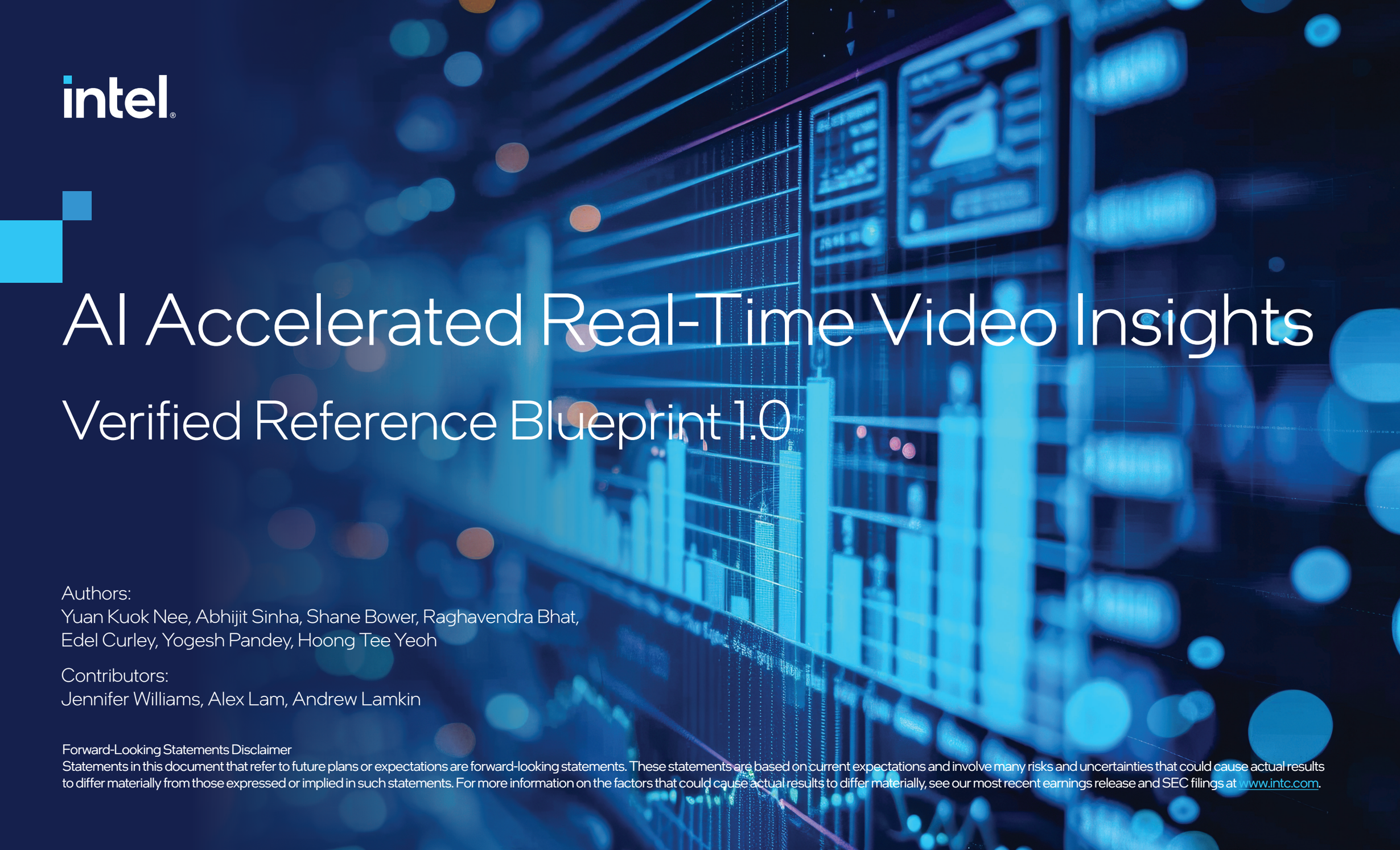


AI Accelerated Real-Time Video Insights

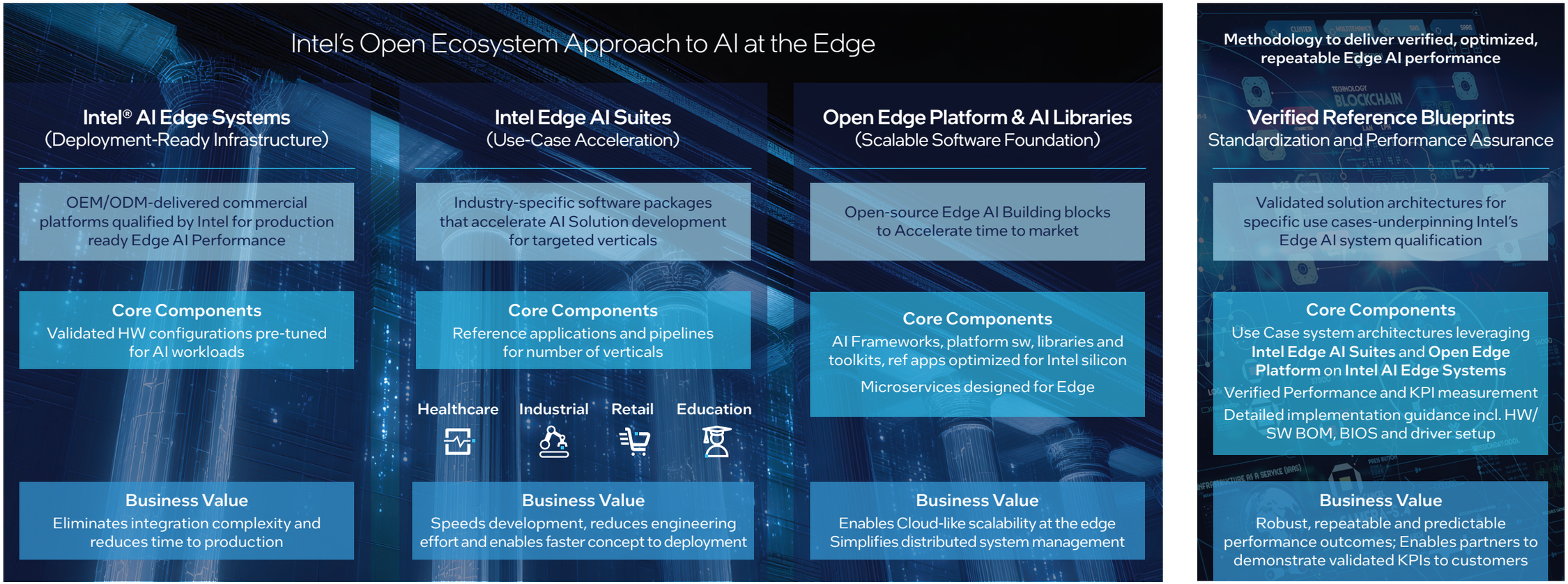
Verified Reference Blueprint 1.0

Authors:
Yuan Kuok Nee, Abhijit Sinha, Shane Bower, Raghavendra Bhat,
Edel Curley, Yogesh Pandey, Hoong Tee Yeoh

Contributors:
Jennifer Williams, Alex Lam, Andrew Lamkin

Forward-Looking Statements Disclaimer
Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.

Intel delivers a silicon roadmap with open, standardized software and validated system solutions to develop Edge AI Solutions. This integrated approach enables customers and partners to **accelerate time-to-market, reduce deployment risk, and scale AI efficiently across distributed edge environments**. At the core of this strategy are modular platform initiatives— including the **Open Edge Platform, Intel® Edge AI Suites, and Intel® AI Edge Systems** — designed to simplify the full lifecycle of Edge AI, from development to deployment and ongoing operations



Qualified systems (Intel® AI Edge Systems).
Optimized software (Open Edge Platform and Edge AI Suites).
Validated architectures (Verified Reference Blueprints).
A broad partner ecosystem (OEMs, ISVs, integrators).

Intel brings together

Result: A cohesive platform that enables organizations to move from pilot to production faster, with:

- Reduced integration complexity.
- Predictable performance and benchmarking.
- Scalable deployment across locations and use cases.
- Lower total cost of ownership over the lifecycle.

