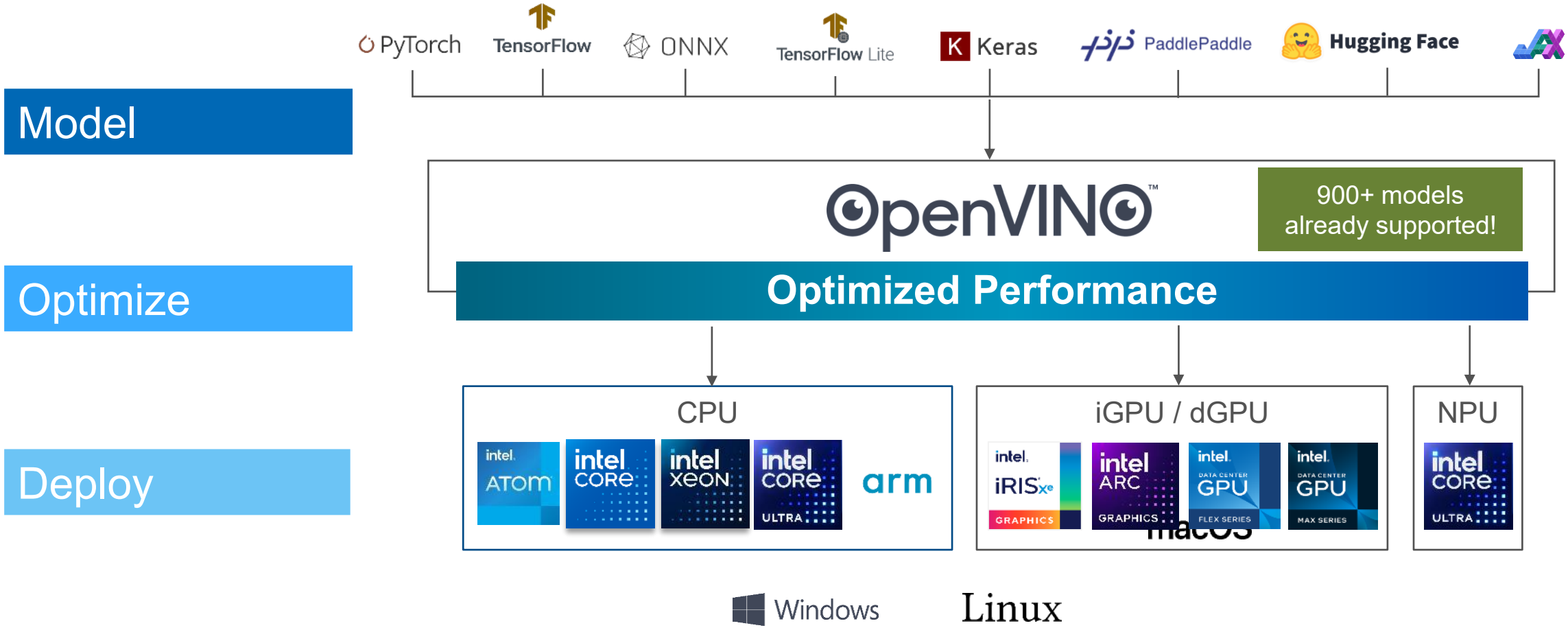
Intel® AI – Optimized for the Edge

Conversion Guide

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

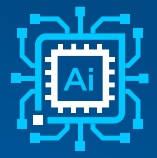
ISV Applications/Solutions

OpenVINO Open-Source Software for AI Inference Optimization and Deployment



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

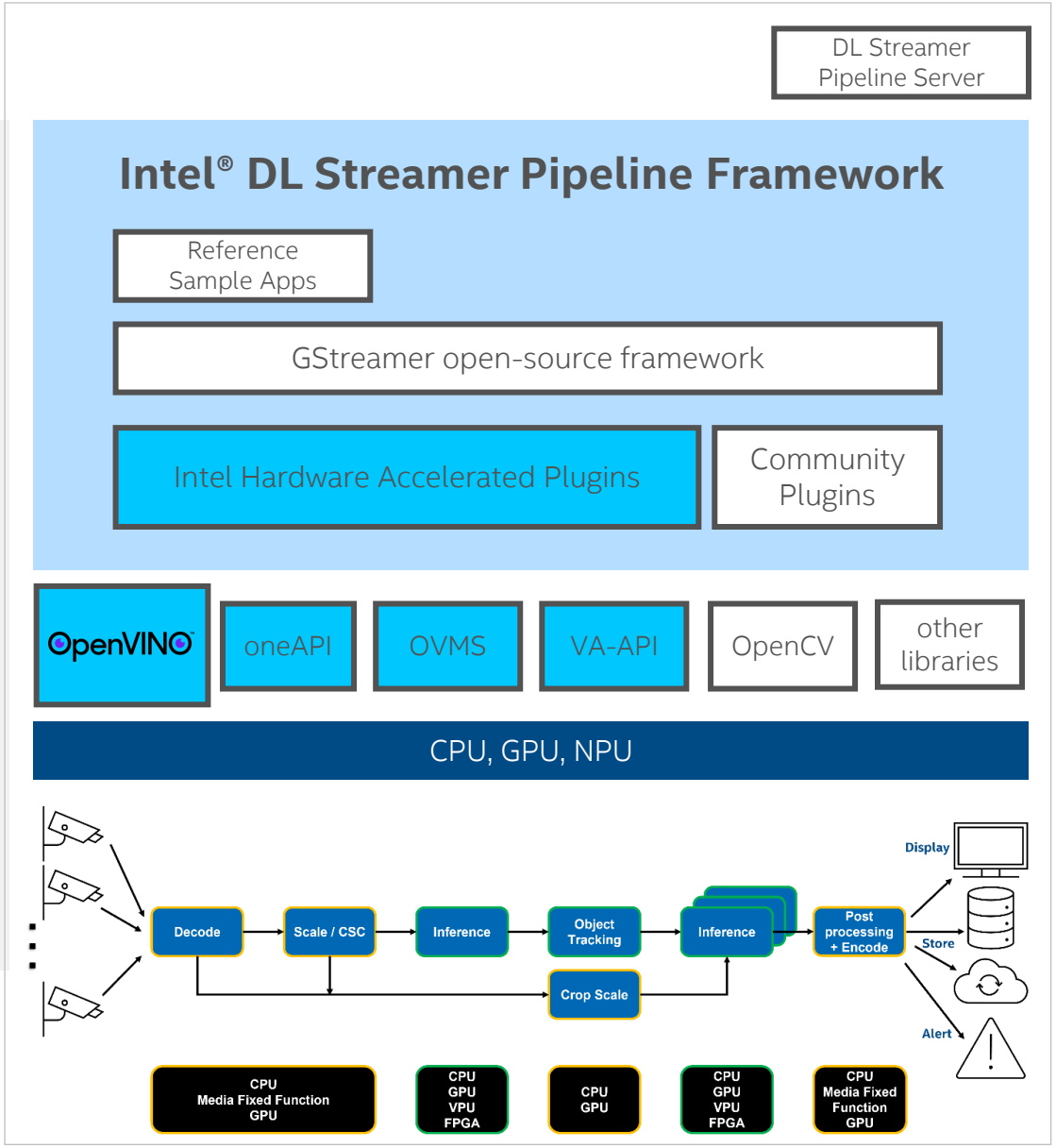
Deep Learning Streamer

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

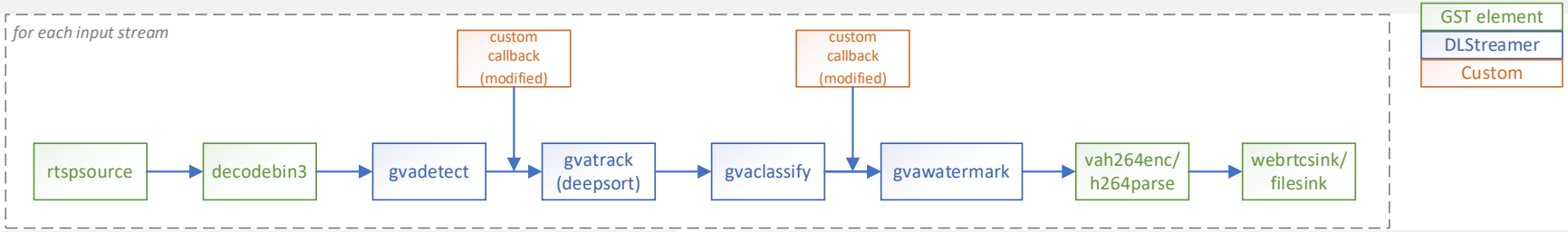
Write once, deploy on any Intel platforms, from Edge to Cloud

