

# Metro AI Suite Software Developer Guide

Version 1.5 – May 2025 Update



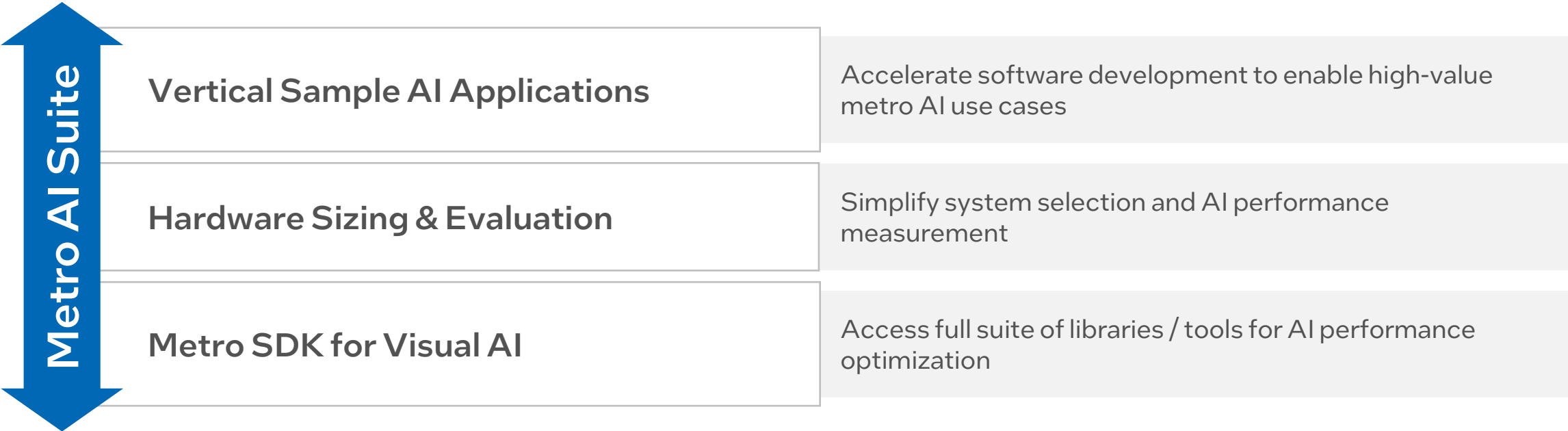
# Metro AI Suite for Software Partners

*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.



**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

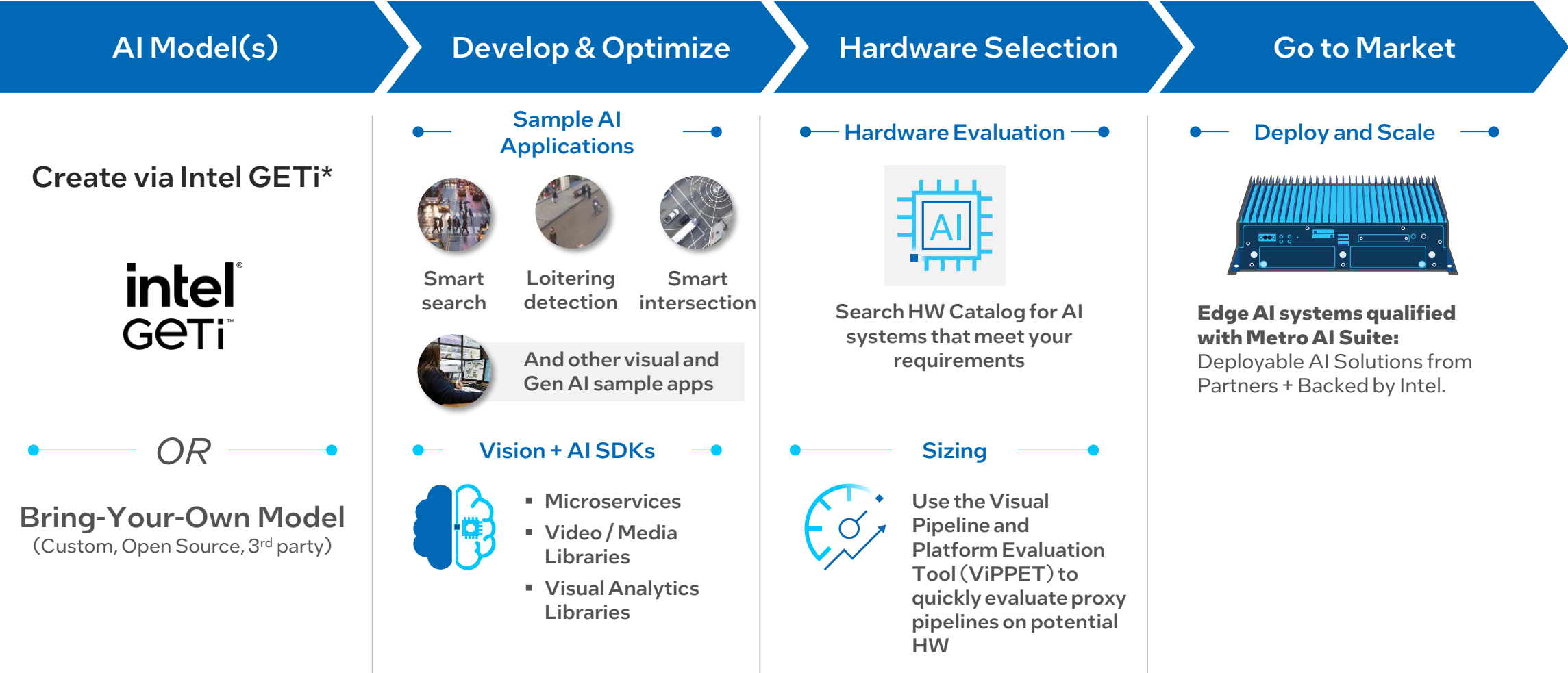


Prioritize open-source, Apache License components



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

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



# Metro SDK Overview

# Metro AI Suite for Solution Development

## Metro SDK



- 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

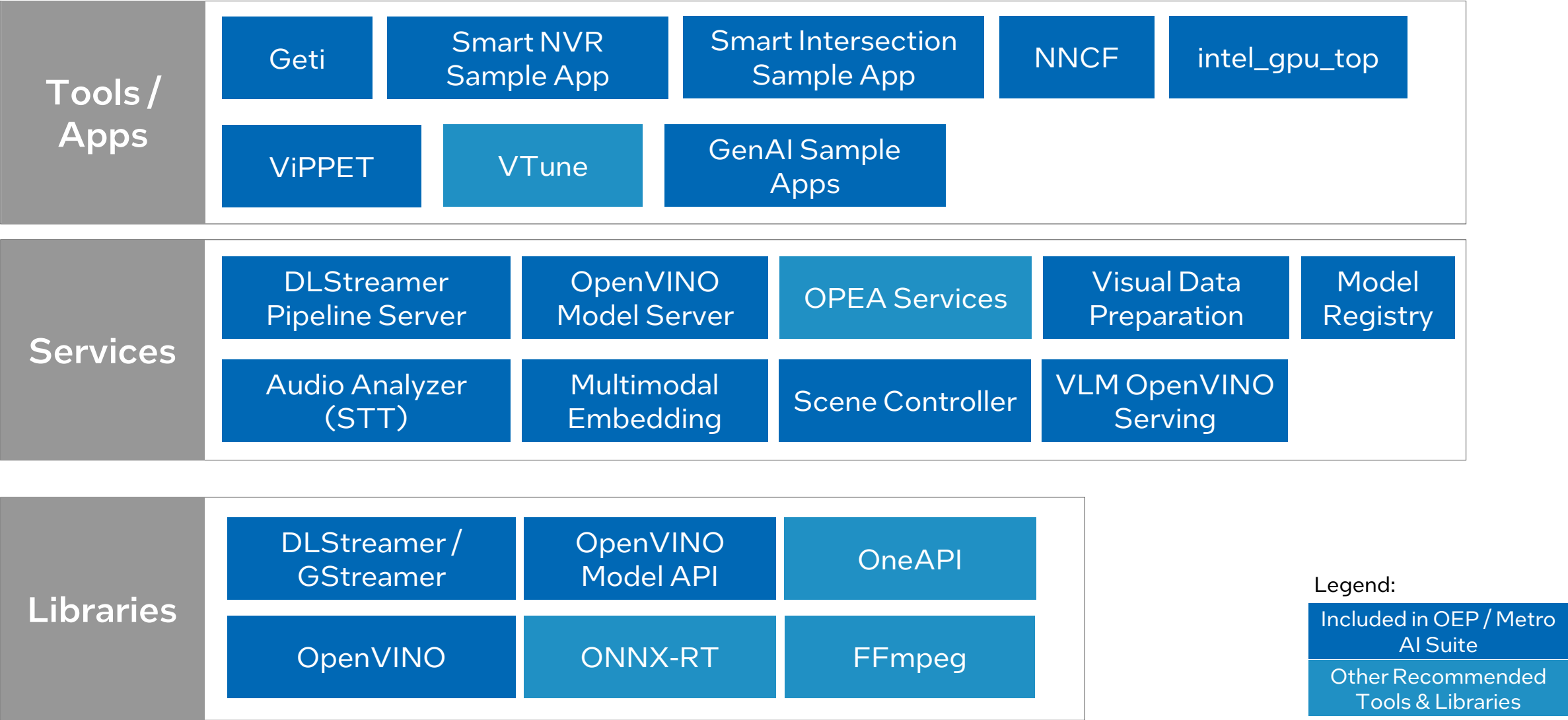
Smart Intersection

Platform Blueprints

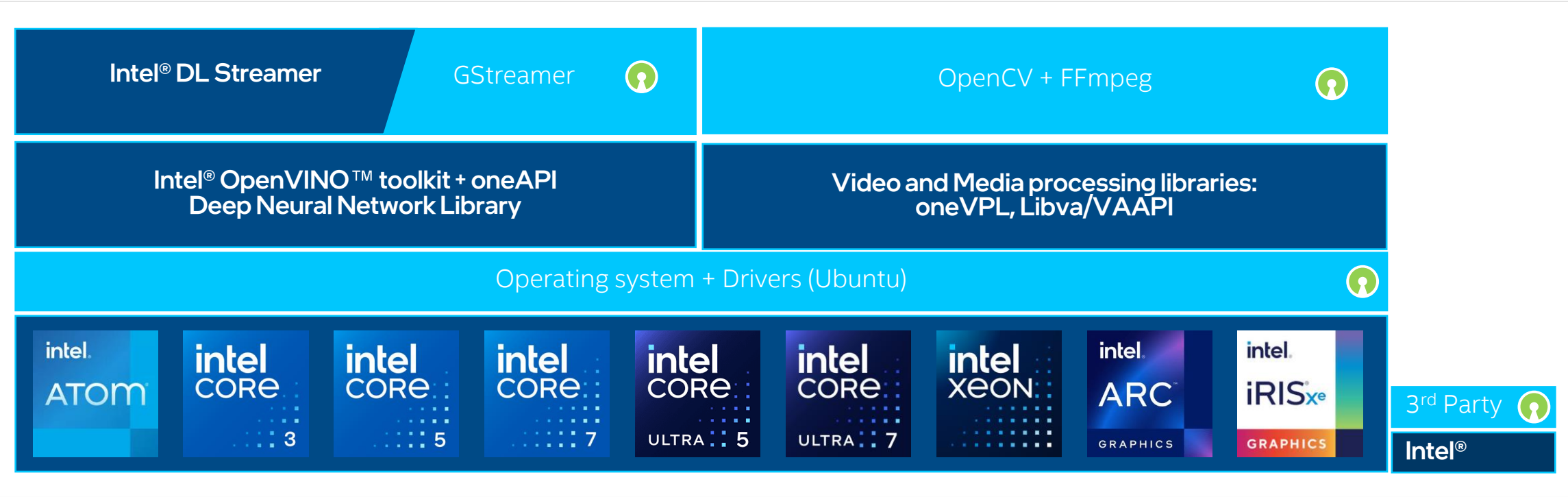
- Quick and easy way to develop Visual and Gen AI pipelines and features
- Use sample app code to apply Intel-optimized best practices
- Review documentation for scalable & heterogenous compute techniques



# Components in Metro AI Suite, Open Edge Platform (OEP), and Other



# Metro AI Suite SDK



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).