Table of Contents

1. Executive Summary

This Verified Reference Blueprint (VRB), "AI Accelerated Real-Time Video Insights," provides a comprehensive framework for deploying advanced scene understanding on high-efficiency edge platforms. By leveraging the Intel® Core™ Ultra Series 3 featuring integrated NPUs and Intel® Arc™ GPUs, this blueprint demonstrates that sophisticated vision-language model (VLM) solutions no longer require the high total cost of ownership (TCO) associated with workstation-class servers and discrete GPUs. This solution optimizes the balance between performance and power, enabling real-time, context-aware captioning at a fraction of the hardware footprint and energy cost of traditional architectures.

To bridge the gap between complex AI research and commercial deployment, Intel offers hardened and verified Intel® AI Edge Systems complemented by Edge AI Suites. These optimized foundations allow independent software vendors (ISVs) and system integrators (SIs/GSIs) to significantly reduce development costs and increase confidence in system reliability. By utilizing these pre-qualified systems, partners can jumpstart the integration of AI functionality into business applications, ensuring predictable performance and a faster path to revenue.

This verified reference blueprint is designed to empower decision-makers and technical architects across key industries where low latency and local processing are critical for safety, privacy, and operational efficiency. By targeting these specific leadership and engineering roles, the blueprint enables cross-functional teams to modernize infrastructure with real-time VLM capabilities while maintaining the reduced TCO of an integrated Intel® Core™ Ultra Series 3 platform.

Revision Number	Description	Revision Date
1.0	Initial release	August 2026

2. Overview

This Verified Reference blueprint demonstrates how Intel® technology enables sophisticated real-time video analytics of Real-Time Streaming Protocol (RTSP) feeds across diverse edge environments. By deploying innovative VLMs, the solution can provide automated scene understanding for **healthcare** (fall prevention and behavioral analysis), **education** (engagement focus), **smart cities** (traffic monitoring), or **manufacturing** (worker safety). These capabilities transform raw video into actionable intelligence, allowing organizations to monitor complex environments with greater accuracy and less manual oversight.

While adaptable across verticals, this blueprint focuses on a **manufacturing** worker-safety scenario to exemplify its technical framework. The solution integrates **YOLO-based object detection** (You Only Look Once) for high-speed event classification with **VLM-generated captioning** to provide rich, descriptive context that minimizes false positives. Furthermore, it incorporates **conversational video search via RAG** (Retrieval-Augmented Generation), utilizing the integrated NPU to power a chat interface. This allows operators to perform "ask the video" queries to instantly retrieve relevant keyframes and timestamps across multiple camera streams for rapid forensic review.

By consolidating advanced AI workloads onto the **Intel® Core™ Ultra Series 3** platform at the edge, this verified reference blueprint eliminates the latency, cost, and privacy concerns associated with server grade platforms in cloud-hosted models. The software architecture intelligently distributes processing across the **CPU, Xe3 GPU, and NPU**, delivering high-performance multi-modal inferencing on a single, cost-effective SoC. This approach provides a viable path for partners to recommend **commercial off-the-shelf (COTS)** systems that achieve server-class performance with a significantly lower TCO and a reduced physical footprint.

The VRB will describe the performance of VLMs on an optimized commercial AI system delivered through an OEM. The performance data was collected using a VLM only pipeline running on the GPU. In order to stress the iGPU to maximum capacity, the Yolo model was not used in performance data collection.

3. Use Cases

The underlying video analytics pipeline featured in this verified reference blueprint is highly versatile, supporting a wide array of use cases across key edge verticals. By providing automated scene understanding, the solution enables critical applications such as **fall prevention and behavioral analysis** in healthcare, **classroom engagement focus** for education, and **real-time traffic monitoring** in smart cities. In the **manufacturing** sector example that this blueprint focuses on, the technology is particularly impactful for enhancing worker safety through automated fall detection and emergency response possibly leading to future fall prevention. While the examples discussed here represent some of the most prominent applications, they are only a subset of the potential deployments made possible by this technology solution.

3.1 Education Vertical: Automated Lecture Summarization and Analytics

Beyond industrial safety, the multi-modal pipeline can be deployed within educational environments — such as lecture halls and classrooms — to automate the aggregation of localized student engagement analytics. Rather than processing ambient classroom audio or conversational dialogue, the system utilizes raw visual data streams to map physical behavioral indicators of active learning and attentiveness.

Depending on the custom prompts provided to the lightweight VLM, the pipeline evaluates environmental dynamics across the classroom layout, generating objective summaries of student attentiveness and tracking real-time pose estimation data. By analyzing behavioral postures, the system automatically counts explicit physical interactions, such as hands raised, head tilt, or students standing up.

By consolidating these visual and localized inference workloads onto a single, power-efficient **Intel® Core™ Ultra Series 3 platform**, educational institutions can access sophisticated classroom engagement analytics locally, maintaining strict student privacy boundaries while eliminating cloud operational costs. Furthermore, while this blueprint focuses strictly on the visual engagement proposition, the broader Intel® Edge AI Suites provide additional, modular Edge AI Libraries and media processing tools — such as advanced speech processing microservices — that can be seamlessly integrated to evolve these core concepts into a comprehensive, multi-sensory classroom analytics solution.

3.2 Smart Cities Vertical: Real-Time traffic Monitoring and Incident Response

In the smart cities vertical, the multi-modal pipeline can be deployed across urban infrastructure to transform traditional traffic surveillance into an intelligent, context-aware monitoring network. While legacy computer vision systems are limited to basic object labeling (e.g., tagging a "car" or "pedestrian"), the fusion of **YOLO** and **a light VLM** architecture delivers comprehensive scene understanding. The object detection block acts as a powerful filter of frames of interest which is fed to a VLM for a detailed analysis. This combination helps address both the "what" and "why" of scene understanding. The pipeline helps describe in natural language complex behaviors, provide richer contexts, and flag potential risks — such as identifying when a child has become separated from a group near a busy transit hub. By converting vast streams of raw video data into structured text descriptions at the edge, municipal traffic managers can leverage large language models (LLMs) to query historical footage using natural language. This drastically flattens the time required for forensic data analysis and makes large-scale urban video data highly searchable and accessible.

Beyond passive observation, the real-time nature of this edge pipeline enables raising alerts through active anomaly detection and dynamic traffic management. The system can immediately identify and describe complex road incidents, such as multi-vehicle accidents, lane blockages, or a motionless person in a high-traffic area, automatically generating descriptive alerts like, *"There is a motionless person lying near exit 3; flagged as a potential medical emergency."* Furthermore, this blueprint aligns with emerging intelligent infrastructure research, such as safety-prioritized meta-controllers like the "Light VLM" framework. By utilizing the VLM to interpret complex, dynamic intersections, the system can assist in adaptive traffic signal control, optimizing urban mobility and safely prioritizing emergency vehicle preemption. Consolidating these dense visual and linguistic workloads onto an energy-efficient **Intel Core Ultra** Series 3 platform ensures that cities can achieve localized, low-latency intelligence without the prohibitive infrastructure costs of discrete GPU servers.

3.3 Healthcare Vertical: Patient Activity Monitoring and Behavioral Analysis

Within healthcare environments, a multimodal edge AI pipeline may provide a foundation for exploring patient activity monitoring and behavioral analysis use cases while preserving privacy through local processing. Combining computer vision models such as YOLO with vision-language models (VLMs), illustrative applications could include identifying and describing observed movement patterns, tracking changes in posture over time, or generating natural-language summaries of activities captured by cameras or other sensors. Potential scenarios might include monitoring patient mobility in care settings, supporting observational workflows in rehabilitation environments, or assisting caregivers with contextual summaries of routine activities.

More advanced exploratory use cases could leverage VLMs to compare observed activity patterns against established baselines and highlight notable deviations for human review. For example, the system might surface changes in movement frequency, mobility, room occupancy patterns, or other observable behaviors that could warrant additional attention. Similarly, organizations could investigate the use of VLM-generated summaries to support hygiene, safety, or operational compliance workflows by documenting observed activities and events. These examples are intended as illustrative applications of vision-language AI and would require appropriate validation, oversight, and integration into clinical processes before being used in healthcare decision-making.

Because the pipeline executes locally at the edge on Intel® Core™ Ultra Series 3 platforms, organizations can evaluate these use cases without requiring continuous transmission of visual data to cloud services. Edge-based processing may help address privacy, latency, bandwidth, and data-governance considerations while enabling near real-time analysis and summarization of visual inputs. As a result, healthcare providers, researchers, and solution developers can explore a range of patient-monitoring and operational-assistance scenarios using multimodal AI architectures while retaining flexibility regarding how insights are reviewed, validated, and acted upon within their specific environments.