Latest DLStreamer content



Why DL Streamer

DL Streamer = GStreamer + Optimized AI



DL Streamer Value Proposition

1. Minimal Change, Maximum Gain
- Built **on top of GStreamer**
- Same pipeline philosophy: elements, pads, caps, buffers

2. Optimized AI Performance
- Out-of-the-box integration with **OpenVINO**
- Hardware acceleration on **Intel CPU, iGPU, dGPU, NPU**
- Zero-copy paths, batching, async inference

3. Production-Ready Video Analytics
- Ready-made elements for:
  - Object detection, classification, tracking
  - Metadata propagation (GVAMeta)
  - ROI handling, analytics overlays

4. Write once, deploy on any Intel platforms, from Edge to Cloud
- Deploy using REST API or command line

Migration from GStreamer → DL Streamer

How hard is it?
Easy and incremental.

What stays the same
- Core GStreamer pipeline
- Decoders, encoders, muxers, sinks
- Pipeline construction (gst-launch, C/C++, Python)

What changes
- Replace or extend elements with DL Streamer plugins:
  - decodebin3 ! Videoconvert
    becomes
    decodebin3 ! gvadetect ! Gvatrack ! Gvaclassify ! gvawatermark
- Add inference configuration (model, device, batch size)

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 !
```

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

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 or concurrent streams for the pipeline

