DELL Technologies

Proof of Concept (PoC)

AI-in-a-Box PoC version 2: Accelerating GenAI Innovations

AI-in-a-Box PoC delivers an integrated solution to explore general-purpose GenAI workloads alongside use-case-specific video analytics. This Proof of Concept (PoC) validates both **Large Language Model (LLM) and Vision Language Model (VLM) customization and inferencing,** while also benchmarking **NVIDIA's Video Search and Summarization (VSS) Blueprint** which integrates multimodal models to process video streams, generate captions, and produce semantic summaries for rapid retrieval.

Quickstart AI Platform

Single node AI-**optimized** Dell PowerEdge to explore AI use cases.

Robust capabilities

This solution provides:

- Large Language Model (LLM) customization and inferencing
- Vision Language Model (VLM) customization and inferencing
- NVIDIA VSS blueprint testing, which offers a reference design for video search and summarization workloads

Joint collaborations

Validated by:

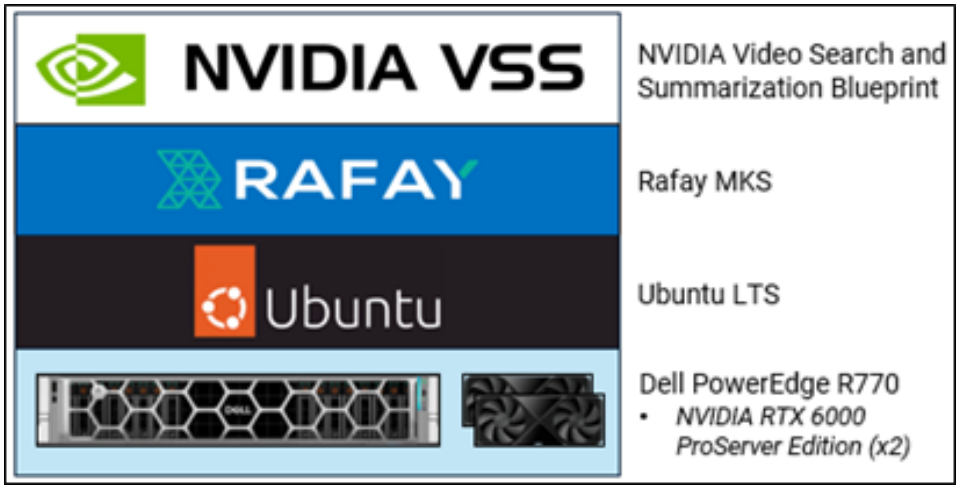
- APJ AI Innovation Hub

Supported by:

- Global Smart Cities
- CSC Singapore

High-level solution architecture

The architecture combines a single node AI-optimized Dell PowerEdge server with NVIDIA GPUs, running Ubuntu OS and Rafay's MKS, an enterprise-grade, upstream Kubernetes distribution from Rafay. By running Rafay MKS, Kubernetes clusters can be monitored and managed centrally through a single pane of glass. Additionally, LLM and VLM-based applications can be quicky deployed, customized, and tested.



Solution components

HW Component	Specification
Server	Dell PowerEdge R770
CPU	Intel® Xeon® 6 Performance 6737P x2 (32C/64T, 2.9 GHz Base)
Memory	32GB DDR-5 DIMM 6400 MT/s x 16
GPU	NVIDIA RTX PRO 6000 Server Edition
Local storage	OS Disk - Local SSD Drives – 894 GB (RAID-1) Data Disk - Local SSD Drives – 14.3 TB (RAID-5)

SW Component	Specification
Operating System	Ubuntu Server 24.04 LTS
Platform	Rafay MKS (Kubernetes v1.32.4)
Model Customization	NVIDIA NeMo 2.0 (nemo:25.11.01)
Model Inferencing	LLM - AIPerf (nvcr.io/nvidia/tritonserver:25.05-py3-sdk) VLM - Genai-Perf (nvcr.io/nvidia/tritonserver:24.10-py3-sdk)
Use Case	NVIDIA Video Search and Summarization Blueprint (Helm Chart v2.4.0)

Model Customization

Model customization (also known as fine-tuning) adapts pre-trained LLM and VLM models for domain-specific tasks without retraining all parameters.