3.4 Manufacturing Vertical: Industrial Worker Safety with Automated Fall Detection and Emergency Response

In the manufacturing vertical, this verified reference blueprint provides a hardened, real-world framework for automated industrial safety monitoring and rapid emergency alerting. Leveraging the **Metro AI Suite**, the solution ingestion engine captures live RTSP video feeds from the warehouse floor to actively enforce safety protocols. Using optimized media creation tools and deep-learning streamer analytics, the architecture processes high-throughput streams locally to minimize edge latency. The primary objective is to accurately capture, identify, and document high-risk safety incidents — such as a worker falling in a warehouse aisle — and instantly broadcast context-rich alerts and respond to queries based on natural language processing (NLP). By establishing a verified reference blueprint using lightweight VLMs, this technology solution proves that robust warehouse surveillance can be fully automated with fewer false positives, ensuring lightning-fast operational decisions where seconds save lives.

To achieve this without the cost and footprint of workstation-grade hardware, the edge workload is consolidated onto a single **Intel Core Ultra Series 3** platform. Frames of interest would be captioned and stored in an Edge database. These frames could be retrieved at a later time for analysis. The multi-modal edge inference pipeline is executed chronologically, as follows :

- **The Entry Stage:** High-speed **YOLO-based object detection and pose estimation** run seamlessly on the integrated **Intel® Arc™ GPU (can be configured for NPU too)**, instantly flagging frames of interest based on configured criteria. These frames are marked for further processing by the VLM.
- **The Contextual Stage:** As the workload moves along the inference pipeline, these same **Xe3 GPU** cores process the downstream execution phase utilizing a lite VLM. This generates highly descriptive, low-latency textual captions of the incident, delivering the critical environmental context that standard vision models miss.
- **The Forensic Stage:** An **NPU-accelerated conversational video search via RAG** takes a prompt and converts it to a "search cue." Security personnel can use the chat interface to make "ask the video" queries, instantly retrieving precise timestamps and keyframes across one or more camera streams.

To establish verifiable key performance indicators (KPIs), the system is benchmarked using reference software from the Edge AI Suites to provide a definitive baseline for edge VLM performance. The primary KPI demonstration of the platform's efficiency is *Captions per Minute* measured over an RTSP stream, alongside precise tracking of time-to-first-token (TTFT) and generation throughput (tokens/second). This data-backed architecture proves that partners can deploy COTS Edge AI systems capable of distributing dense multi-modal workloads intelligently, eliminating the need for expensive discrete GPU server deployments.

4. System Architecture

4.1 Architecture Overview

Searching video footage for critical events or incidents is time consuming. Storing video footage is expensive. This verified reference blueprint proposes an Edge architecture that leverages video summarization software in the Metro AI Suite to caption images, generate search vectors, and trigger alerts on edge hardware. Using VLMs, captions are generated and stored for retrieval. An object detection model determines frames of interest to be provided to the VLM for captioning. The captions are stored in an Edge database and can be retrieved for forensic analysis to determine important segments in the video. The "Raise Alerts" feature can output an alert based on the prompt to a UI.

Traditional models only label objects (e.g., "car," "person"). VLMs go further by describing the context, behavior, and potential risks in natural language, e.g., "One person is lying on the ground, seemingly unconscious or injured." This provides richer information to human operators. This architecture implements a light VLM approach where only frames of interest are passed to the VLM to be captioned. Three different VLMs are profiled in this VRB.

The proposed architecture would be divided into two functional components:

- Captioning and Alert Generation Pipeline implemented at an Edge location
- Centralized Control Center implemented at a location so that it could query a number of Edge locations

A diagram of the architecture is shown in Figure 1.

(Note: There are two functions in the architecture that use prompts. One prompt is used to raise an alert, and another would be used in the Centralized Control Center to query the edge database.)

Architecture Components

Captioning and Alert Generation Pipeline: a pipeline that operates in almost real time. Using YOLOv8, if an object or event of interest is detected, the frames of interest (still images) are sent to a VLM for captioning. The caption is converted to an embedding (a search vector) using the Qwen3 Embedding 0.6B model and is stored for retrieval. Alerts could be generated for instant action depending on the prompt used. The YOLOv8 model and the model used to create the embedding could run on the NPU; but for this version of the VRB, they run on the integrated Intel® Arc™ GPU (iGPU).

- Object detection using YOLOv8
  - Light VLM captioning:
  - OpenGVLab/InternVL2-1B
  - OpenGVLab/InternVL2-2B
  - Google/Gemma-3-4B-IT
- Caption conversion to embeddings service that converts caption in clear text into numerical vectors that capture their semantic meaning.
  - Qwen3-Embedding-0.6B

Edge vector database to store embeddings (search vectors) and frames of interest (still images) that were captioned. (The VDMS project at https://github.com/IntelLabs/vdms handles both vector and visual data storage using a PMGD persistent memory graph database).

Centralized Control Center: takes a query input "Is there a person lying on the ground?" and converts it to a search vector which is used to query the edge vector database. The query is converted to an embedding (search vector) and is sent to VDMS for context retrieval. Still images and captions that are similar to the query are returned from VDMS. The LLM caption summarization would be invoked by the RAG orchestrator and would summarize the results returned from VDMS. The LLM summarization could be offloaded to NPU.

- Caption conversion to embeddings – (same function as above)
- RAG Orchestrator — takes a search vector, retrieves similar vectors with images and captions from VDMS, summarizes results using an LLM, and returns a response enriched with relevant still images.
- LLM Caption summarization
  - Microsoft/Phi-3.5-mini-instruct