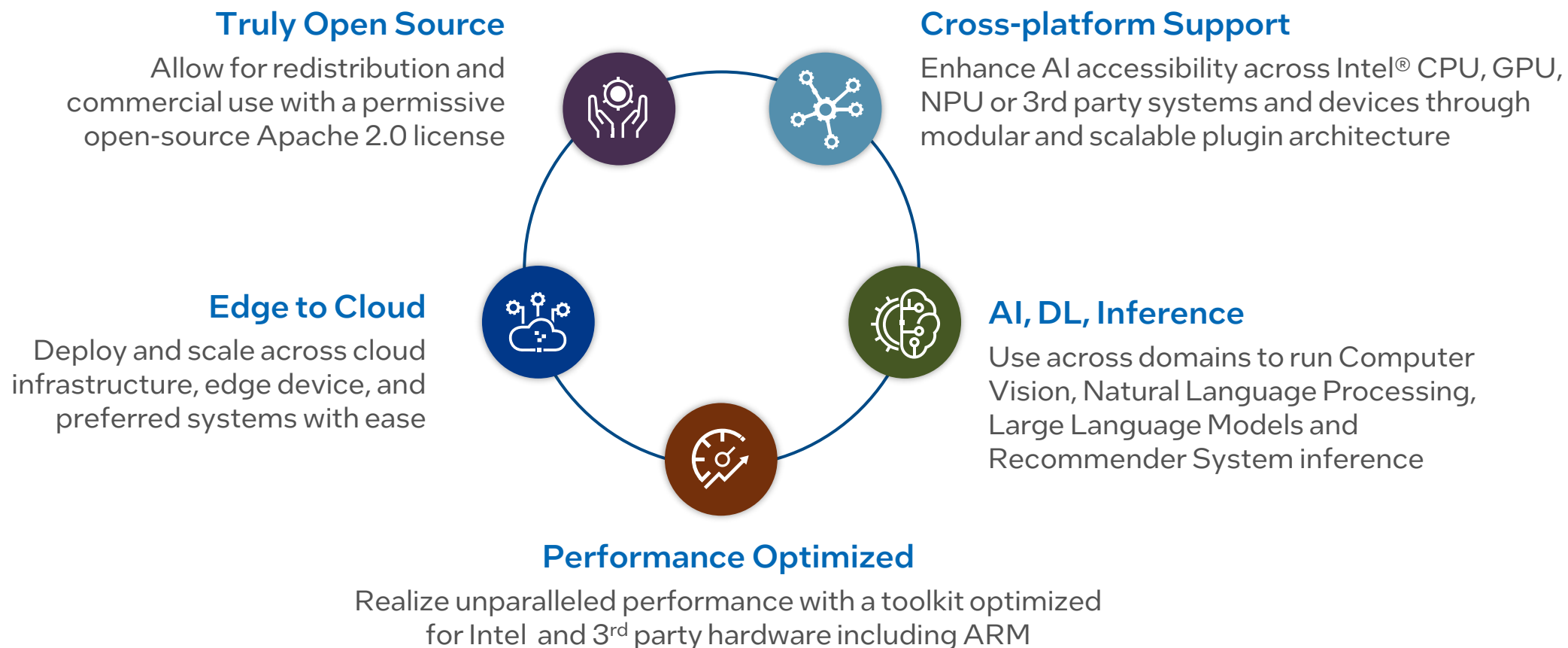
It also includes enhanced versions of OpenCV and FFMpeg to speed up your Edge AI solutions development. Configure your application end-to-end with flexible AI capacity and reference video analytic pipelines for fast development.

# Intel® AI – Optimized for the Edge

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

ISV Applications/Solutions

# OpenVINO™ Toolkit Value Differentiators



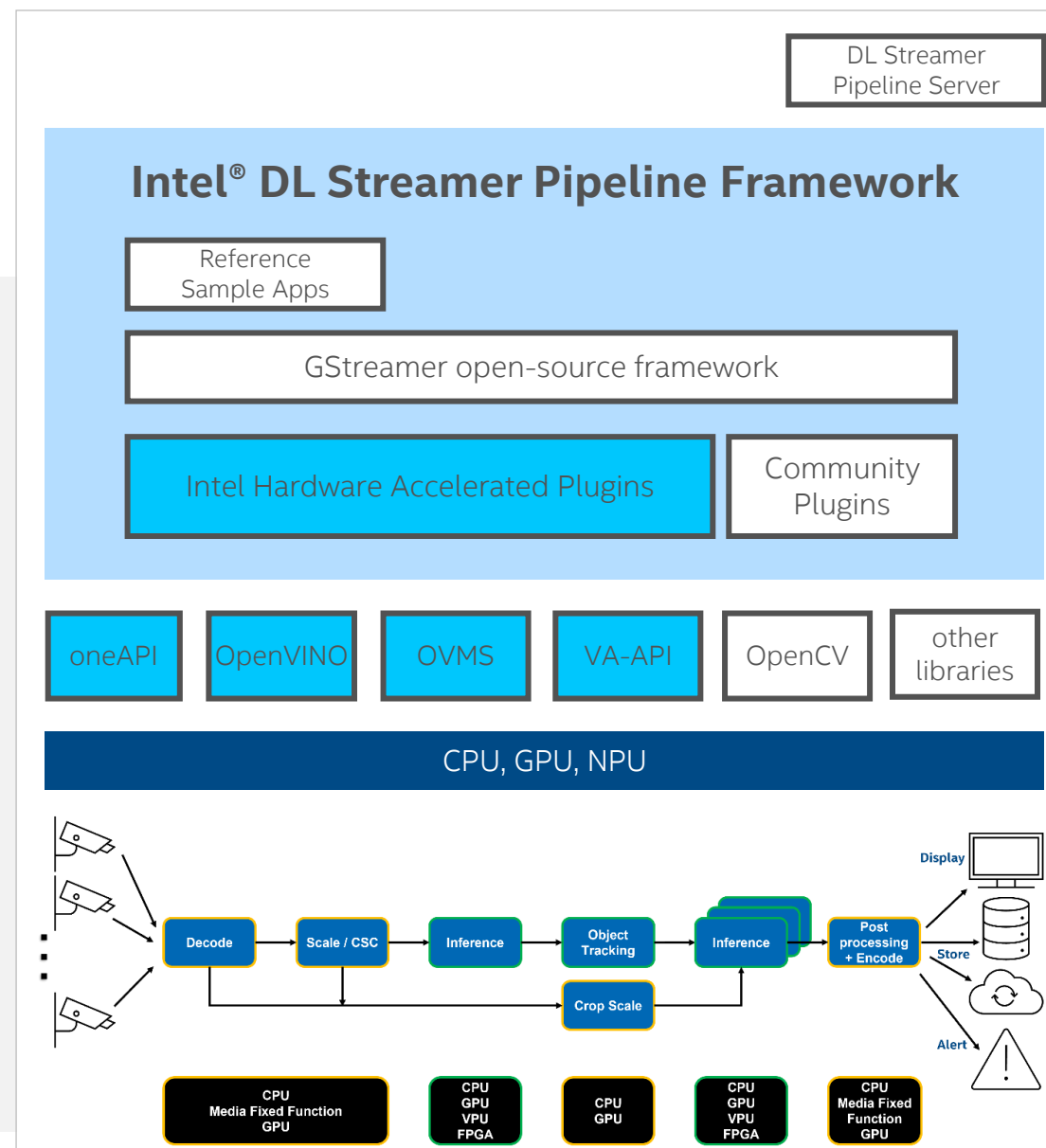
# Intel® DL Streamer



## What is DL Streamer?

- **Intel® Deep Learning Streamer (Intel® 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