```
[optimizer] [    INFO] - ============================== SUMMARY ==============================
[optimizer] [    INFO] - Optimized pipeline found with 109.97 fps improvement over the original pipeline.
[optimizer] [    INFO] - Original pipeline FPS: 45.36
[optimizer] [    INFO] - Optimized pipeline: filesrc location=1192116-sd_640_360_30fps.mp4 ! decodebin3 ! vapostproc ! v
ideo/x-raw(memory:VAMemory) ! gvadetect model=/home/pbartosi/models/public/yolo11s/INT8/yolo11s.xml model-instance-id=in
f0 pre-process-backend=va-surface-sharing device=GPU batch-size=4 ! queue ! gvafpscounter ! fakesink sync=false filesrc
location=1192116-sd_640_360_30fps.mp4 ! decodebin3 ! vapostproc ! video/x-raw(memory:VAMemory) ! gvadetect model=/home/p
bartosi/models/public/yolo11s/INT8/yolo11s.xml model-instance-id=inf0 pre-process-backend=va-surface-sharing device=GPU
batch-size=1 ! queue ! gvafpscounter ! fakesink sync=false
[optimizer] [    INFO] - Optimized pipeline FPS: 155.33
[optimizer] [    INFO] - ================================================================
```

45 fps → 155 fps

DLStreamer speaks Gstreamer Natively

Inference results stored in GstAnalyticsRelationMeta - readable by any GStreamer element, zero adaptation needed

WRITES
DL Streamer elements

- **gvadetect** → GstAnalyticsODMtd — Object detection (bbox, label, confidence, rotation)
- **gvatrack** → GstAnalyticsTrackingMtd — Object tracking (ID, timestamps)
- **gvaclassify** → GstAnalyticsClsMtd — Classification (confidence + label)

+ GstAnalyticsGroupMtd (17× KeypointMtd + skeleton edges) - written by gvadetect or gvaclassify when using a pose estimation model

WRITE →

GStreamer Buffer
GstAnalyticsRelationMeta (standard upstream container)

- **GstAnalyticsODMtd** — bbox, label, confidence, rotation
  - RELATE TO → **GstAnalyticsTrackingMtd** — Object tracking (ID, timestamps)
  - CONTAIN → **GstAnalyticsClsMtd** — Classification (confidence + label)
  - CONTAIN → **GstAnalyticsGroupMtd** — Ordered group of metadata

READ →

READS
Any GStreamer element

- **Gvawatermark** — renders bboxes, labels, skeleton overlays
- **Custom Gstreamer Python Element** — Python callback
- **Custom sink** — any 3rd-party GStreamer sink standard GstAnalytics API
- **Non DLStreamer tool** — upstream GStreamer ecosystem element - fully compatible

No vendor lock-in
Standard upstream GstAnalytics API – not a proprietary metadata format