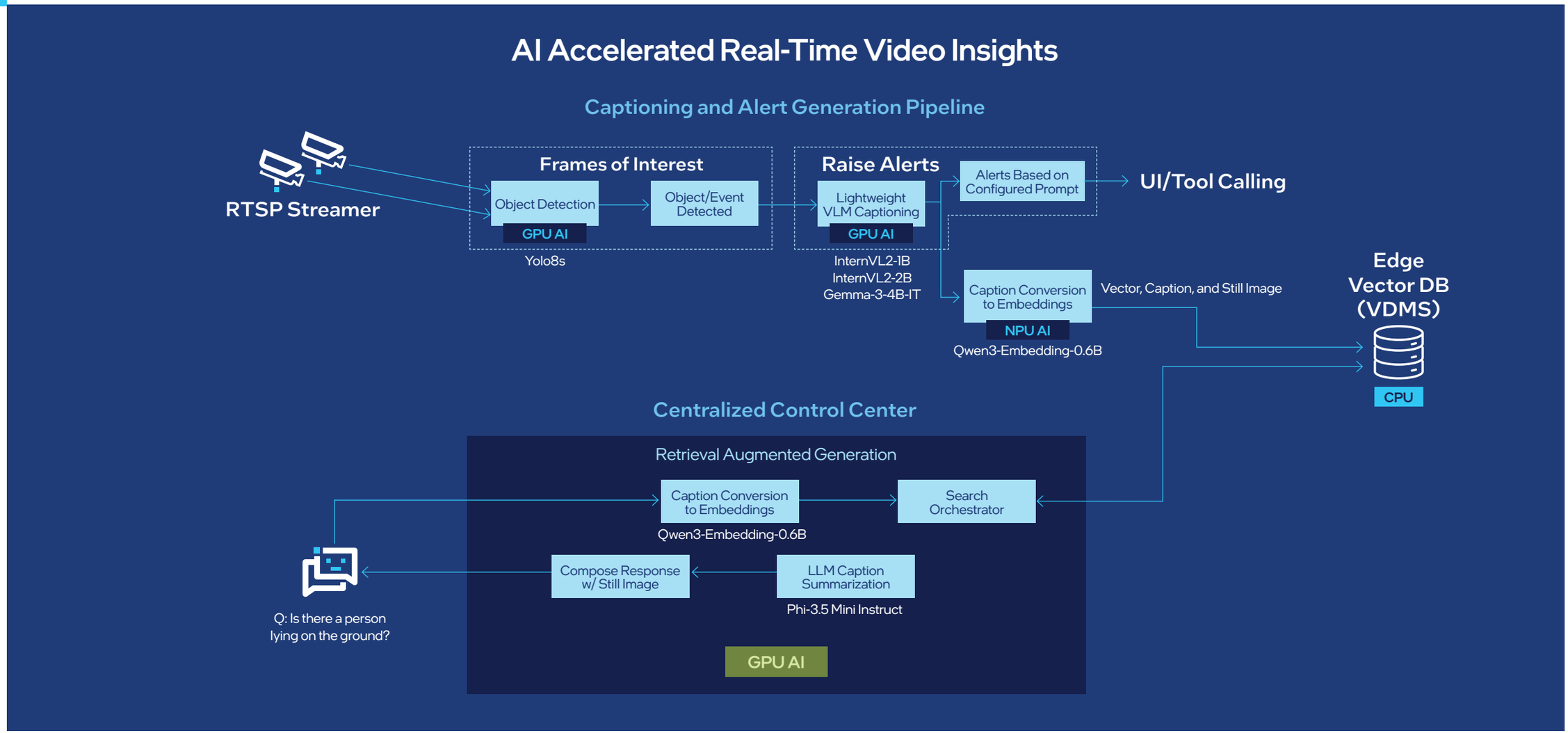


Figure 1: Verified Reference Blueprint System Architecture

4.2 Software Architecture

This Verified Reference Blueprint deploys microservices from Intel's Edge AI Suites to perform AI-powered captioning on live video streams with Intel's Deep Learning Streamer (DL Streamer) and Intel's OpenVINO™. The microservices from the Edge AI Suites are described below. The diagram of the software architecture is shown in Figure 2.

- **live-video-captioning:** Main live video captioning at the edge; configure and launch video pipelines with a web UI.
- **dlstreamer-pipeline-server** Pipeline execution microservice with vision AI models and VLM.
- **live-metrics-service:** ViPPET-based reusable microservice for system metrics collection and real-time streaming.
- **live-alert-agent:** AI-powered alerting for live video streams with VLM; generate real-time alerts based on natural language prompts.
- **multimodal-embedding-serving:** Embedding-vectors storage and retrieval microservice.
- **chat-question-and-answer:** Foundational RAG pipeline that allows users to ask questions and receive answers, querying embedding serving microservice and generate answers using LLM inference

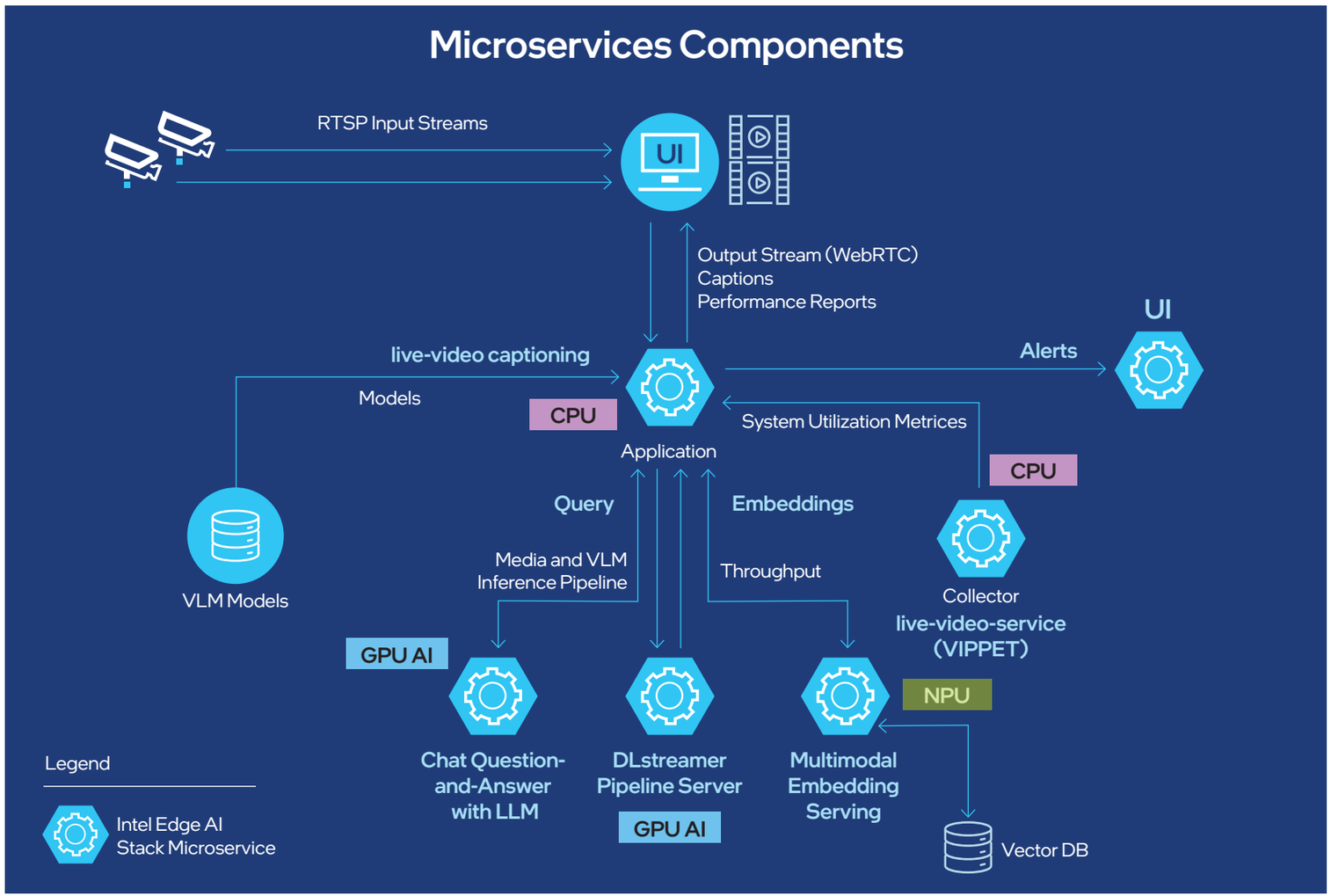


Figure 2: Microservices Components

4.3 Enabling Software Frameworks and Example GUI

This VRB makes use of the live video captioning and the live video captioning RAG sample applications. The links to the software are:

Metro AI Suite - https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite

Live video captioning - https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite/live-video-analysis/live-video-captioning

Live video captioning RAG - https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite/live-video-analysis/live-video-captioning-rag

Video Alert Agent - https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite/live-video-analysis/live-video-alert-agent

Below is additional reference software available in the Metro AI Suite.

Smart NVR - https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite/smart-nvr

Video Search and Summarization - https://github.com/open-edge-platform/edge-ai-suites/tree/main/metro-ai-suite/live-video-analysis/live-video-search