For this PoC, the solution uses NVIDIA NeMo 2.0 Framework with parameter-efficient fine-tuning (PEFT) and supervised fine-tuning (SFT) strategies for model customization.

Robustly Validated

Rigorously tested for model customization with widely used industry and enterprise grade LLM and customization technique.

LLM Customization Benchmark

Test configuration

Parameters	Settings
Model	Llama 3 8B Instruct
Framework	NVIDIA NeMo 2.0
Customization Modes	PEFT (LoRA), SFT

Time to customize (fine-tune) model in minutes (1000 steps)

Model	No. of GPUs	LoRA	SFT
Llama 3 (8B)	1	278	N.A.
	2	138	447

Note:

- These timings exclude the loading of the model, the dataset, and model validation.
- For LoRA multi-GPU fine-tuning, data parallelism is used.
- Dataset used for Model Customization—SQuAD: 100,000+ Questions for Machine Comprehension of Text
- Based on the nemo:25.11.01 container image.

Model Inferencing

Model inferencing is the process of evaluating pre-trained LLM and VLM models for performance under production-like conditions.

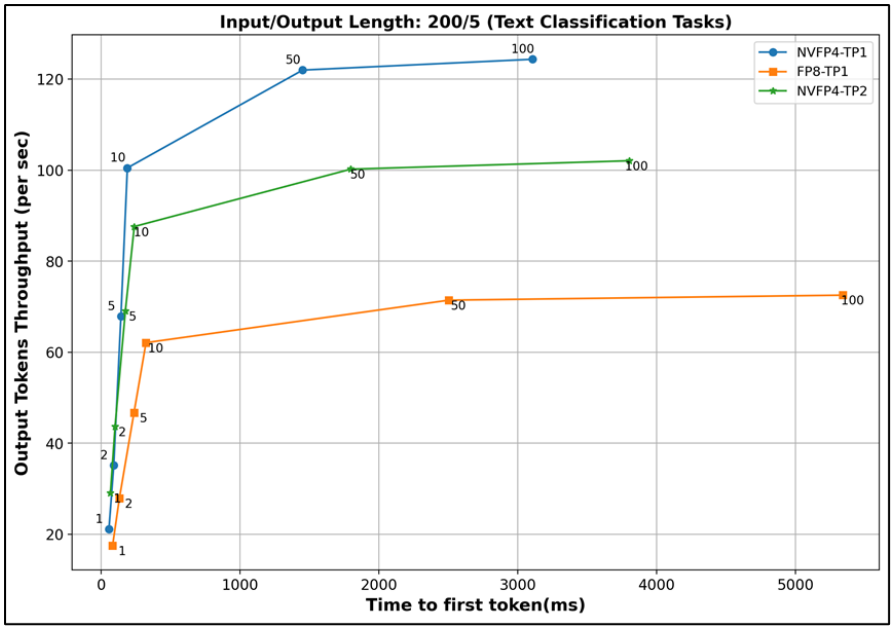
For this PoC, the solution uses NVIDIA NIM microservices for serving models and genai-perf to benchmark latency and throughput metrics.