Future-proof
Aligned with GStreamer

C, C++, Python – one API
gives identical metadata access in every language

New

AI CODING AGENT (PREVIEW): From plain language to a running pipeline

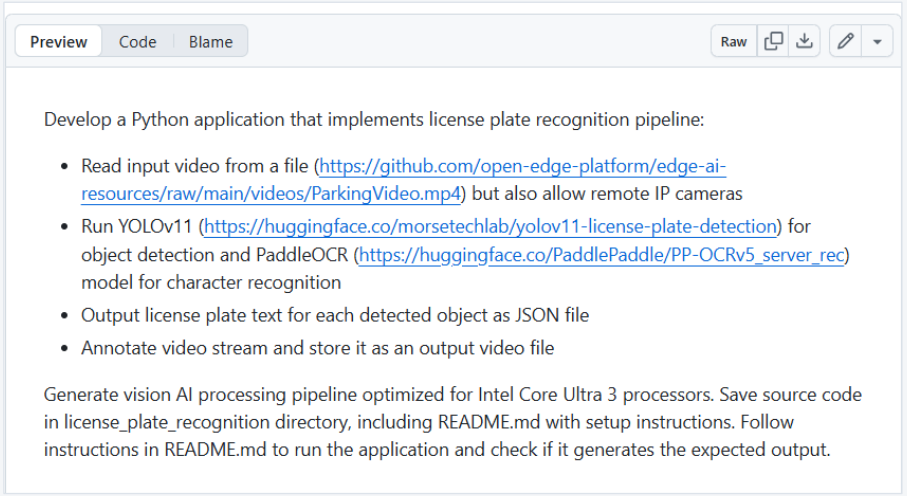
A skill that equips your AI coding agent — GitHub Copilot Chat, Cursor, or Claude Code — to turn a natural-language description into a complete, ready-to-run DL Streamer project.

Describe the pipeline. The agent ships the project.

Describe what you want; the agent generates a complete DL Streamer project — code, configuration, and run instructions — ready to execute on Intel hardware.

- ✓ **Prompt to running app in minutes**
  A written request becomes a working pipeline in minutes.
- ✓ **Lower barrier to entry**
  Accelerate development without deep platform specialization
- ✓ **Faster prototyping on Intel**
  Evaluate and iterate faster on Intel platforms.

PROMPT · RESPONSE



Example: License Plate Recognition — built end-to-end with the agent

Explore details →

New

AI CODING AGENT (PREVIEW): DeepStream to DL Streamer Migration

The same AI coding agent reproduces NVIDIA's DeepStream License Plate Recognition workflow on Intel hardware

One prompt recreates the pipeline on Intel

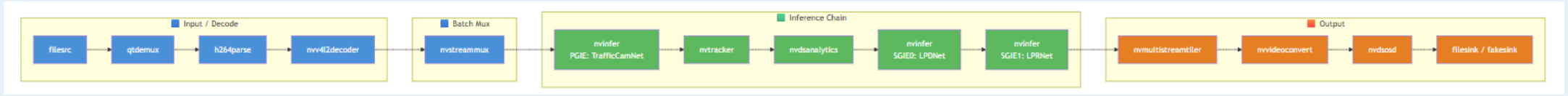
From a single prompt, the agent analyzes the DeepStream License Plate Recognition app and generates an equivalent DL Streamer project — reproducing the end-to-end workflow on Intel hardware.

✓ **Runs on any Intel accelerator**
Targets Intel GPU, CPU, or NPU, selected with a --device argument.

✓ **Standard GStreamer metadata**
Uses the GStreamer GstAnalytics metadata API instead of the vendor-specific pyds SDK.

✓ **Open, optimized models**
Open YOLO11 and PaddleOCR exported to OpenVINO IR (INT8) replace the proprietary TrafficCamNet model.

BEFORE · ORIGINAL DEEPSTREAM PIPELINE



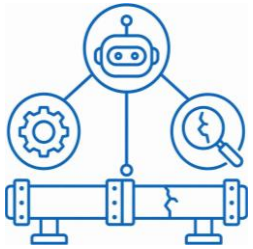
AFTER · CONVERTED PIPELINE ON INTEL HW · DL STREAMER

Example License Plate Recognition, recreated end-to-end on DL Streamer

New

Solution Blueprints

Intel-validated software blueprints enabling Agentic Predictive Maintenance for critical infrastructure and Digital Twins for smart buildings



Vision-AI Based Agentic Predictive Maintenance

Multi-agent reasoning for failure detection, prediction, policy enforcement, and inspection audits.