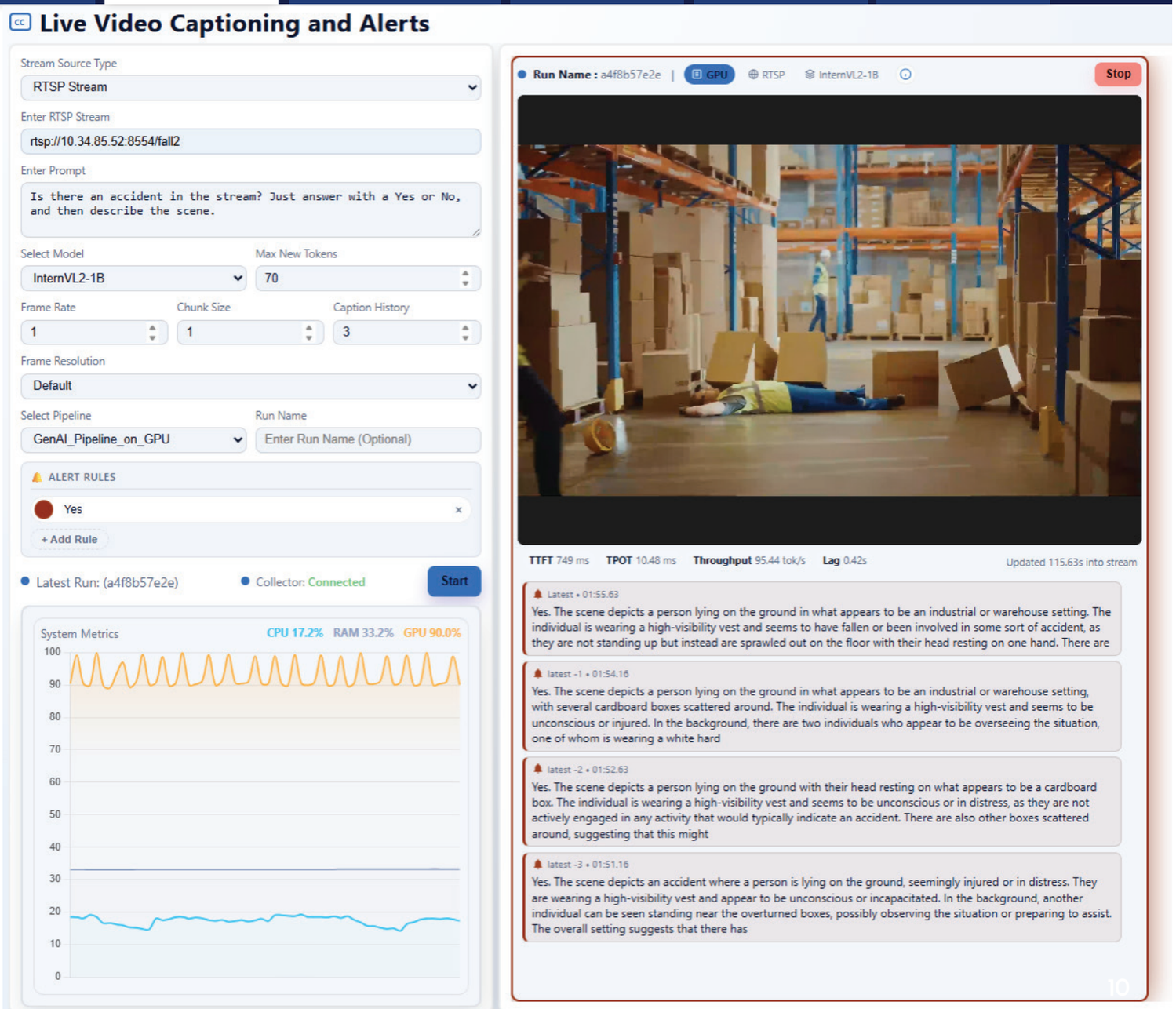


Figure 3: Example Captioning GUI

4.4 Verified Performance Pipeline

This verified reference blueprint includes performance data collected from COTS hardware, specifically systems qualified under Intel® AI Edge Systems using the hardware configuration specified in Section 5.1.

Note: the iGPU frequency runs at maximum dynamic frequency for sustained periods of time. The software stack — including the operating system, drivers, and frameworks — is based on widely available components and is detailed in Section 5.2.

The verified performance results presented later were obtained using the processing pipeline illustrated in Figure 4. This pipeline differs from the previously described architecture, as it was designed to specifically stress the iGPU. The resulting performance metrics are documented in the Verified Performance section.

Steps in the flow diagram are as follows:

1. RTSP frames (1080p video streams) are sent to the video decoder.
2. The YOLOv8 model was not used for performance profiling. Keyframes are decoded and then sent to a VLM for captioning. The VLM outputs metadata which is published over the MQTT bus. The metadata that is published to the Multimodal embedding service container is then converted to embeddings. The embeddings are stored in a vector database. The number of captions that are output to the Edge database are recorded.
3. The TTFT is the time taken to output the first token from the VLM.
4. The Throughput (tokens/s) measures the rate of token generation from the VLM.

*Disclaimer: Data was collected using 1080p video streams under optimal model settings and configuration for vision-language model (VLM) inferencing.

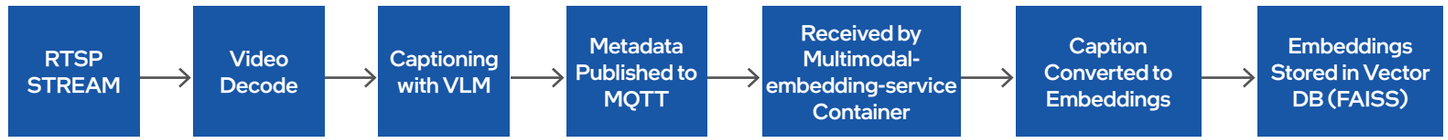


Figure 4: Flow diagram for the verified performance data methodology

5. Hardware and Software Configurations and System Tuning

This blueprint defines the precise hardware, software, and firmware configurations required to maximize multi-modal inference throughput on a COTS Edge AI platform. Rather than relying on specialized, proprietary architectures, this solution achieves optimized Edge AI performance on standard OEM hardware through targeted system tuning, including optimized BIOS settings, operating system parameters, and granular compute allocations. By detailing these "known good" configurations, the blueprint establishes a fully validated reference foundation, significantly reducing engineering overhead, improving performance predictability, and streamlining application onboarding for end users.

This high-performance outcome is made possible by the tight integration of hardware-level tuning with Intel's comprehensive software ecosystem. The end-to-end media and AI pipeline leverages core components of the **Intel Open Edge Platform** including the **Edge AI Suites**, deploying modular microservices and specialized development tools. By harmonizing these hardware optimizations with Intel's publicly available software stack, system integrators can deploy out-of-the-box edge systems that unlock optimum performance.

5.1 Hardware BOM

The hardware configuration of the commercial hardware used in this VRB is detailed below.

Table 1. Hardware Configuration

Ingredient	Requirement
Processor	Intel® Core™ Ultra Series 3 X7 358H with 4 P-cores, 8 E-cores, 4 LP E-cores,
Memory	64GB LPDDR5, 8533 MT/s
Network	Intel® Ethernet Network Adapter i226-V/LM/IT (2.5 Gbps)
Storage (Boot/Capacity Drive)	1 TB or equivalent
iGPU	Intel® Arc™ graphics (part of SoC)
NPU	Intel® AI Boost (part of SoC)

5.2 Software BOM

The software drivers and frameworks utilized to generate verified performance are shown in the table below.

Table 2. Software Configuration

AI Models	Vision Language Model (VLM)	OpenGVLab/InternVL2-1B OpenGVLab/InternVL2-2B Google/Gemma-3-4B-IT
	Large Language Model (LLM)	Microsoft/Phi-3.5-mini-instruct
	Embedding Model	QwenText/qwen3-embedding-0.6b
	Object Detection	YOLOv8s
Frameworks and Runtimes	OpenVINO Toolkit	OpenVINO GenAI 2026.0.0
	DL Streamer Pipeline Server	2026.0.0-ubuntu24
	VDMS Vector DB	intellabs/vdms:v2.11.0
	intel/multimodal-embedding-serving	1.3.2
	FFMPEG	6.1.1
	MQTT Broker	2.0
	ViPPET Collector	2025.2.0
	MediaMTX (WebRTC signaling)	1.16.2
	NGINX (reverse proxy)	N/A
Low-Level Libraries & Drivers	Intel Media Driver (mesa-va-driver)	25.4.2
	Intel NPU Driver	1.32.1
	Intel Graphics and OpenCL Driver	26.14.37833.4
Development Tools	Python	3.12.3
Containerization Tools	Docker	29.2.1
Operating System and Kernel		Ubuntu 24.04 LTS, 6.17.0-23-generic
Open Edge Platform		2026.1

5.3 BIOS Tuning

For optimal performance, Intel recommends the BIOS settings in this table.

Table 3. BIOS Settings

BIOS Setting	Purpose
PCIe BAR Resize	Ensure iGPU is allocated maximum memory
CPU Turbo on	Enable CPU to enter Turbo state
cTDP	Performance mode

5.4 Platform Security

It is recommended that Intel® Boot Guard Technology be enabled on Edge AI hardware to verify the platform firmware as suitable during the boot phase.

In addition to protecting against known attacks, all Intel® Accelerated Solutions recommend installing the Intel® Trusted Platform Module (Intel® TPM). The Intel TPM module enables administrators to secure platforms for a trusted (measured) boot with known trustworthy (measured) firmware and OS. This allows local and remote verification by third parties to advertise known safe conditions for these platforms through the implementation of Intel® Trusted Execution Technology (Intel® TXT).

5.5 Side Channel Mitigation

Intel recommends checking your system's exposure to the "Spectre" and "Meltdown" exploits. This reference implementation has been verified with Spectre and Meltdown exposure using the latest Spectre and Meltdown Mitigation Detection Tool, which confirms the effectiveness of firmware and operating system updates against known attacks.