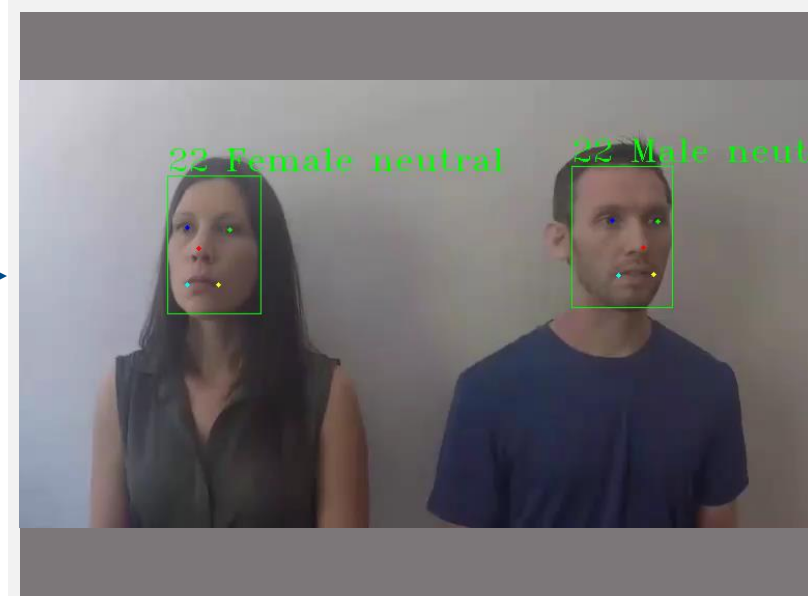
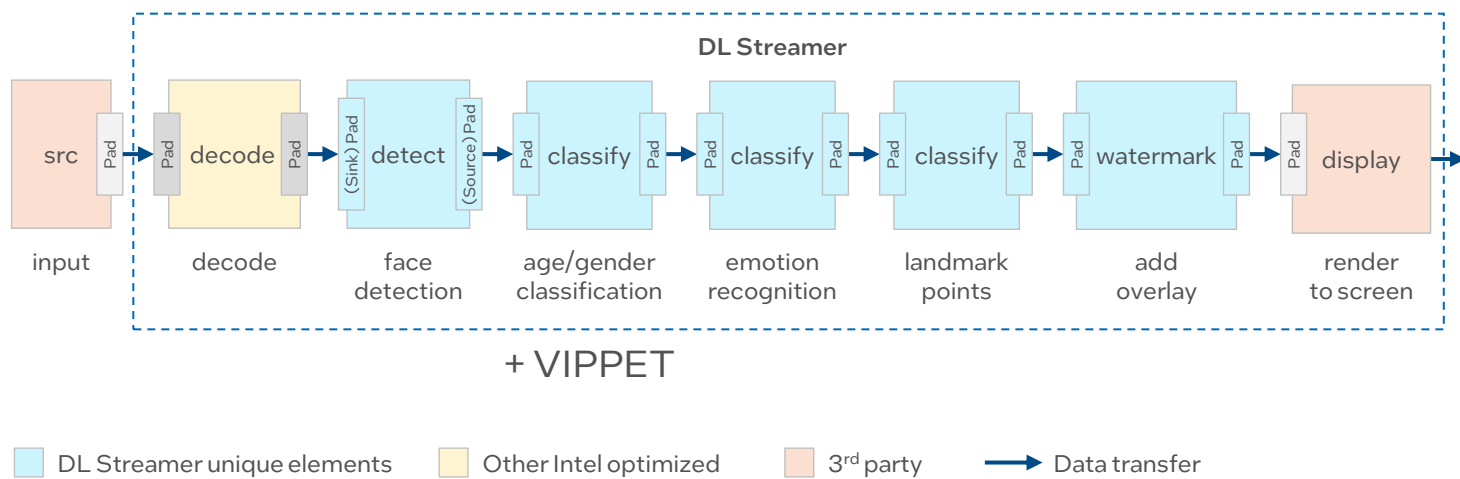
**Write once, deploy on any Intel platforms,  
from Edge to Cloud**



# Intel® 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](https://github.com/openvinotoolkit/open_model_zoo)

# Porting Example

[https://dlstreamer.github.io/dev\\_guide/converting\\_deepstream\\_to\\_dlstreamer.html](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
<a href="#">nvinfer</a>	<a href="#">gvadetect, gvaclassify, gvainference</a>
<a href="#">nvdsosd</a>	<a href="#">gvawatermark</a>
<a href="#">nvtracker</a>	<a href="#">gvatrack</a>
<a href="#">nvmsgconv</a>	<a href="#">gvametaconvert</a>
<a href="#">nvmsgbroker</a>	<a href="#">gvametapublish</a>

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
nvv4l2decoder	decodebin3
nvv4l2h264dec	vah264dec
nvv4l2h265dec	vah265dec
nvv4l2h264enc	va264enc
nvv4l2h265enc	va265enc



## DL Streamer drives E2E AI Performance on Intel

- Utilizes OpenVINO™ inference optimization
- Efficient handling of memory spaces and minimization of copies
- Inference batching across video streams
- Easy assignments for 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



- 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

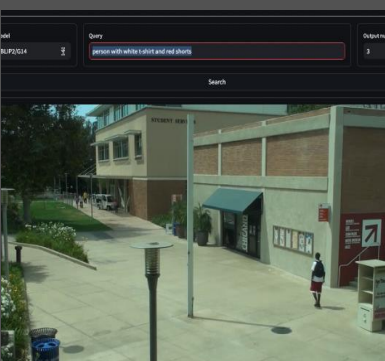
Smart Intersection

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

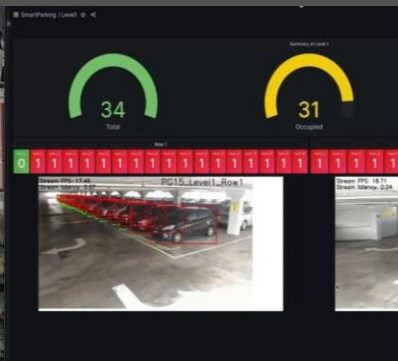
Smart Search (Image Search by Text)



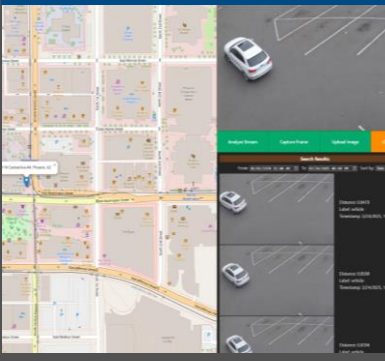
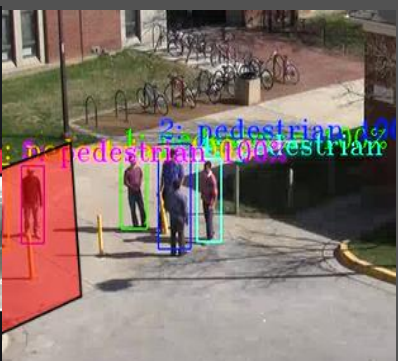
Smart Intersection



Smart Parking



Loitering Detection



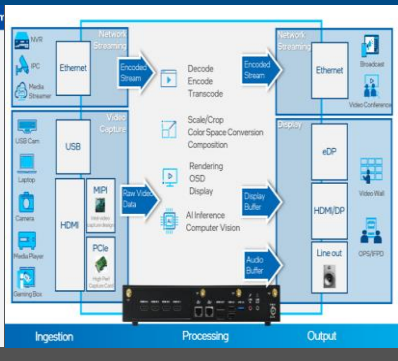
Reidentification (Image Based Video Search)



Video Search and Summarization



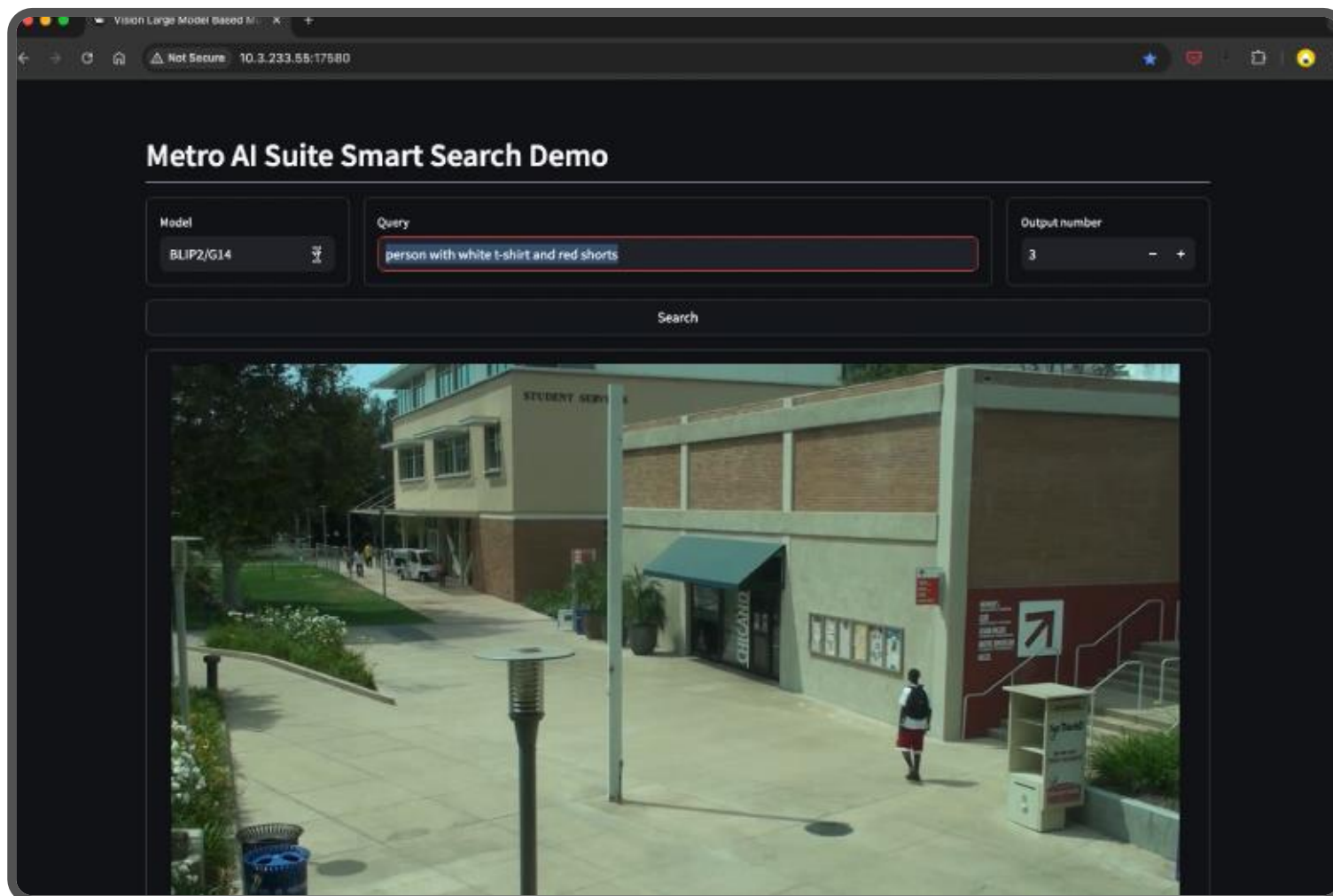
Visual Q&A



Platform Blueprints

## Why Use AI Sample Apps?

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



## Gen AI Based Smart Search

This application uses multi-modal large models to do image retrieval with text query (“Image Search by Text”), enabling Gen AI “Smart Search” functionality in NVR or VMS systems.