- Vision inference, multi-agents that can be deployed in one Edge Platform or across distributed edge deployment
- Detects six pipeline defect classes in real time
- Interactive chat for report generation and audit compliance check
- Uses Public Open Data Set

Explore details →

Multimodal Predictive Maintenance

Fuses edge vision with chemical-sensor data at inference time for more reliable detection in safety-critical environments.

- Real-time multi-modal image and sensors data fusion
- Cuts false positives under ambiguous conditions
- Degrades gracefully when a camera or sensor fails
- Uses Public Open Data Set

Explore details →

Smart Building Digital Twin

Intel SceneScape unifies multi-camera vision into a live 4D digital twin where agents reason on real-world scene data, not pixels.

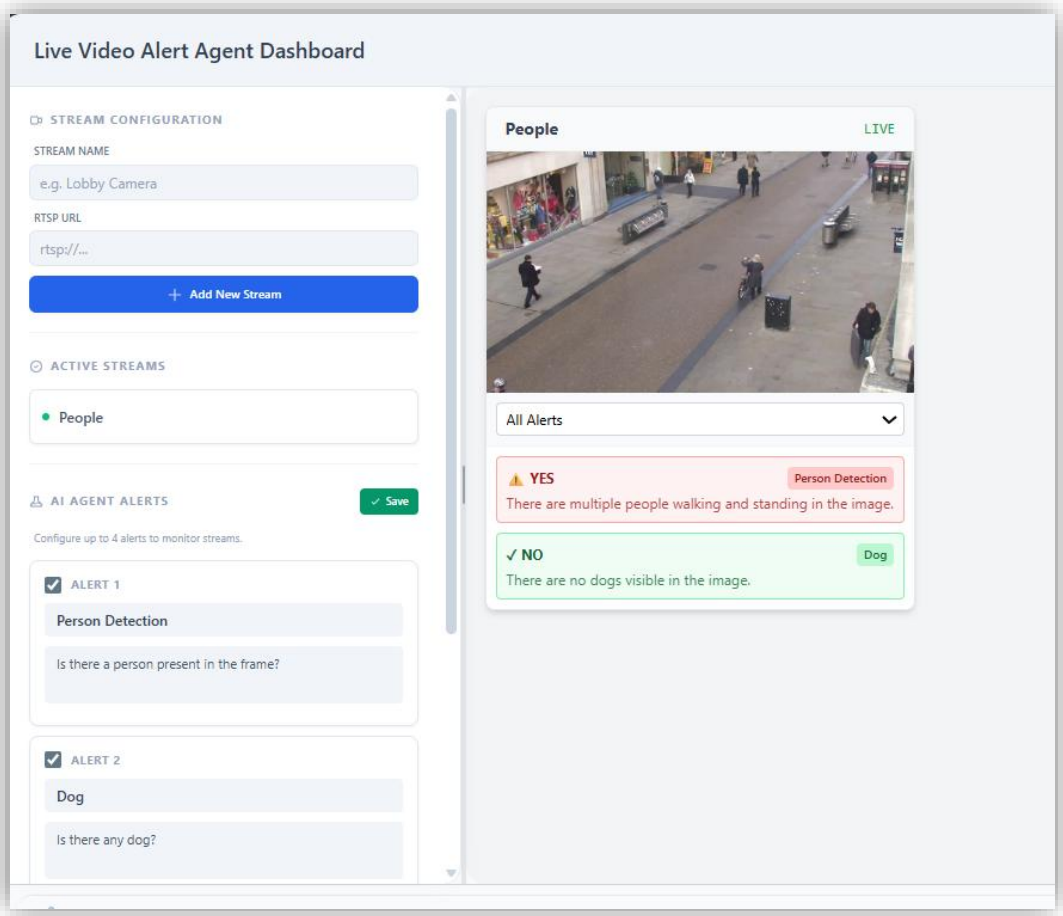
- Tracks every person, object, and door with metric position and history
- Powers access control, occupancy, and safety beyond video analytics

Explore details →

Agentic AI

Live Video Alert Agent

AI-powered video alerting for RTSP streams using OpenVINO™ Vision Language Models and natural-language prompts



RTSP camera streams are analyzed in real time with user-defined alert prompts and low-latency dashboard events.

Key Features:

- **Dynamic alert prompts:** Define and modify alerts in real time through the UI without redeploying.
- **Real-time event broadcasting:** Alerts are displayed to a unified dashboard with low latency.
- **Modular architecture:** Decoupled ingestion, VLM analysis, and event broadcasting enable scalable deployment.
- **Telemetry:** Dashboard tracks CPU, GPU, and memory utilization, with health/readiness/metrics endpoints for operational visibility.

Supported Intel Platforms:

- Intel® Core™ Ultra 2 and 3 processors with integrated GPU, Intel® Xeon® platforms with/without Intel® Arc™ GPU

Explore details →

Smart Traffic Intersection Agent

Edge AI service for real-time intersection monitoring, directional traffic density, and VLM-powered traffic insights

Intersection cameras and traffic data streams are analyzed at the edge to estimate density, surface alerts, and support driving recommendations.

Key Features:

- **Real-time intersection analytics:** Monitor traffic conditions and directional vehicle density at the edge.
- **VLM-powered insights:** Use vision language models to interpret scene context and traffic scenarios.