The spectre-meltdown-checker tool is available for download at GitHub Spectre-Meltdown Checker.

6. Verified Performance Data

Our verified performance data is collected on commercial platforms that are tuned and optimized for performance. The testing was focused on showing the performance of the iGPU. To establish baseline KPIs for multi-modal edge processing, the platform was evaluated on processing density measured in **Captions per Minute (CPM)** using 1080p video streams. Testing concentrated on distinct operational thresholds: an unconstrained **single-stream baseline** to determine peak generation velocity, and a **max-density workload** as streams are added to identify the physical compute limits of the consolidated SoC. The VLM performance was also documented in terms of throughput (Tokens/s) and Time to First Token Latency (TTFT). The reference software used to generate the performance data utilized the Live Video Captioning pipeline in release 2026.1 of the Live Video Analysis reference software in the *Metro AI Suite*.

The models are optimized using OpenVINO at BF16 precision on Intel® Core™ Ultra Series 3 X7 358H. Individual system results may vary as power and performance are affected by use, configuration, and other factors A summary of the Verified Performance data is shown in Table 4.

The benchmarking as shown in Figure 5 highlights a clear relationship between model parameter size and architectural scaling behavior:

- **InternVL2-1B** (Optimized Edge Scaling): Leveraging a highly efficient, compact parameter footprint, this model demonstrated exceptional multi-stream concurrency. It successfully maintained all required operational KPIs while scaling smoothly from a single baseline up to a 4-stream maximum allocation.
- **InternVL2-2B & Gemma-3-4B-IT** (High-Capacity Threshold): Due to the increased compute demands, dense parameter structures, and higher memory bandwidth requirements of these larger architectures, scaling reached its physical threshold at a 3-stream maximum allocation. Attempting a fourth concurrent stream on these models forced processing metrics below our minimum acceptable KPI, defining the absolute boundary for single-node deployment consolidation.

Table 4. Verified Performance Summary Table

Models	No. of 1090p Video Streams	Captions per Minute	Time to First Token Latency TTFT (ms) per stream	Tokens/s	Perf/W (Tokens/s/W)
InternVL2-1B	4	58	2727	15	12
	3	62	1900	29	15
InternVL2-2B	3	34	3911	15	16
Gemma3-4B-IT	3	22	1070	7	13

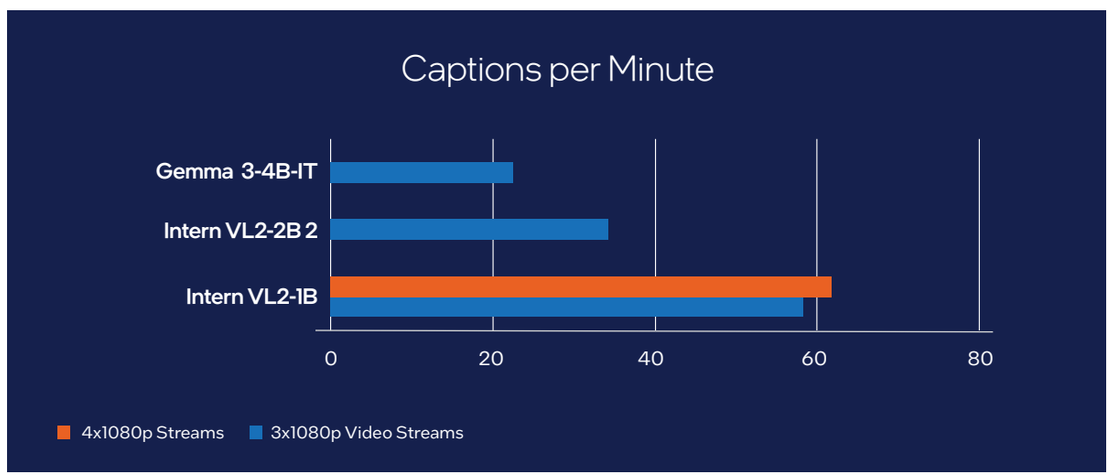


Figure 5: Verified Performance Summary

6.1 Captions Per Minute

The test methodology utilized is outlined in Section 4.4 and strives to utilize the iGPU to maximum capacity.

InternVL2-1B – 62 Captions/min for 3 x 1080p Video streams

- 58 Captions/min for 4 x 1080p Video streams

InternVL2-2B – 32 Captions/min for 3 x 1080p Video streams

Gemma3-4B-IT – 23 Captions/min for 3 x 1080p Video streams

An interesting observation arises when comparing the single-stream maximum captions per minute performance to the cumulative maximum captions per minute when utilizing the maximum number of supported streams. In many cases, the cumulative throughput with multiple streams is close to or even exceeds the single-stream maximum, indicating efficient parallelization and resource utilization.

For example:

- **InternVL2-1B:** Intel Core Ultra Series 3 X7 358H can support a cumulative throughput of 62 captions per minute across three concurrent 1080p video streams using the InternVL2-1B Multi-modal model at BF16 precision for an AI Accelerated Real-time Video Insights workload. It can also support a cumulative throughput of 58 captions per minute across four concurrent 1080p video streams using the InternVL2-1B Multi-modal model at BF16 precision for an AI Accelerated Real-time Video Insights workload.
- **InternVL2-2B:** Intel Core Ultra Series 3 X7 358H can support a cumulative throughput of 34 captions per minute across three concurrent 1080p video streams using the InternVL2-2B Multi-modal model at BF16 precision for an AI Accelerated Real-time Video Insights workload.
- **Gemma-3-4B-IT:** Intel Core Ultra Series 3 X7 358H can support a cumulative throughput of 23 captions per minute across three concurrent 1080p video streams using the Gemma-3-4B-IT Multi-modal model at BF16 precision for an AI Accelerated Real-time Video Insights workload.

These results highlight that while larger models reduce overall frequency of captioning, it provides more detail object-of-interest and scene description for better recall accuracy. This trade-off between model size, accuracy, and throughput should be carefully considered when designing and deploying captioning systems, balancing the need for accuracy with performance requirements.

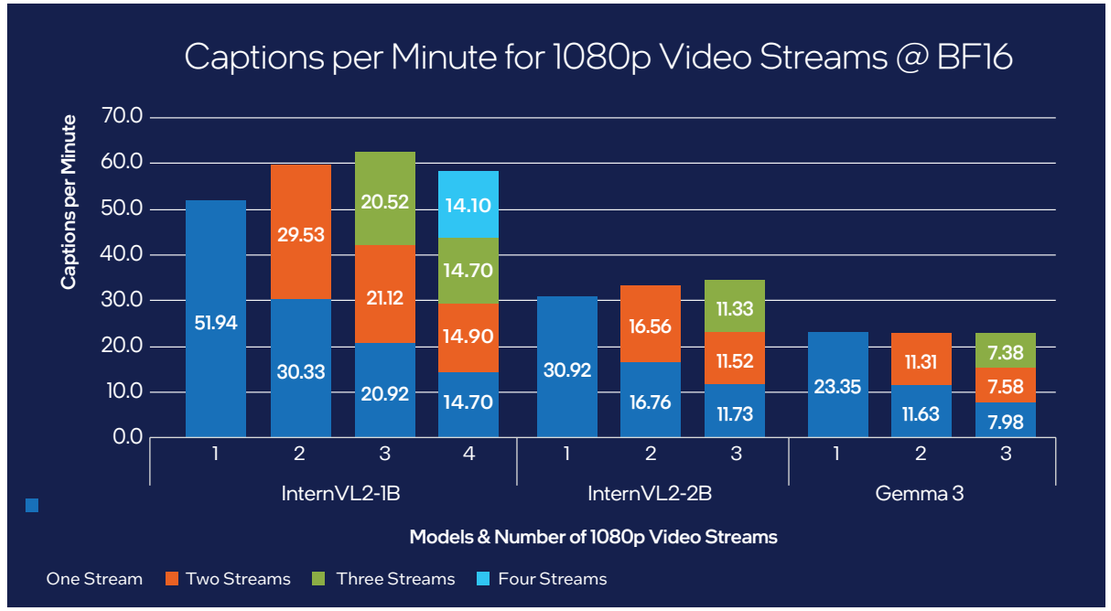


Figure 6: Verified Performance Data Captions Per Minute

6.2 VLM Performance

The throughput of the VLMs was measured using the methodology described in Section 4.4 and is charted below. The throughput is defined in terms of tokens/second and is shown in Figure 6 "Verified Performance data for Throughput Tokens/s."