### Key Features:

- **Reduce manual search time:** speed up investigations
- **Augment capabilities:** Expand the possibilities of what can be searched for
- **Integrations include multi-modal LLMs CLIP and BLIP2, with results displayed in Web UI**

### Supported Intel Platforms:

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

[Download](#)

# 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 Metro AI Suite DLStreamer) enables composability and reconfiguration

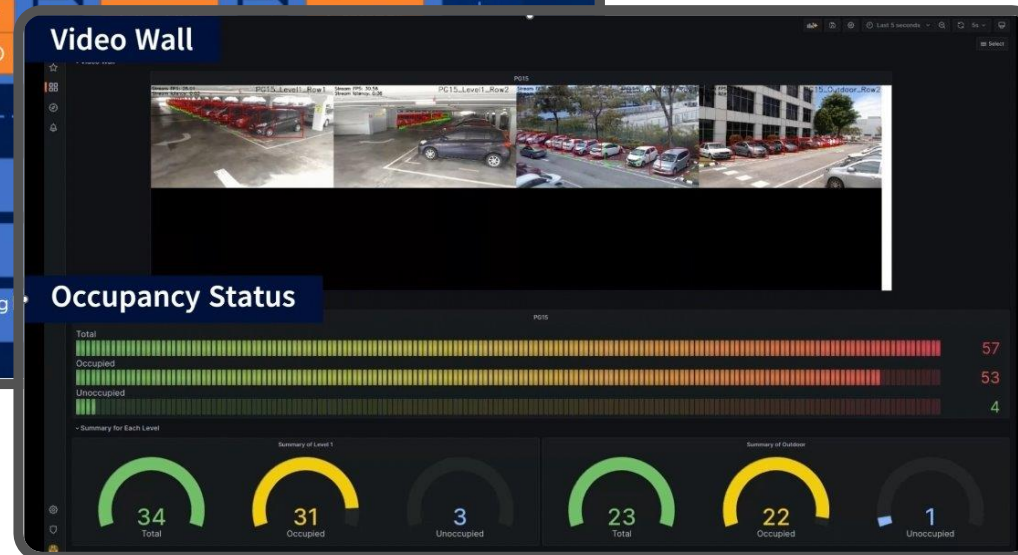
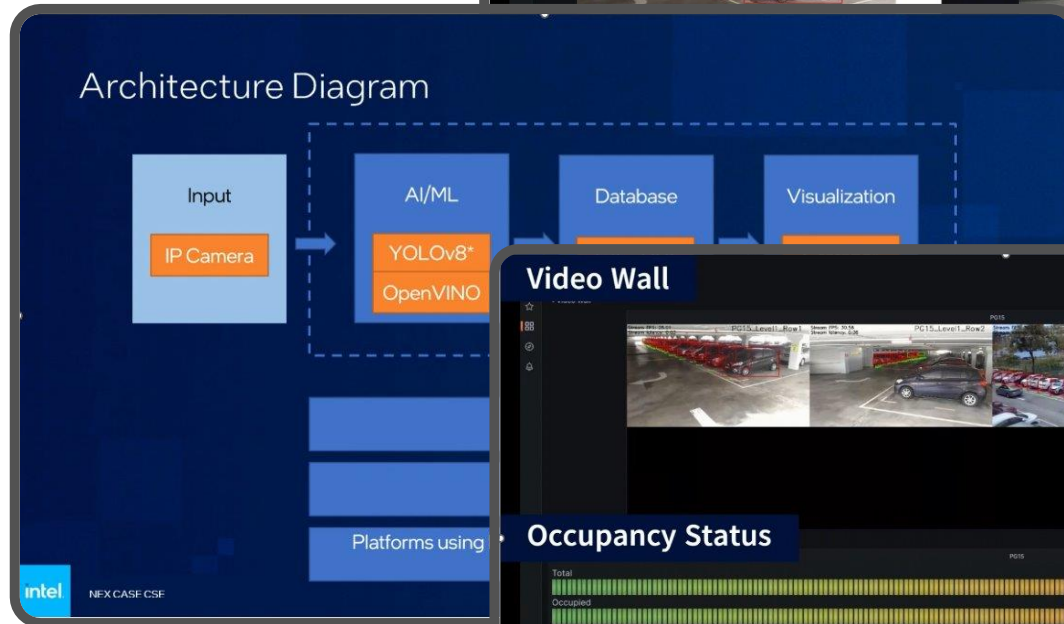
## Supported Intel Platforms:

- Intel® Core™, Intel® Core™ 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 Rapid RI** low-code app framework

## Supported Intel Platforms:

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

[Download](#)

# 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 Rapid RI** low-code app framework

## Supported Intel Platforms:

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

[Download](#)

Last 1 second



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

# Reidentification (Image based Video Search)

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

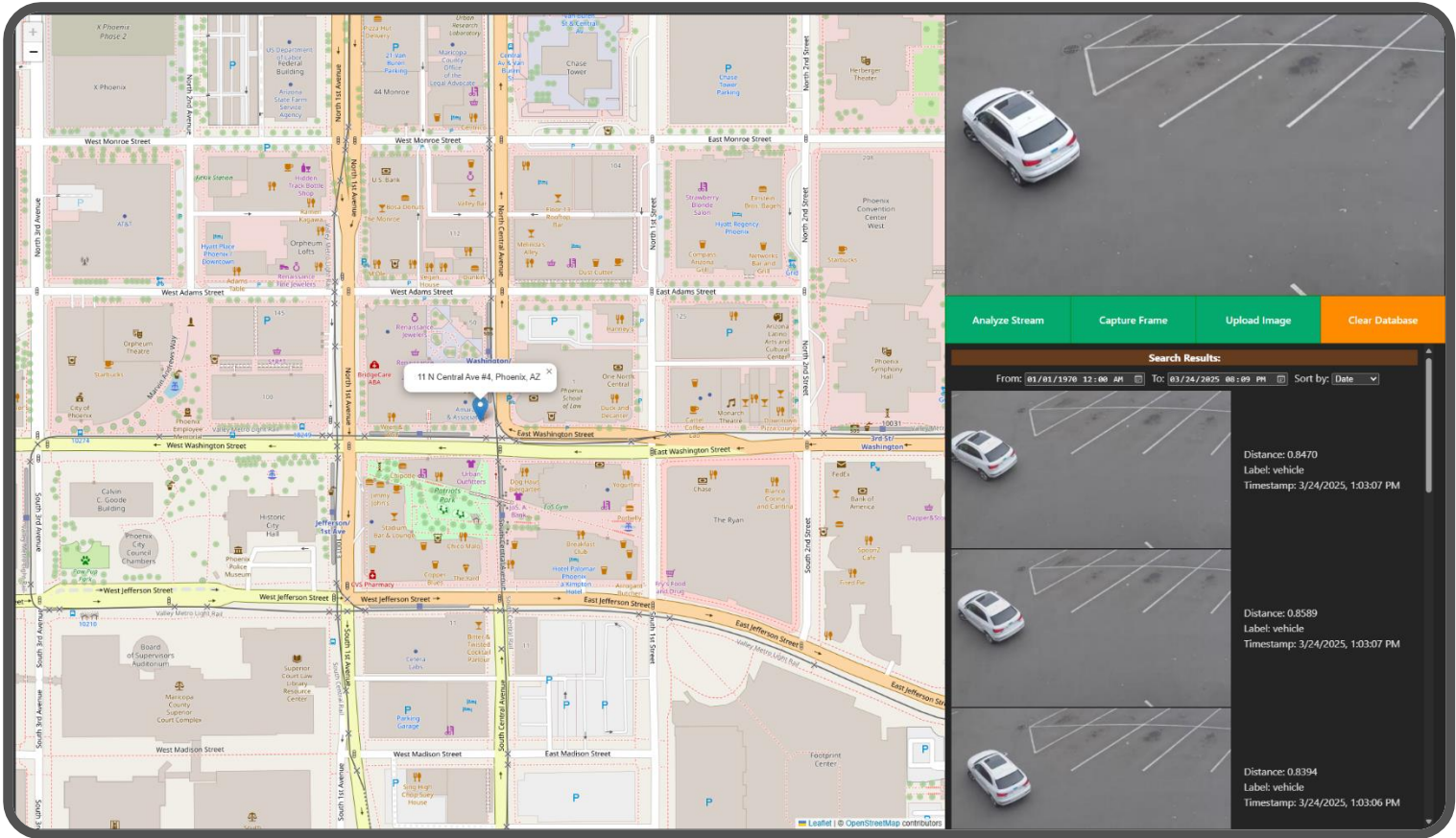
## Key Features:

- **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, ImageIngstor**

## Supported Intel Platforms:

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

[Download](#)





# 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.

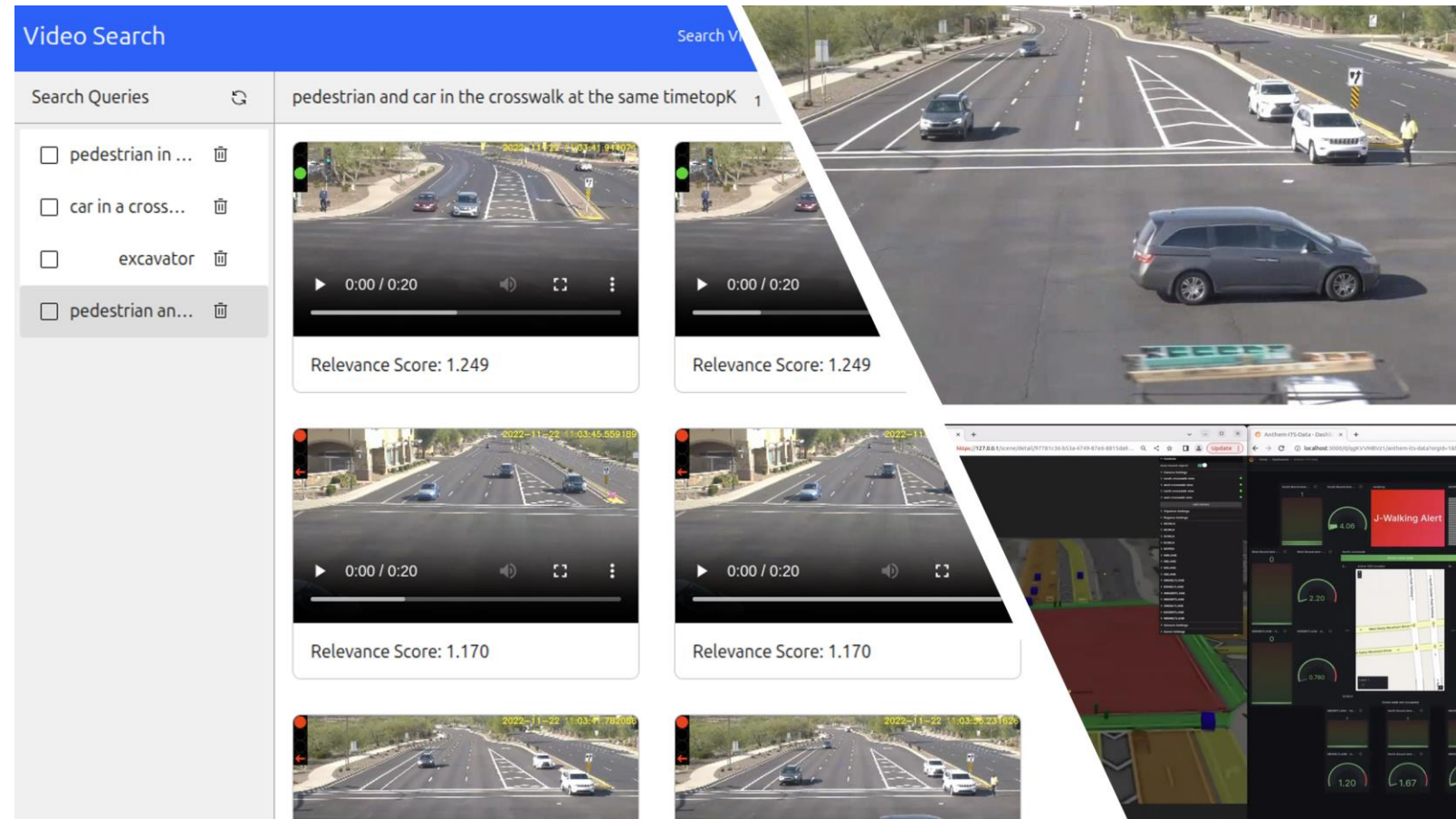
## Key Features:

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

[Download](#)





# Video Search and Summarization (Continued)

A developmental sample application that demonstrates summarization of video streams.