- **Traffic data streaming:** MQTT broker support streams traffic data; camera images are managed for downstream services.
- **Open interfaces:** REST APIs and a UI help configure workflows, review outputs, and integrate other agents.

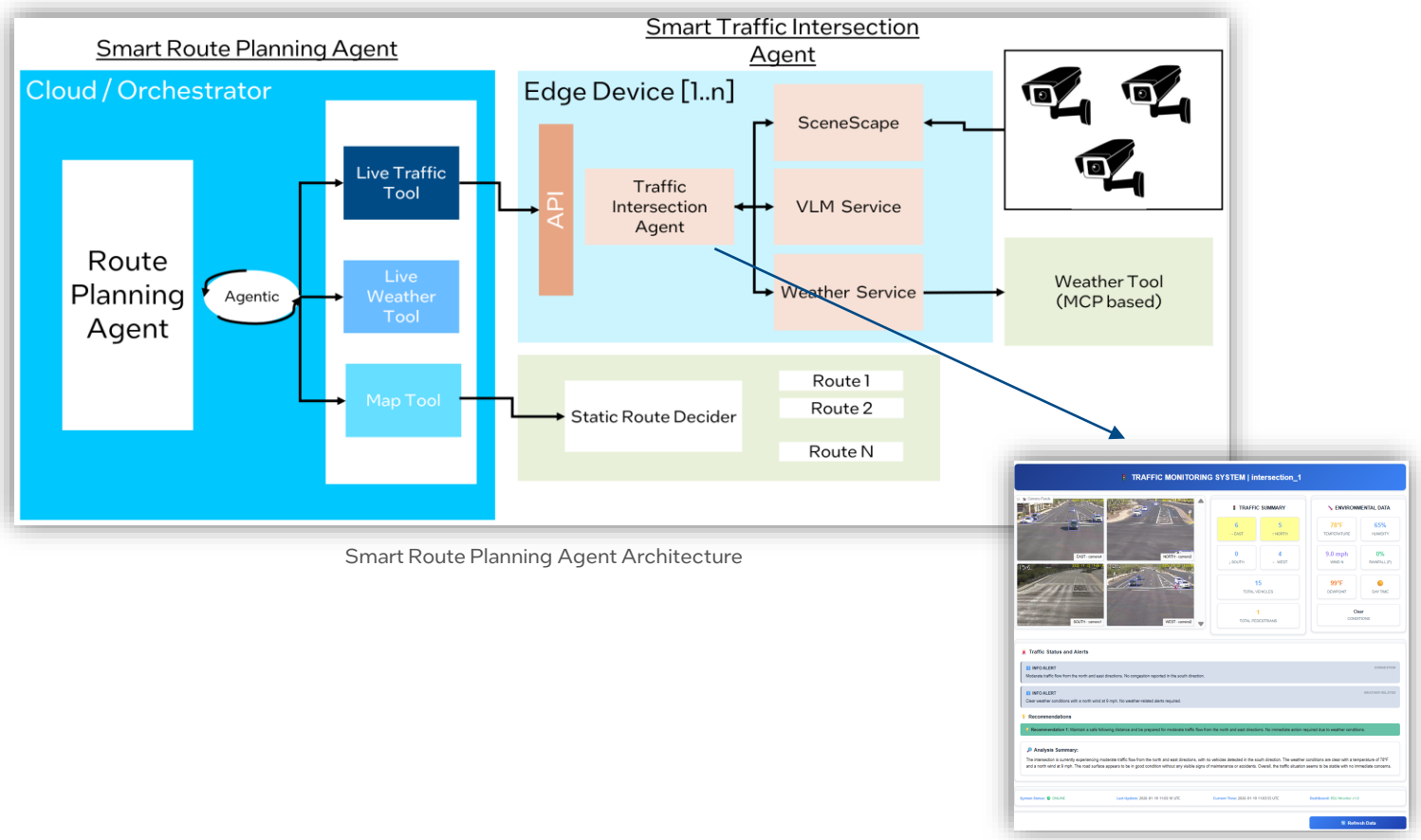
Supported Intel Platforms:

- Intel® Core™ Ultra 2 and 3 processors with integrated GPU or Intel® Xeon® processor

Explore details →

Smart Route Planning Agent

AI-powered route optimization agent that uses multi-agent communication to analyze traffic intersections and find incident-free paths between source and destination in real-time.



Smart Route Planning Agent Architecture

Key Features:

- **Smart Route Planning Agent (Cloud)** orchestrates edge agents and recommends optimal routes using live traffic data
- **Smart Traffic Intersection Agent(Edge)** analyzes traffic patterns, counts vehicles, and identifies congestion causes using SceneScape + VLMs
- **Integrated Tools:** Weather services, live traffic data, and mapping for comprehensive decisions

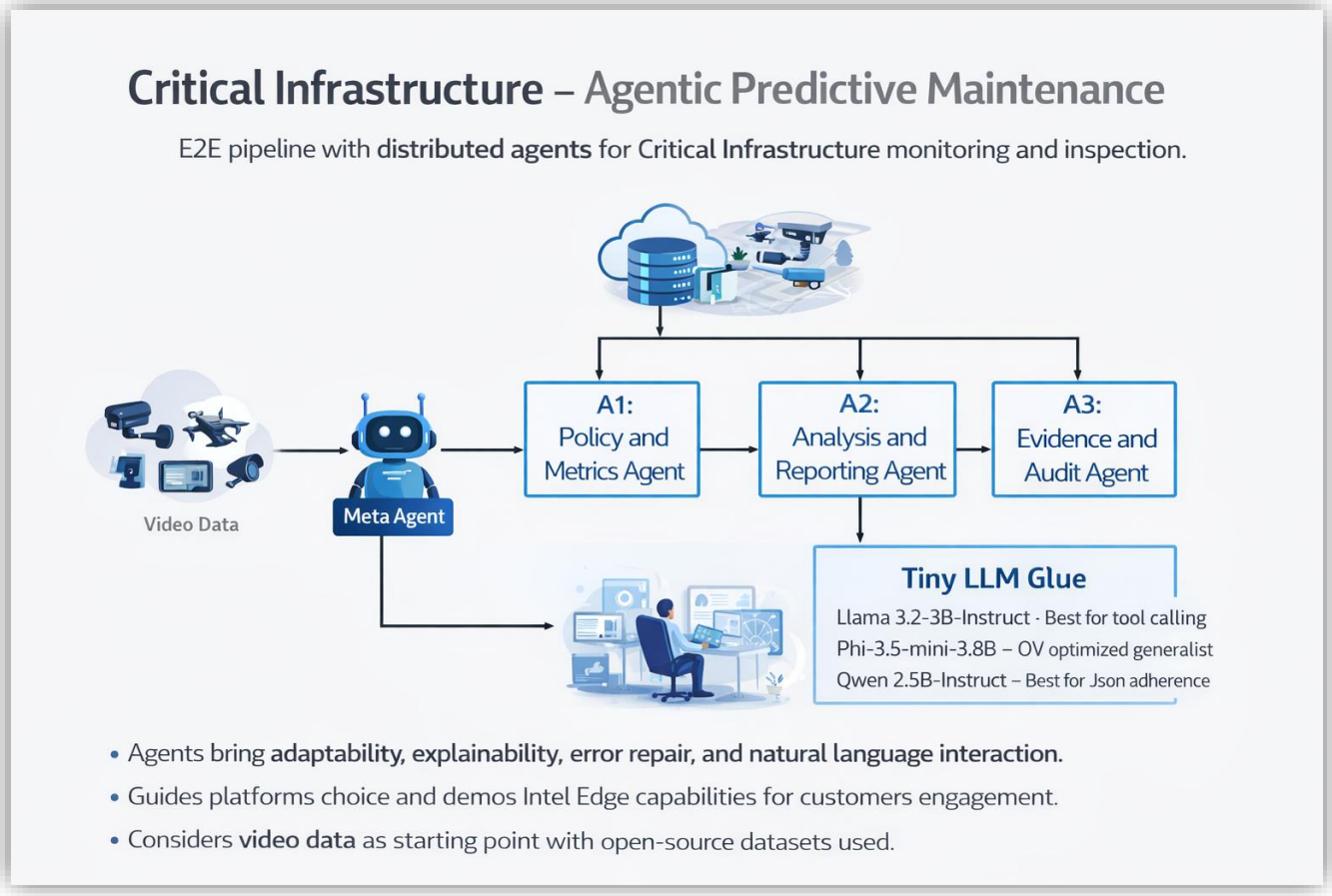
Supported Intel Platforms:

- Intel® Xeon® 6 processors

Explore details →

Vision-AI Based Agentic Predictive Maintenance

An end-to-end pipeline powered by distributed AI agents for infrastructure monitoring and inspection.



Key Features:

- **Performs real-time defect detection** from images and video at the edge while supporting industrial inspection scenarios
- Uses a **multi-agent reasoning pipeline** to analyze detections, assess severity, enforce policies, and generate structured, auditable outputs beyond simple AI inference

Supported Intel Platforms:

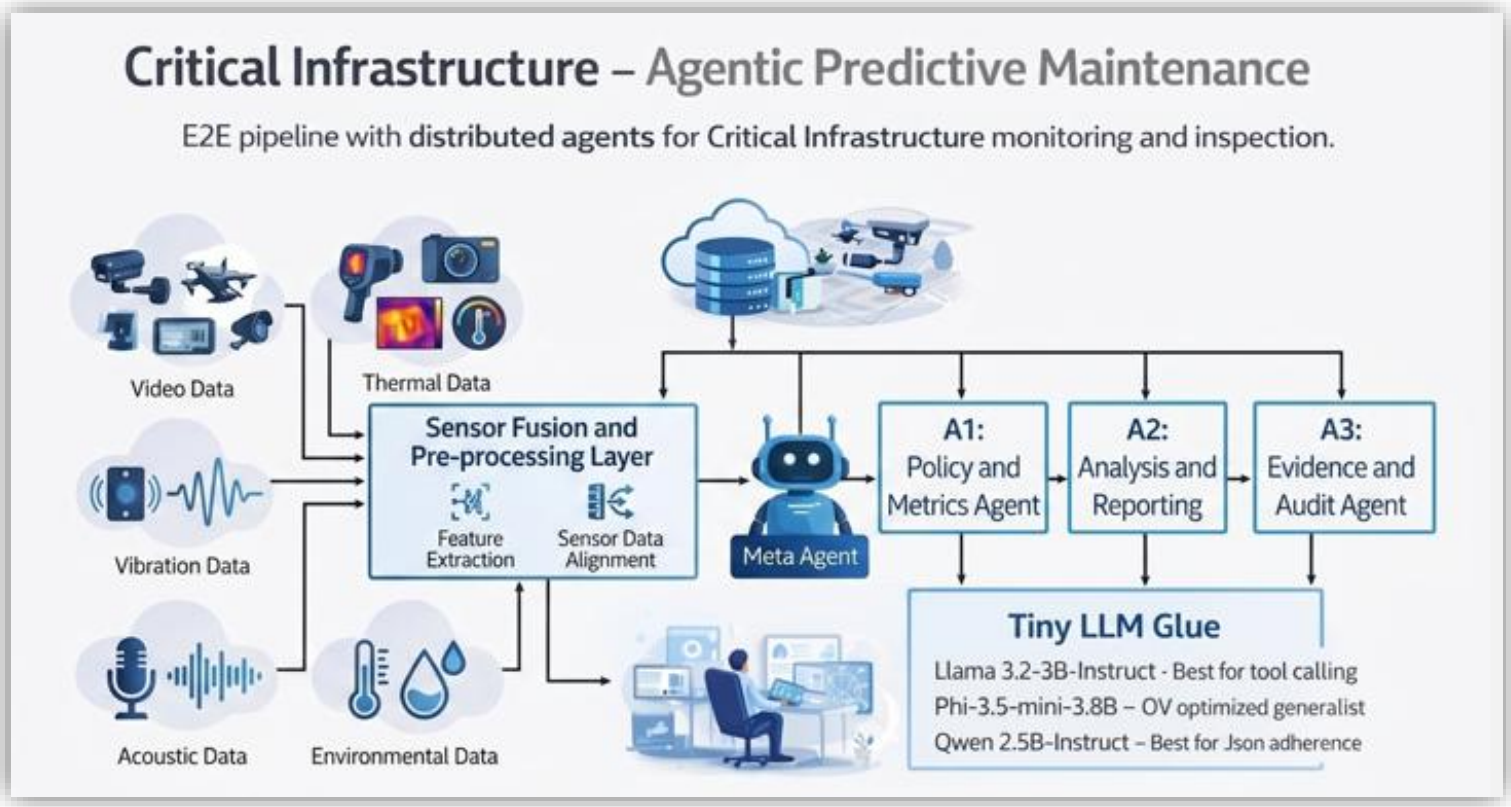
- Intel® CPUs with integrated graphics (12th Gen or later)

Explore details →

New

Multimodal Predictive Maintenance

Multi-modal Agentic pipeline powered by distributed AI agents for infrastructure monitoring and inspection leveraging data from sensors and cameras



Key Features:

- Fuses edge camera vision with chemical sensors for a single, more reliable gas classification with fewer false positives
- Real-time multi-modal image and sensor data fusion that cuts false positives under ambiguous conditions and degrades gracefully when a camera or sensor fails

Supported Intel Platforms:

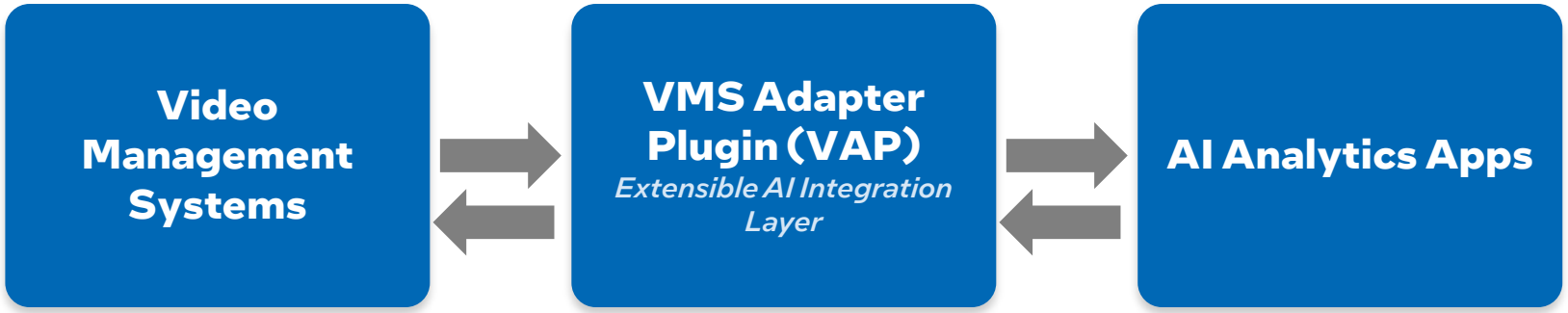
- Intel® CPUs with integrated graphics (12th Gen or later)

Explore details →

New