- **InternVL2-1B:** Intel Core Ultra Series 3 X7 358H can support 15 tokens/s per stream for an AI Accelerated Real-Time Video insights workload while running four 1080p video streams using the InternVL2 1B Multi-modal model on iGPU at BF16 Precision.
- **InternVL2-2B:** Intel Core Ultra Series 3 X7 358H can support 15 tokens/s per stream for an AI Accelerated Real-Time Video insights workload while running three 1080p video streams using the InternVL2 2B Multi-modal model on iGPU at BF16 Precision.
- **Gemma-3-4B:** Intel Core Ultra Series 3 X7 358H can support 7 tokens/s per stream for an AI Accelerated Real-Time Video insights workload while running three 1080p video streams using the Gemma 3 4B IT Multi-modal model on iGPU at BF16 Precision.

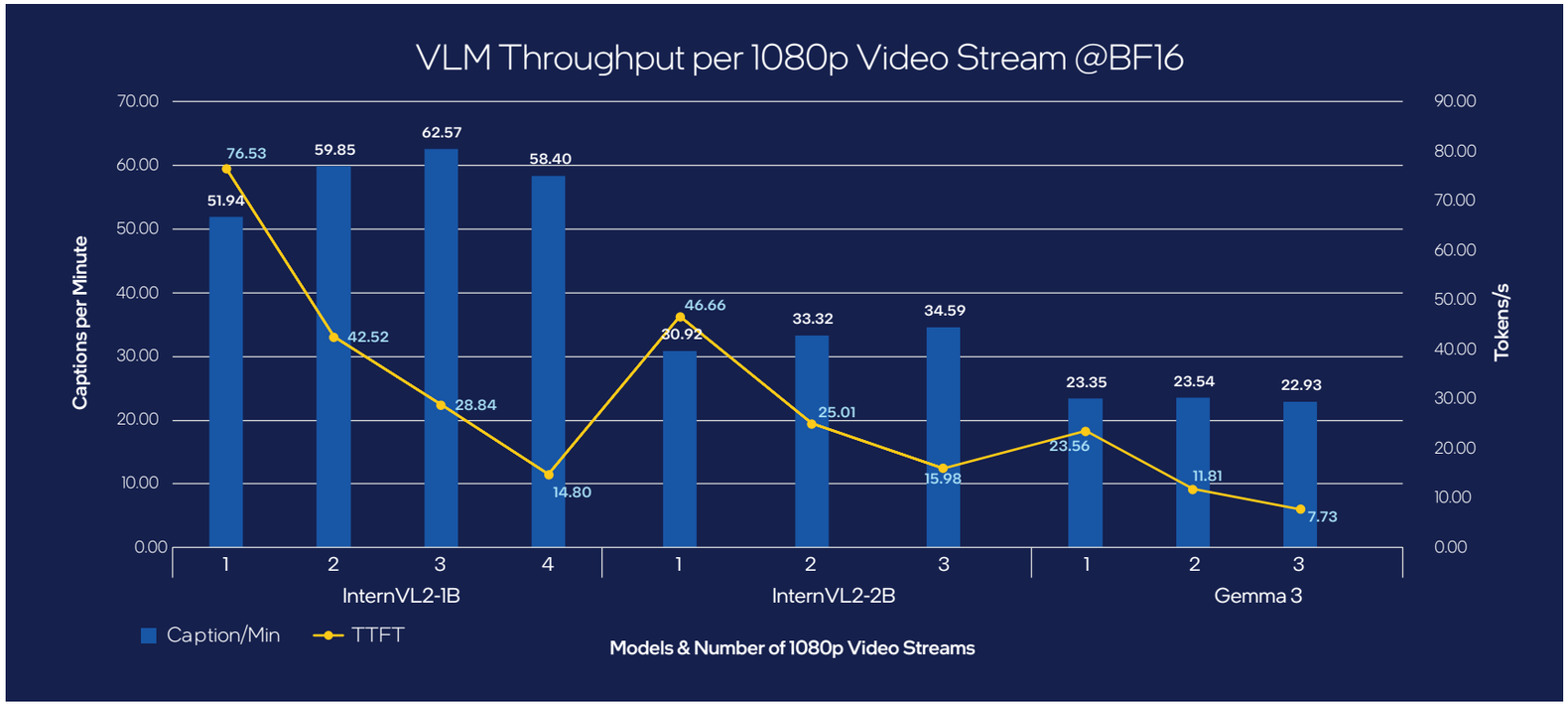


Figure 7: Verified Performance data for Throughput Tokens/s

The time to first token (TTFT) latency for the models was measured. The metric described is in milliseconds.

InternVL2-1B: The TTFT latency of Intel Core Ultra Series 3 X7 358H is 2727 ms per stream for an AI Accelerated Real-Time Video Insights workload while running four 1080p video streams using the InternVL2 1B Multi-modal model on the iGPU at BF16 precision.

InternVL2-2B: The TTFT latency of Intel Core Ultra Series 3 X7 358H is 3911 ms per stream for an AI Accelerated Real-Time Video Insights workload while running three 1080p video streams using the InternVL2 2B Multi-modal model on the iGPU at BF16 precision.

Gemma3: The TTFT latency of Intel Core Ultra Series 3 X7 358H is 1070 ms per stream for an AI Accelerated Real-Time Video Insights workload while running three 1080p video streams using the Gemma 3 4B IT Multi-modal model on the iGPU at BF16 precision.

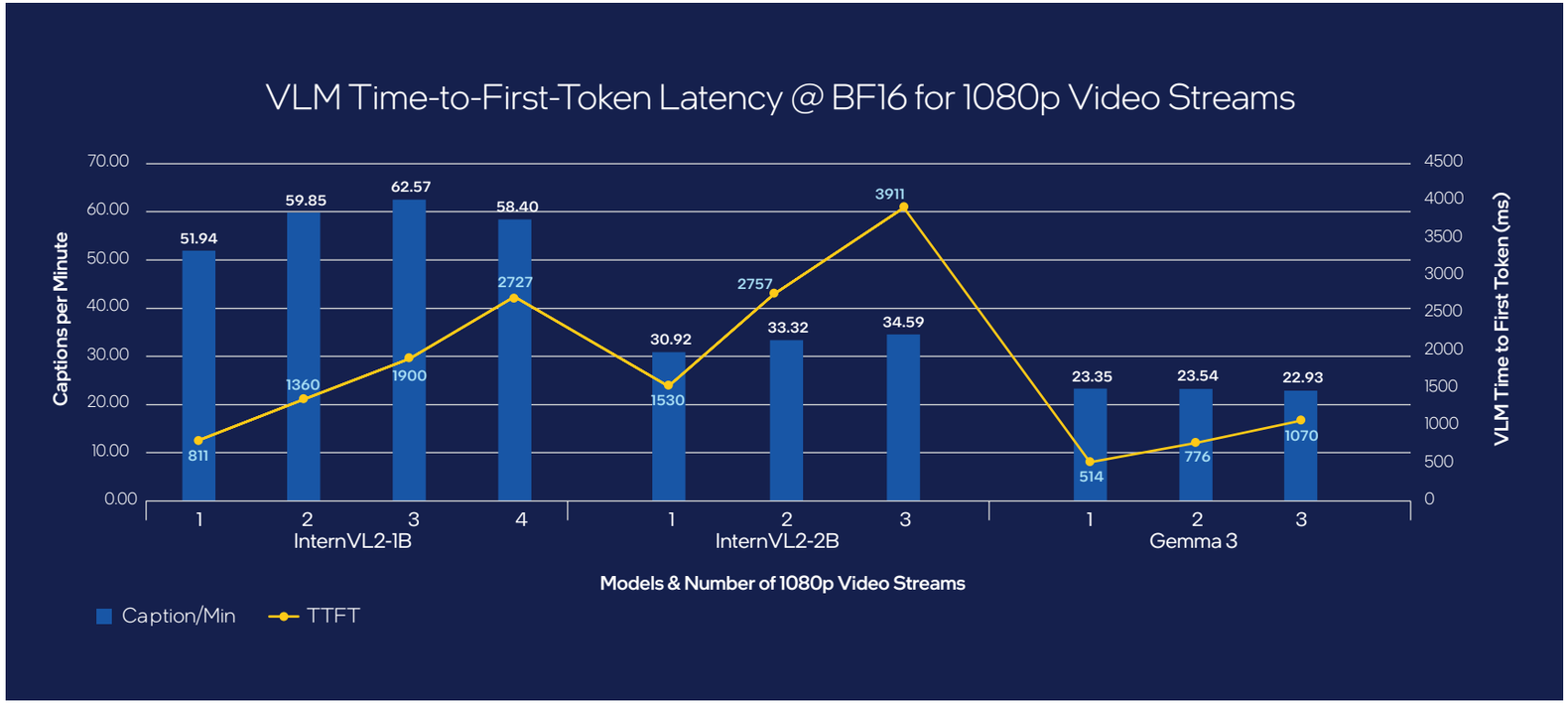


Figure 8: Verified Performance Data for Time to First Token Latency

The power dissipated by the iGPU while executing the pipeline as described in Section 4.4 is detailed below. The metric used is power per stream.