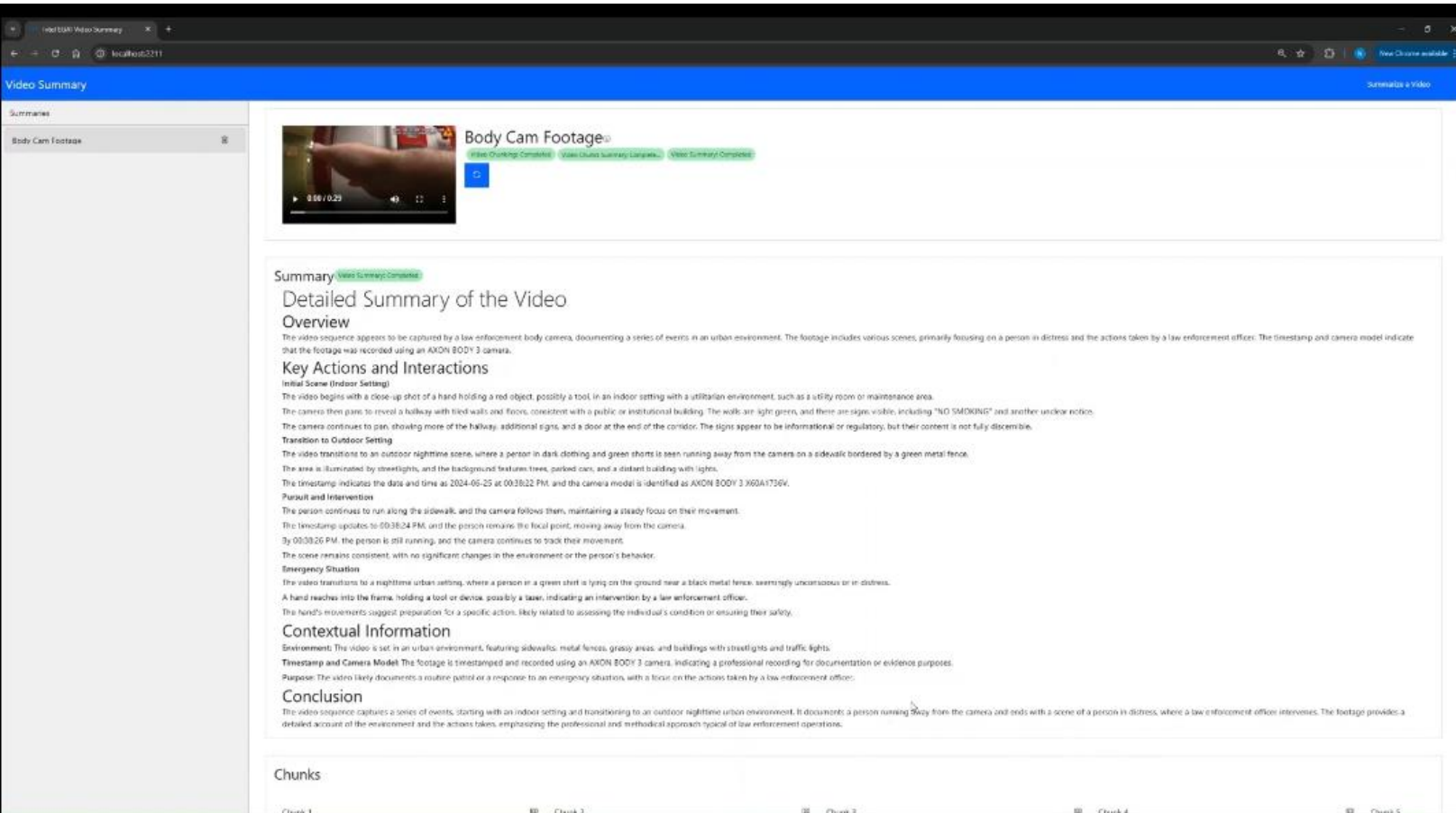
## Key Features:

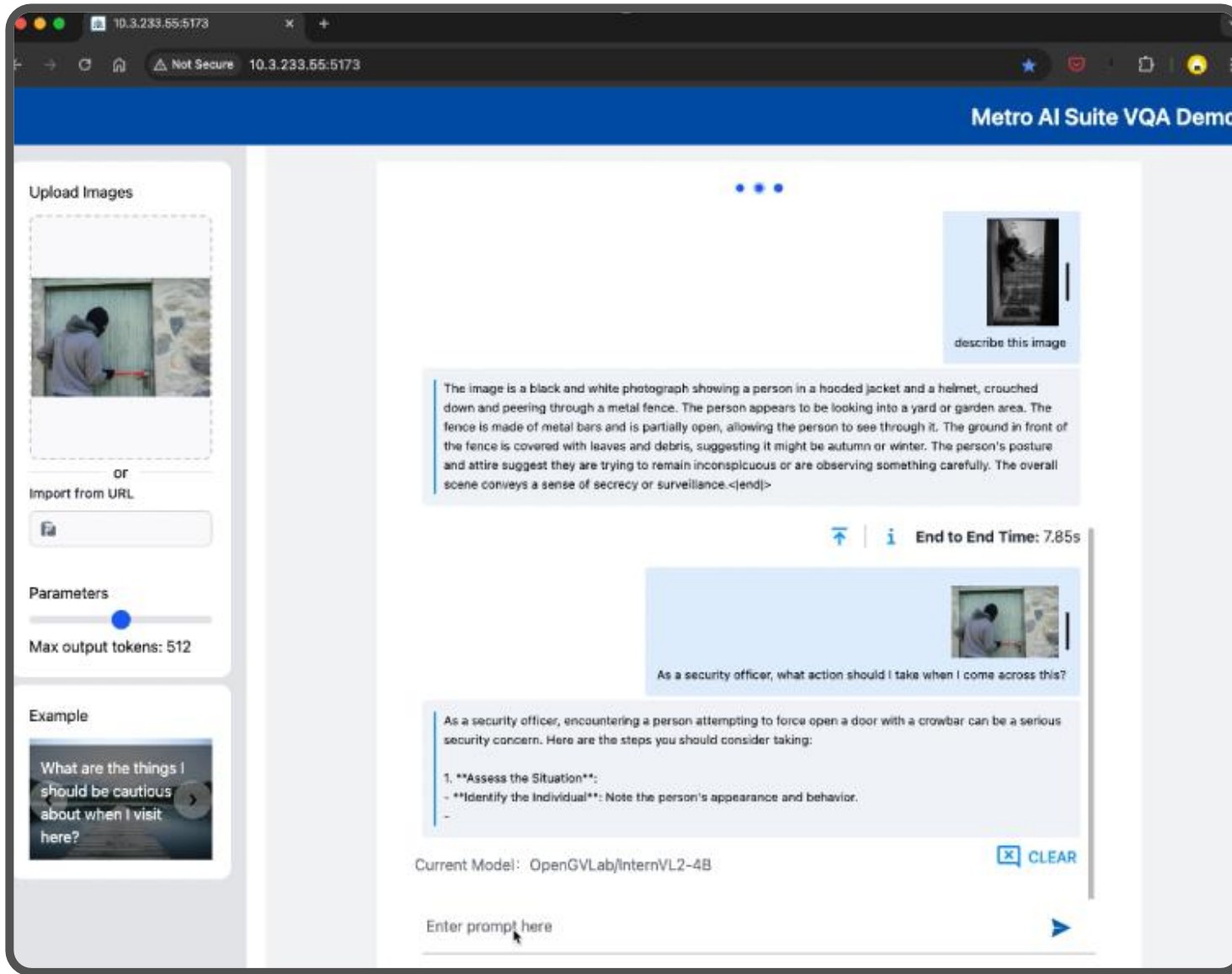
- **Video Summarization:** Using Vision Language Models (VLMs), Computer Vision, and Audio Analysis, the application distills key information into brief synopses from large volumes of data within long-form videos.

## Supported Intel Platforms:

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

Download





# Gen AI Based Video Q&A (VQA)

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**

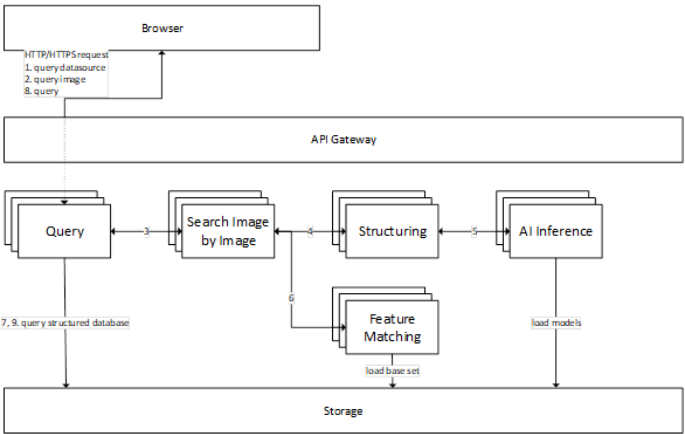
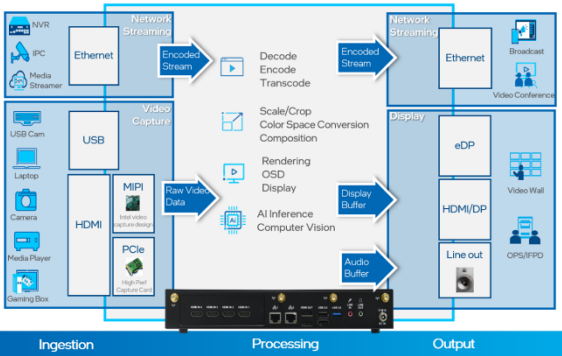
## Supported Intel Platforms:

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

[Download](#)