Integration with Video Management Systems (VMS)

Bring AI analytics into your existing VMS.



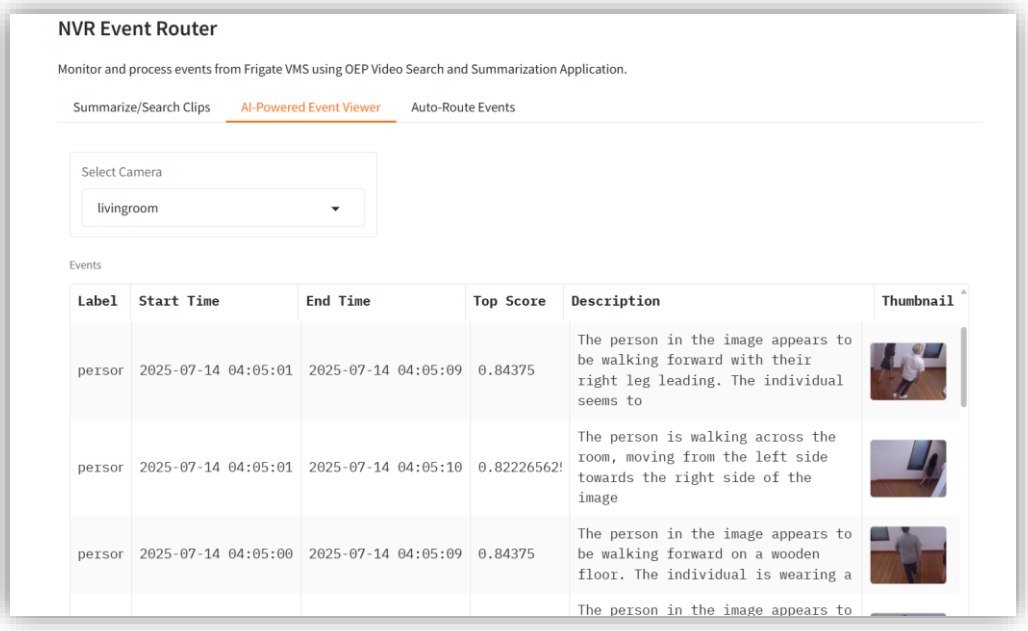
An extensible integration layer connecting video systems to AI analytics

- **Integrate** VMS with AI pipelines — **unlock AI insights**
- **Orchestrate** multi-camera analytics at — **scale effortlessly**
- **Return** results into workflows — **drive decisions**
- **Deploy** with containers — **accelerate time-to-value**

Explore details →

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.



Quickly search, summarize, and understand video streams using AI-powered video analytics.

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® CoreTM, Intel® Xeon® and Intel® ArcTM A-series Graphics, such as A770 GPU

Explore details →

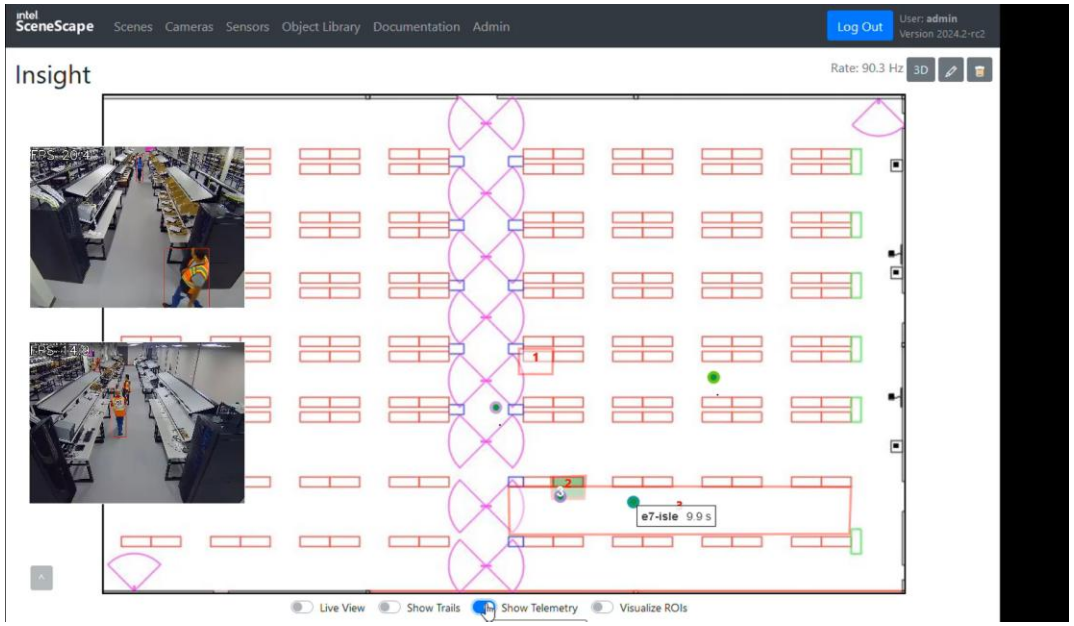
Physical AI

What is SceneScape?

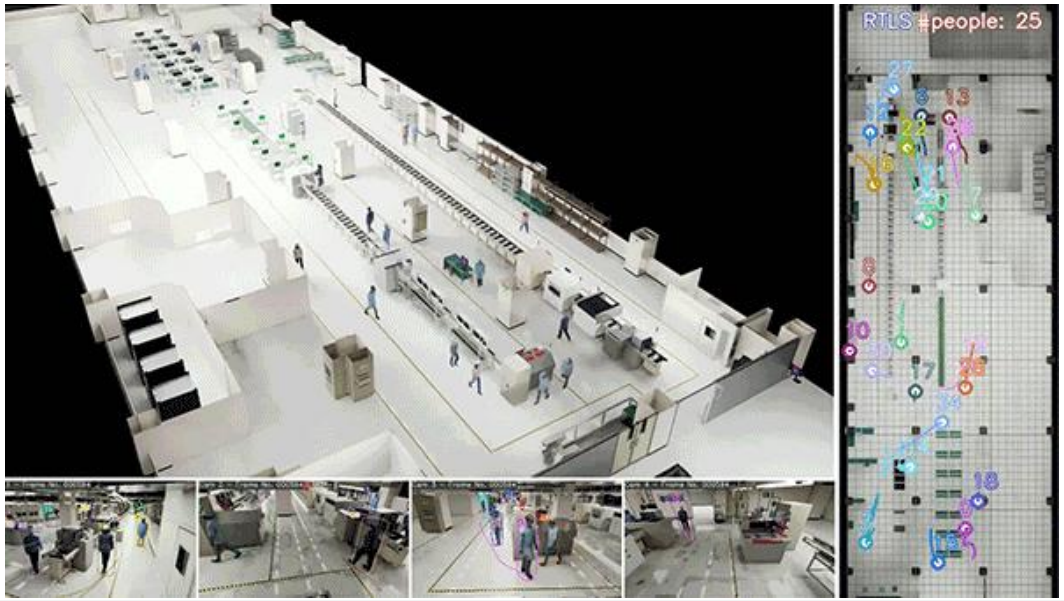
- SceneScape is a **software framework** that reaches beyond vision-based AI to realize spatial awareness from sensor data.
- It transforms data from many sensors to create and provide live updates to a 4-dimensional digital twin of a physical space.
- This spatial data can be applied to use cases including 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
1000 concurrent objects

Intel® Scenescape:



The competition:



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

New

Reliable Tracking in Crowded, Real-World Scenes

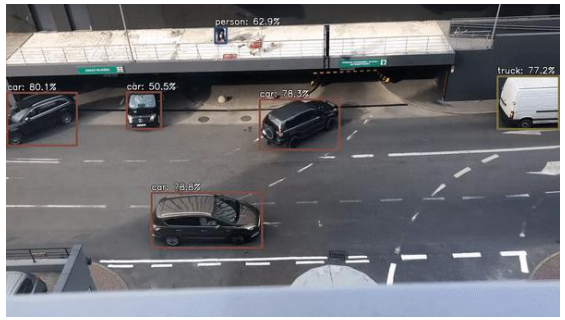
Stronger re-identification and accurate position through partial occlusion — even from a single camera.

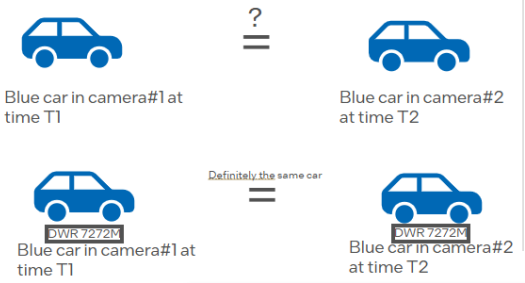
Re-ID Extension

Supports **modern embedding models like CLIP**, adds **cosine similarity**, and publishes **track state** (pending, no match, matched).

Key Benefits

- More accurate recognition across cameras and time
- Trustworthy unique counts with explicit match status
- Drop in new models without re-architecting the pipeline





Occlusion Mitigation

Uses **pose keypoints** to keep position accurate when subjects are **partially blocked**. Works for **any object type**.

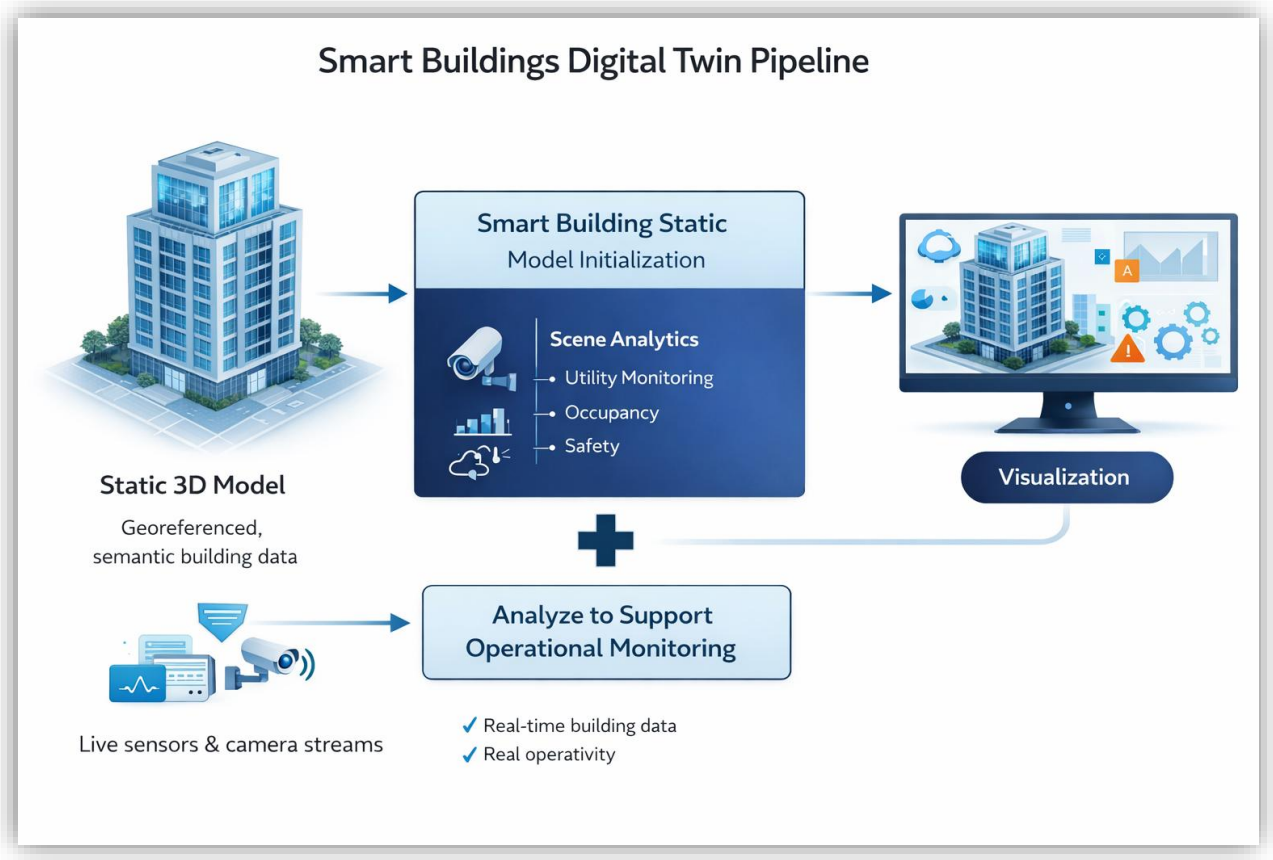
Key Benefits

- Identities don't drop or swap behind obstacles
- Reliable tracking in crowded, cluttered scenes
- Extends beyond people to any tracked object

Get Started with SceneScape

Smart Buildings Digital Twin Blueprint

Advanced pipeline that creates a high-fidelity virtual replica of smart buildings