InternVL2-1B: The Power per stream of Intel Core Ultra Series 3 X7 358H is 12W/stream for an AI Accelerated Real-Time Video Insights workload while running four 1080p streams using the InternVL2 1B Multi-modal model on the iGPU at BF16 precision.

InternVL2-2B: The Power per stream of Intel Core Ultra Series 3 X7 358H is 16W/stream for an AI Accelerated Real-Time Video Insights workload while running three 1080p streams using the InternVL2 2B Multi-modal model on the iGPU at BF16 precision.

Gemma-3: The Power per stream of Intel Core Ultra Series 3 X7 358H is 13W/stream for an AI Accelerated Real-Time Video Insights workload while running three 1080p streams using the Gemma 3 4B IT Multi-modal model on then iGPU at BF16 precision.

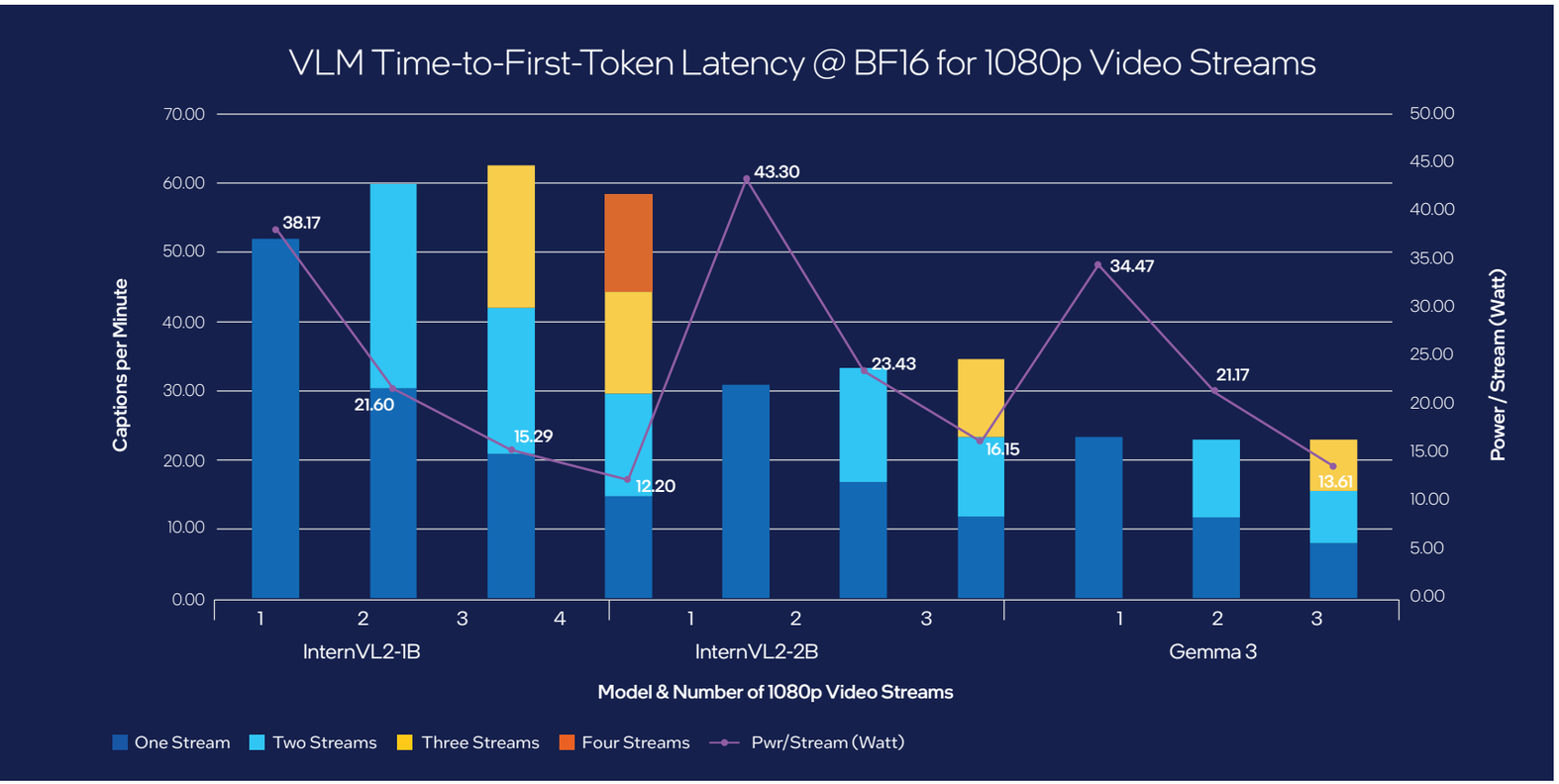


Figure 9: Verified Performance Data Power/Stream

7. Conclusion

This **VRB** describes a recipe (HW BOM and SW BOM) that delivers verified, repeatable performance for an AI Accelerated Real-Time Video Insights use case. It will serve as a definitive roadmap for organizations seeking to transition complex AI workloads from the data center to the edge. By consolidating high-performance vision-language modeling onto a single, power-efficient SoC, this framework eliminates the traditional requirement for server-grade workstations and expensive, power-hungry discrete GPUs. This architectural shift significantly reduces TCO by removing recurring cloud-based token costs and lowering overall energy consumption. Beyond the financial advantages, the solution provides an essential layer of data sovereignty and security, ensuring that sensitive video and audio feeds remain processed locally, meeting the stringent privacy requirements of healthcare, education, and public safety sectors.

The versatility of this pipeline — proven across diverse verticals ranging from industrial worker safety to automated classroom summarization — is a testament to the robustness of Intel's **Open Edge Platform**. This ecosystem provides the necessary building blocks, including modular microservices and performance-optimized SDKs like the **Edge AI Suite**, to simplify the development cycle for independent software vendors and system integrators. This blueprint is not a theoretical exercise; it is backed by **verified data** and benchmarking that empowers partners to deploy COTS systems with total confidence in their performance and reliability.

Ultimately, the synergy of the **Intel Core Ultra Series 3** provides a singular, transformative solution for the next generation of edge computing. By intelligently distributing workloads across the **integrated Intel Arc GPU** for high-throughput detection and the **dedicated NPU** for conversational RAG interfaces, Intel delivers a complete, multi-modal AI stack in a single package. This integrated approach allows partners to modernize their infrastructure with state-of-the-art VLM capabilities at a fraction of the hardware footprint and cost of traditional architectures, redefining the possibilities of real-time intelligence at the edge.

InternVL2-1B

4 camera streams @1080p with cumulative throughput of 58 captions per sec

VLM Throughput = 15 tokens/s for 4x1080p streams

VLM TTFT = 2727ms

VLM Power/stream = 12W

InternVL2-2B

3 camera streams @1080p with cumulative throughput of 34 captions per sec

VLM Throughput = 16 tokens/s for 3x1080p streams

VLM TTFT = 3911ms

VLM Power/stream = 16W

Gemma3-4B_IT

3 camera streams @1080p with cumulative throughput of 22 captions per sec

VLM Throughput = 8 tokens/s for 3x1080p streams

VLM TTFT = 1070ms

VLM Power/stream = 13W

8. References

8.1 List of Figures

8.2 List of Tables

Notices & Disclaimers

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visiting the Intel Resource and Documentation Center.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

No product or component can be absolutely secure.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

0626/EC/MESH/PDF 367774-001US

intel