# Platform Blueprints

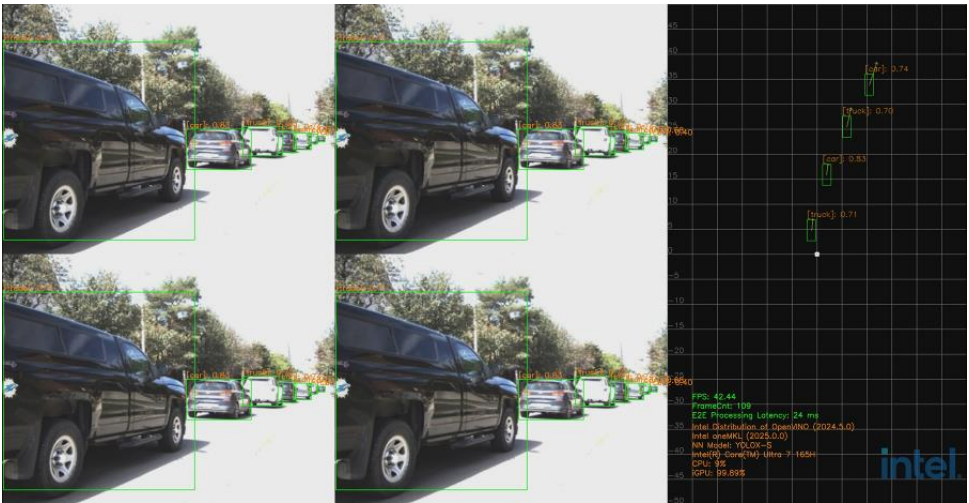
Intel® Video Processing Platform



## Platform Blueprints:

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

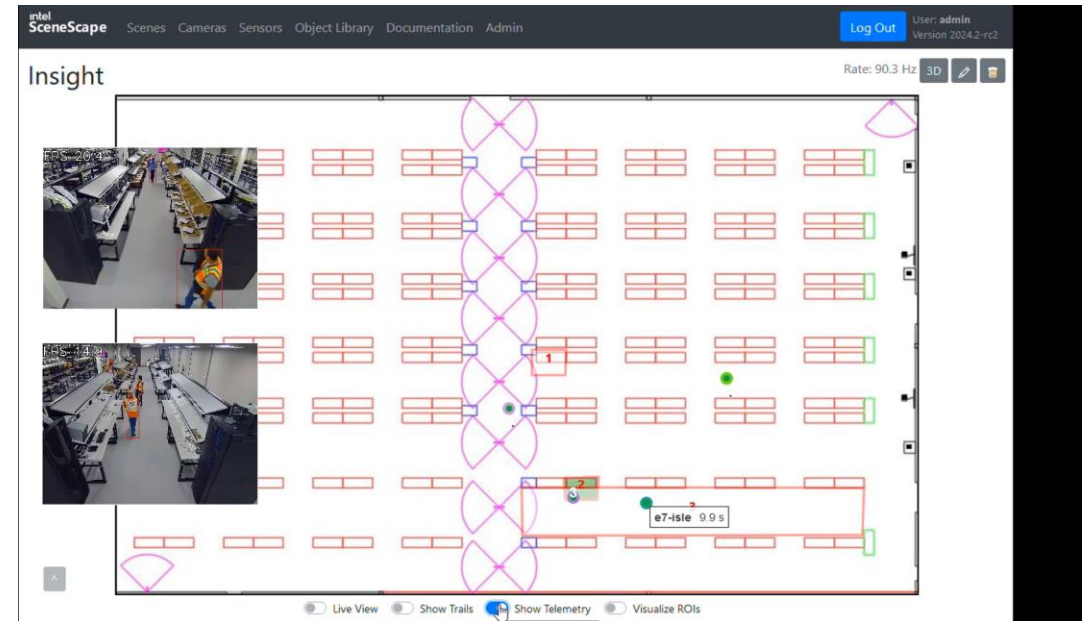
- [Video Processing Platform \(VPP\)](#)
- [Video Analytics Server](#)
- [Sensor Fusion for Traffic Management](#)



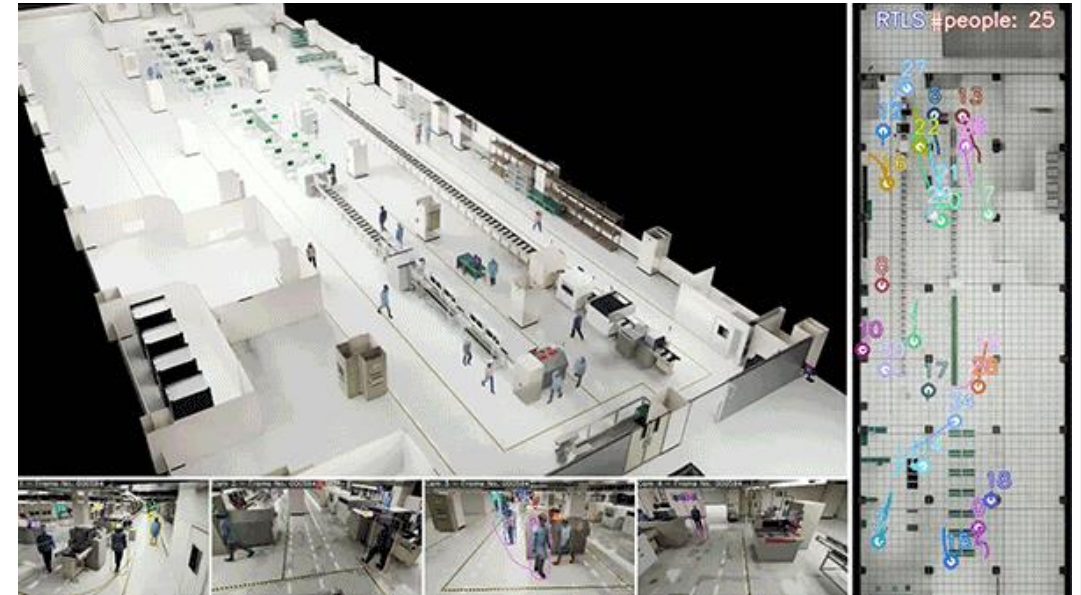
# Additional Tools

# What is Intel® Scenescape?

- Intel® 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 is bringing Physical AI to spaces like intersections, warehouses, stores, and campuses.



The competition:





# 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

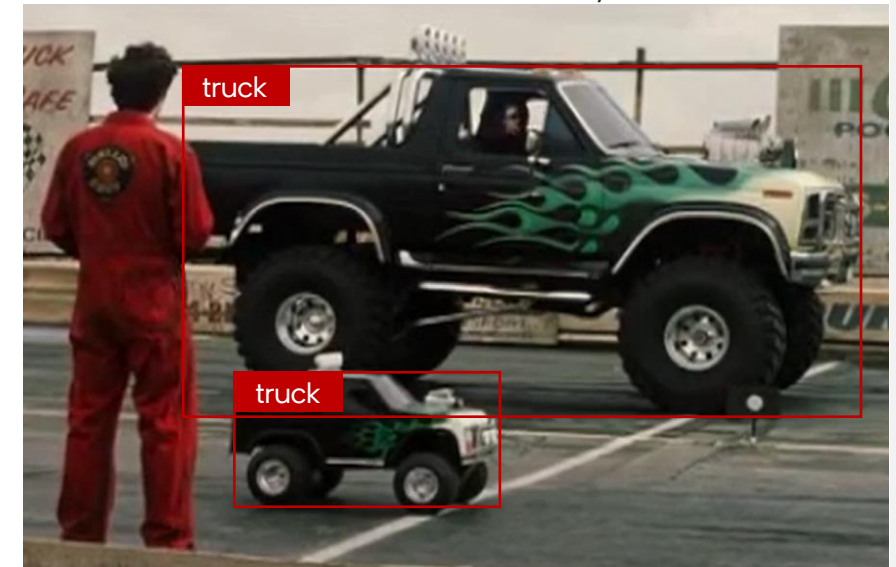


SceneScape solves these real-world problems, even on single cameras, by utilizing world coordinates and camera calibration

Which cars are parked properly?



Which truck is the toy?



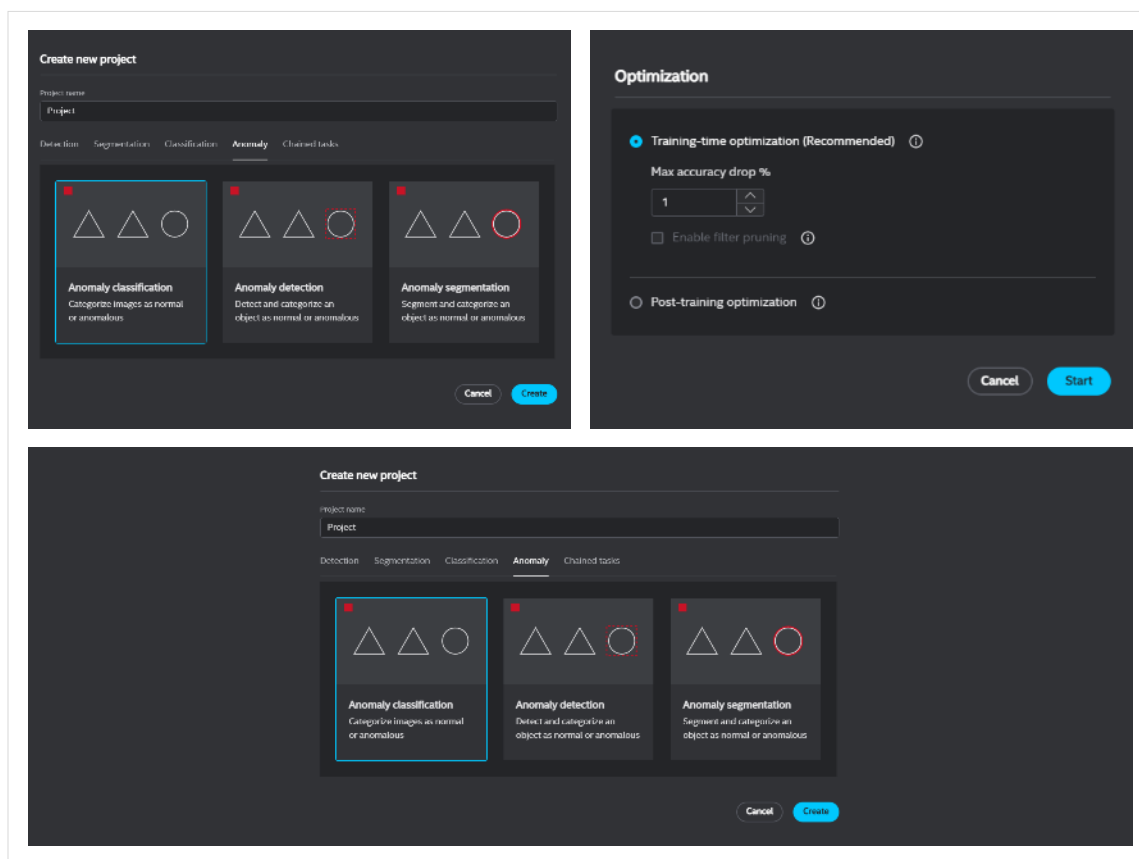
How fast is each car going?