Key Features:

- Unifies multi-camera feeds into a single **4D digital twin** — a live metric 3D scene where every person, object, and door has precise position, velocity, and history
- Fuses edge vision AI with **environmental and identity sensor data** in real-world metric coordinates with SceneScape
- Runs **intelligent agents on scene data, not pixels** — powering access control, tailgating and object security, plus room occupancy, safety, and comfort

Supported Intel Platforms:

- 10th Gen or newer Intel® Core™ (i5 or higher); 2nd Gen or newer Intel® Xeon®

Explore details →

Partner Support Programs

CASE Enabling & Support Model

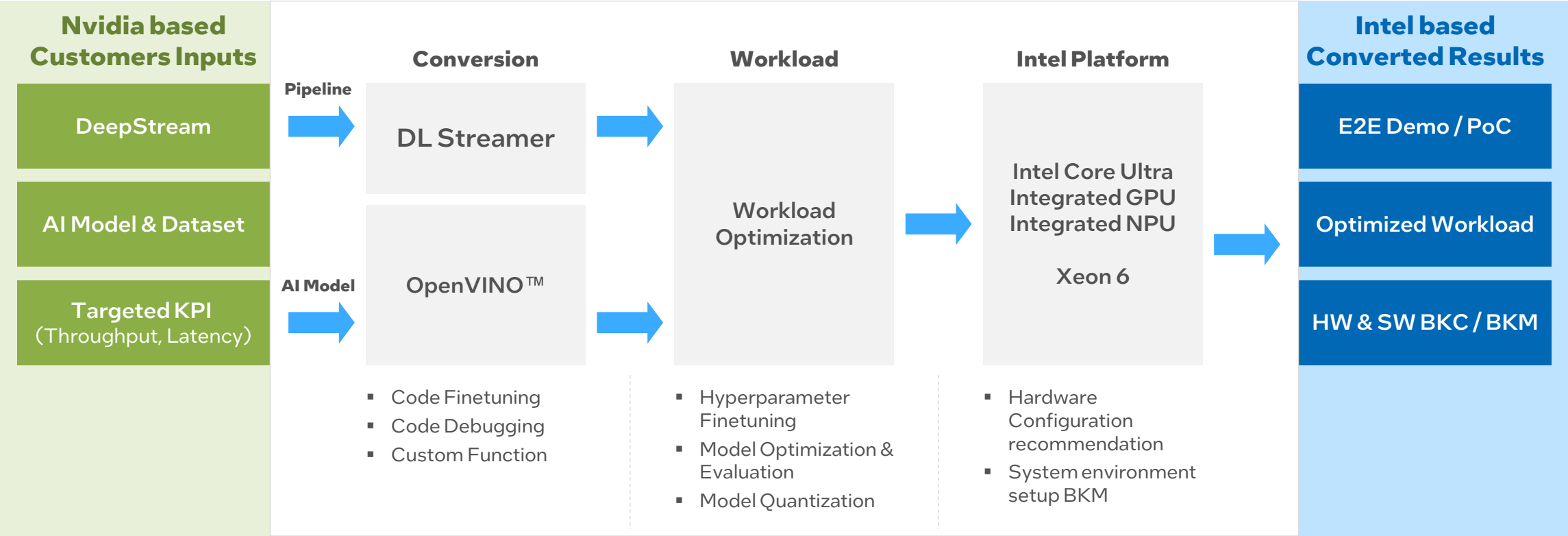
Enabling & Support Level	Intel Support Scope	Support Tools & Channels	Support Examples
Intel® Premium Enabling Services	▪ Per SLA/ TSoW – Custom, Customer Specific ▪ Custom POC/ Demo/ RI ▪ Onsite Support ▪ Priority on Early Access ▪ Pre-launch Preview	Intel Premier Support Portal (IPS) & Email & Phone & 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	**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
People/Face Detection, Image Classification	**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
Gen AI/LLM, Big Data Analytics	**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
Optical Character Recognition, Secured AI	**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
Speech to Text Recognition, Sound Classification				

Conversion Journey and Intel Support*

*Details varies based on project needs



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

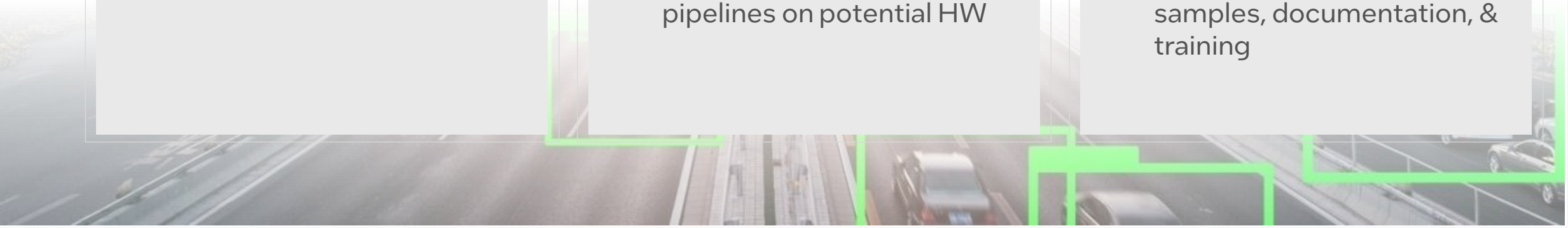
- **Jumpstart AI feature development** by reviewing Proxy Pipelines, Sample Apps, and Blueprints

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

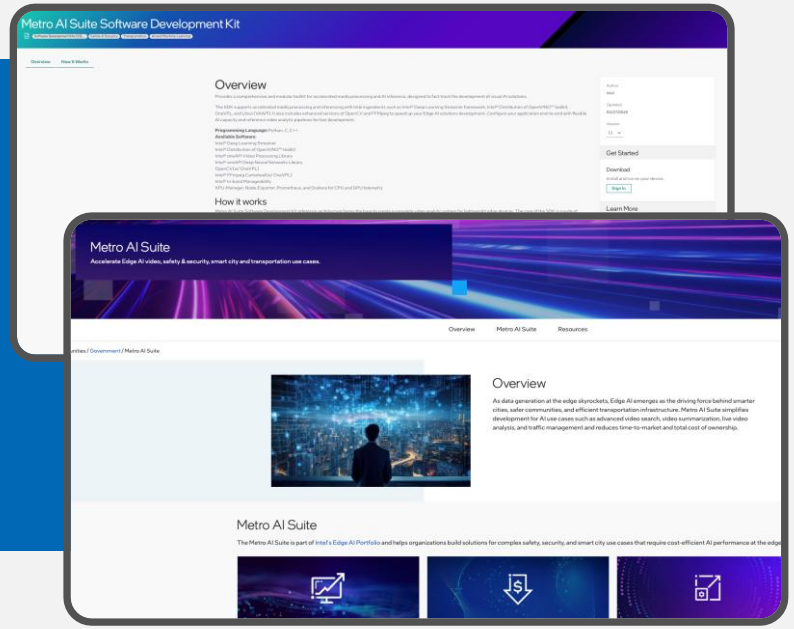


Take the Next Step

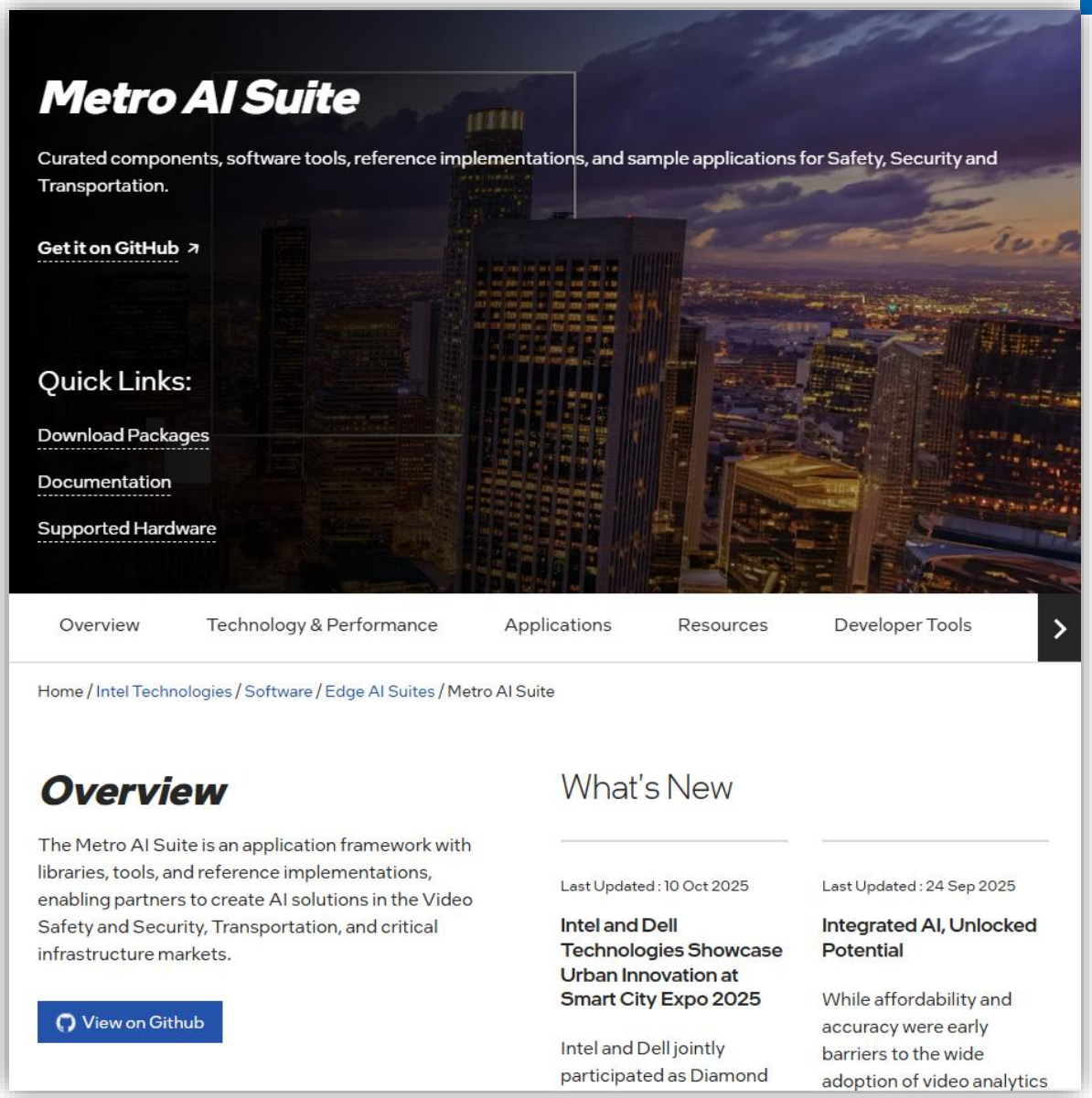
- Software Collection: **starting point** for Video Analytics & AI solution development
- Enables Software Developers to **quickly add Visual & GenAI to existing solutions**
- Helps **maximize performance** of market leading edge AI silicon from Intel

Metro AI Suite

- intel.com/metroai
- Documentation
- Github page



- **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



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; Memory16GB (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.