Get Started with Scenescape

# intel<sup>®</sup> Geti<sup>™</sup> : Powerful AI for Everyone

From data input to optimization and model export, the Intel<sup>®</sup> Geti<sup>™</sup> software platform enables teams to create vision AI models more efficiently – [geti.intel.com](https://geti.intel.com)



## Develop vision models with the Intel<sup>®</sup> Geti<sup>™</sup> platform:

- **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<sup>™</sup>:

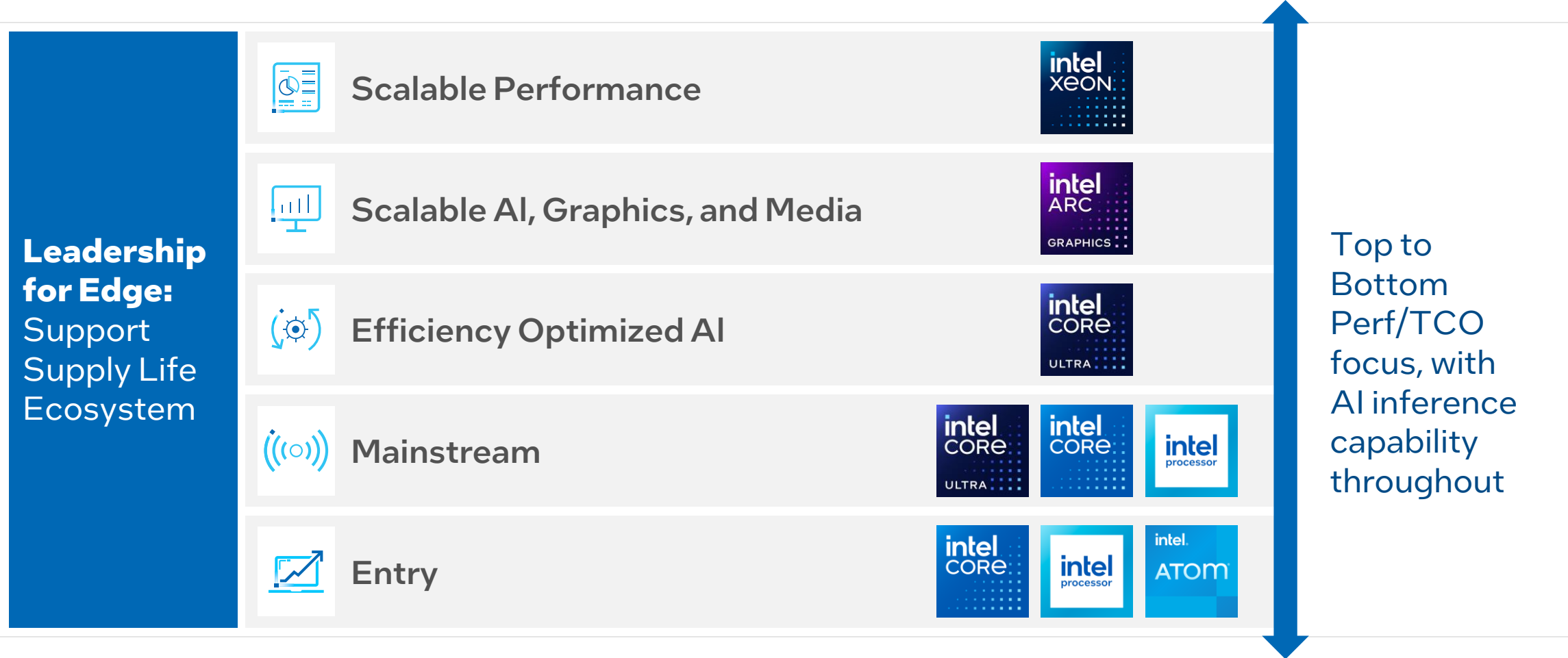
- **Built-in OpenVINO<sup>™</sup> optimizations** – Easily optimize and quantize trained models with inferencing software that automatically detects available compute across Intel CPUs and GPUs

# Recommended Systems



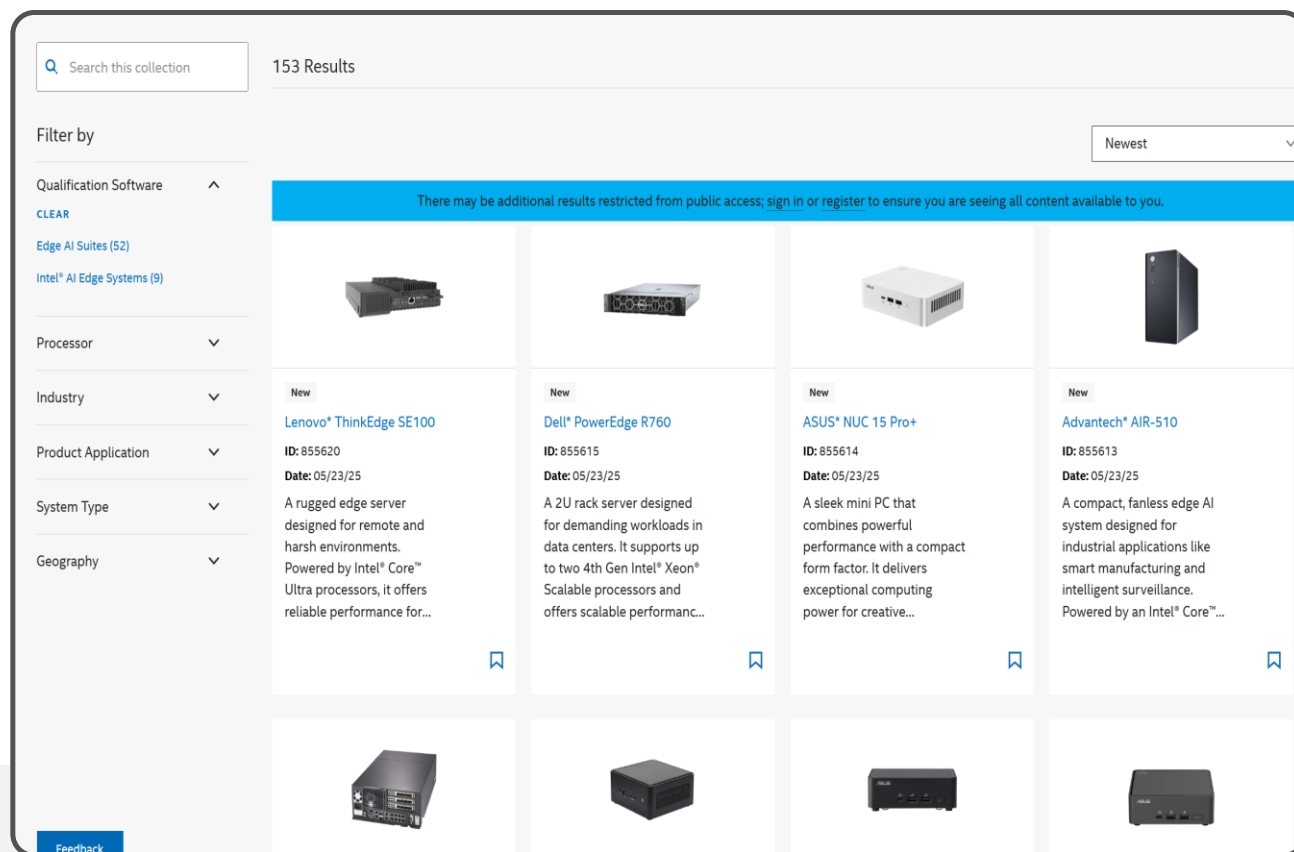
# Intel Edge Product Portfolio

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**



# Find Qualified Systems in the Recommended Hardware Catalog

## Metro AI Suite Recommended Hardware Catalog



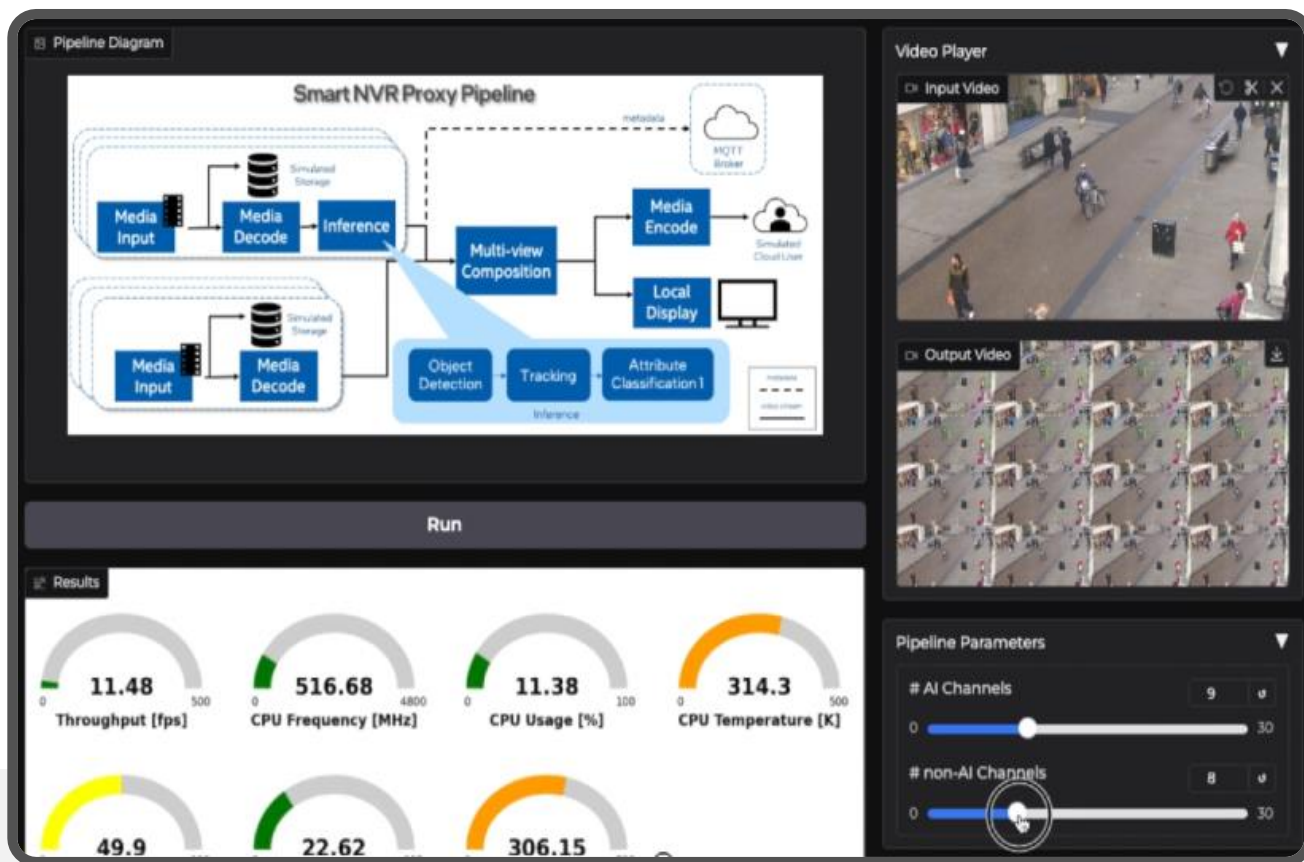
### Catalog Benefits for SW Partners:

- 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



### About ViPPET

- Evaluate AI system performance for Visual AI workloads (# of streams, etc.)
- Answer 'What-if' permutations for model, pipeline, & system performance tradeoffs
- End to End (Video + AI) workload is much closer to real world than TOPS or Inference-only

### Recommended Usage











- Quickly evaluate system performance using proxy models
  - Run on each system to get detailed E2E performance for video & most common AI models
  - Understand system capacity for Visual AI workloads

# Partner Support Programs

# CASE Enabling & Support Model

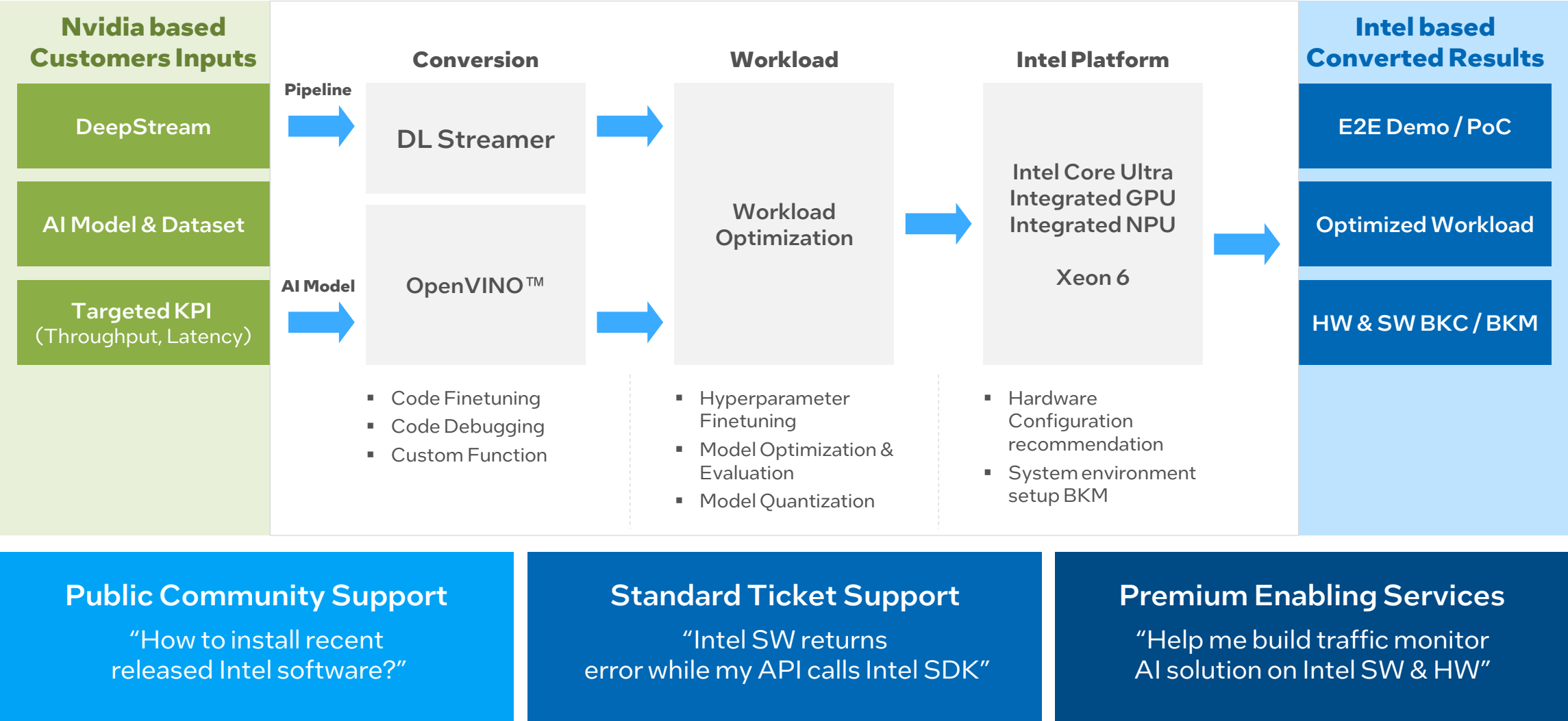
Enabling & Support Level	Intel Support Scope	Support Tools & Channels	Support Examples
<b>Intel® Premium Enabling Services</b>	<ul style="list-style-type: none"> <li>▪ Per SLA/ TSoW – Custom, Customer Specific</li> <li>▪ Custom POC/ Demo/ RI</li> <li>▪ Onsite Support</li> <li>▪ Priority on Early Access</li> <li>▪ Pre-launch Preview</li> </ul>	Intel Premier Support Portal (IPS) & Email & Phone & Meetings & As defined in TSOW	“Help me build traffic monitor AI solution on Intel SW & HW”
<b>Standard Ticket Support</b>	<ul style="list-style-type: none"> <li>▪ Dedicated AE, NDA collateral</li> <li>▪ Ticket based case</li> <li>▪ Reference Implementation</li> <li>▪ Training/ Workshop, Remote Access</li> <li>▪ Time bound, SLA based on ticket priority</li> </ul>	Intel Premier Support Portal (IPS) & Email	“Intel SW returns error while my SW is calling Intel stack”
<b>Public Community Support</b>	<ul style="list-style-type: none"> <li>▪ Post Launch Support</li> <li>▪ Community support</li> <li>▪ Self Enabling Tool</li> <li>▪ Public &amp; limited NDA Collateral</li> <li>▪ How-to Video</li> <li>▪ Webinar</li> </ul>	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
 <b>Object Detection</b>  <b>Segmentation</b>	<b>Convert to Intel</b>	<ul style="list-style-type: none"> <li>▪ CUDA source codes</li> <li>▪ AI model</li> <li>▪ KPI (Throughput, Latency)</li> </ul>	<ul style="list-style-type: none"> <li>▪ CUDA &gt; SYCL migration</li> <li>▪ Throughput &amp; Latency Optimization</li> <li>▪ Platform recommendation</li> <li>▪ Evaluation on Intel Hardware</li> </ul>	<ul style="list-style-type: none"> <li>▪ Demo/POC</li> <li>▪ Optimized workload</li> <li>▪ HW/SW BKC &amp; BKM</li> </ul>
 <b>People/Face Detection</b>  <b>Image Classification</b>	<b>LLM/ ML on Intel</b>	<ul style="list-style-type: none"> <li>▪ LLM Model Requirement</li> <li>▪ Current/Target Platform</li> <li>▪ KPI (Throughput, Latency)</li> </ul>	<ul style="list-style-type: none"> <li>▪ LLM Model Recommendation</li> <li>▪ Model Quantization</li> <li>▪ Workload Optimization</li> <li>▪ Platform recommendation</li> <li>▪ Sample application</li> </ul>	<ul style="list-style-type: none"> <li>▪ Demo/POC</li> <li>▪ Optimized workload</li> <li>▪ HW/SW BKC &amp; BKM</li> </ul>
 <b>Gen AI/LLM</b>  <b>Big Data Analytics</b>	<b>Securing AI Model &amp; Dataset</b>	<ul style="list-style-type: none"> <li>▪ KPI (Throughput, Latency)</li> <li>▪ Use Case</li> <li>▪ AI Model &amp; Sample Dataset</li> <li>▪ Security Requirement</li> </ul>	<ul style="list-style-type: none"> <li>▪ End-to-end encrypted training &amp; inferencing</li> <li>▪ Protecting the NN model in Secure Enclave</li> <li>▪ Model copyright protection through Watermark/fingerprint technology</li> </ul>	<ul style="list-style-type: none"> <li>▪ Demo/POC</li> <li>▪ Encrypted Model &amp; Dataset</li> <li>▪ HW/SW BKC &amp; BKM</li> </ul>
 <b>Optical Character Recognition</b>  <b>Secured AI</b>	<b>Model Training &amp; Inference Optimization</b>	<ul style="list-style-type: none"> <li>▪ KPI (Throughput, Latency)</li> <li>▪ Use Case</li> <li>▪ Sample Dataset</li> <li>▪ Trained Model</li> </ul>	<ul style="list-style-type: none"> <li>▪ Model Optimization &amp; Quantization</li> <li>▪ Hyper Parameter Tuning</li> <li>▪ Performance Optimization &amp; Demonstration</li> <li>▪ Platform recommendation</li> <li>▪ Reference Configuration</li> </ul>	<ul style="list-style-type: none"> <li>▪ Demo/POC</li> <li>▪ Optimized workload</li> <li>▪ HW/SW BKC &amp; BKM</li> </ul>
 <b>Speech to Text Recognition</b>  <b>Sound Classification</b>				

# Conversion Journey and Intel Support\*

\*Details varies based on project needs



# 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

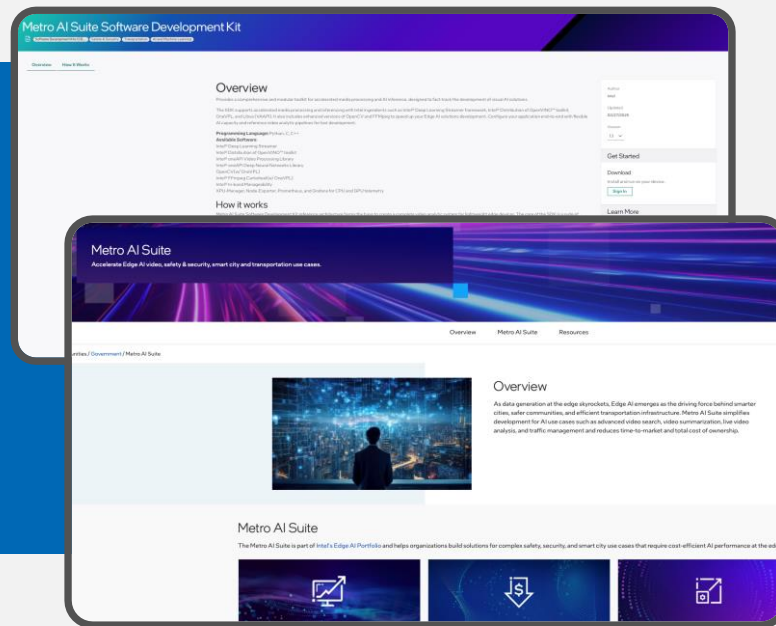
- ▶ **Gain performance improvements** via SDK libraries
- ▶ **Refer to resources**, such as optimized code samples, documentation, & training

# 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](https://intel.com/metroai)
- [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