LLM Inferencing Benchmark

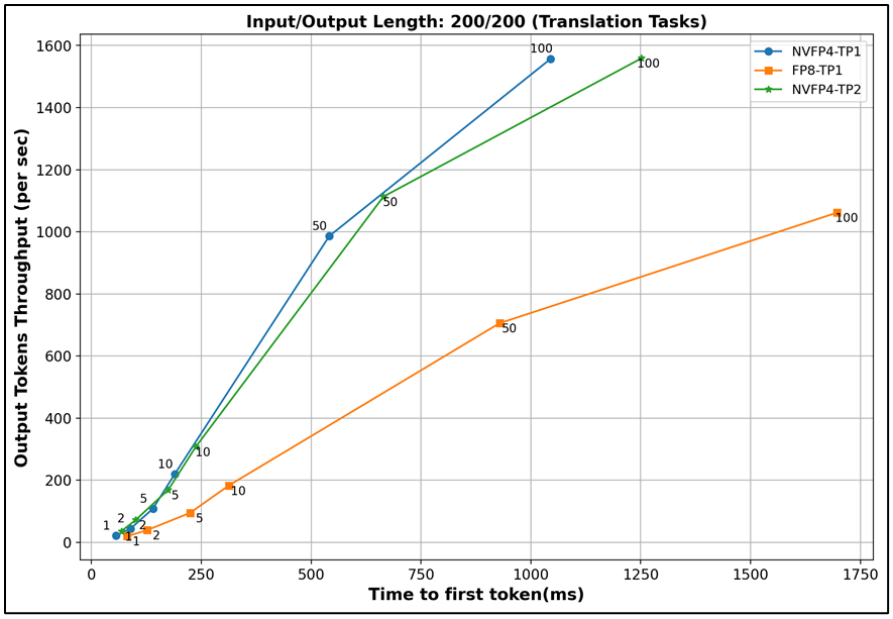
Test configuration

Parameters	Settings
Model	Llama 3.3 70B Instruct
Concurrencies	1, 2, 5, 10, 50, 100
Deployment Modes	NVFP4,FP8,NVFP4 (TP2)
LLM Task Profiles	200/5 – Text Classification Tasks
	200/200 – Translation Tasks
	1000/200 – Text Summarization Tasks
	100/1000 – Content Generation Tasks

Result 1—Input/Output Length: 200/5 (Text Classification Tasks)



Result 2—Input/Output Length: 200/200 (Translation Tasks)

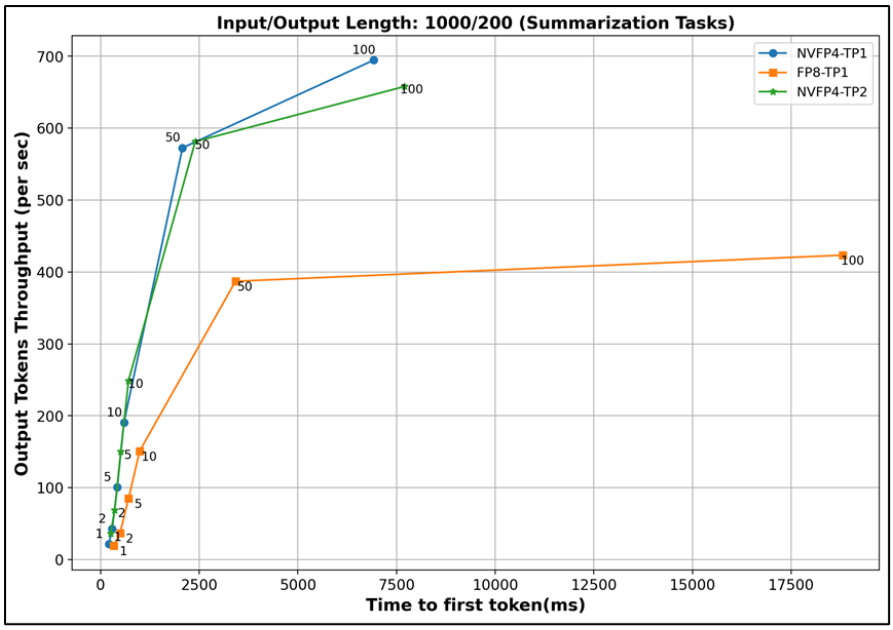


Analysis Of Quantization And Parallelism

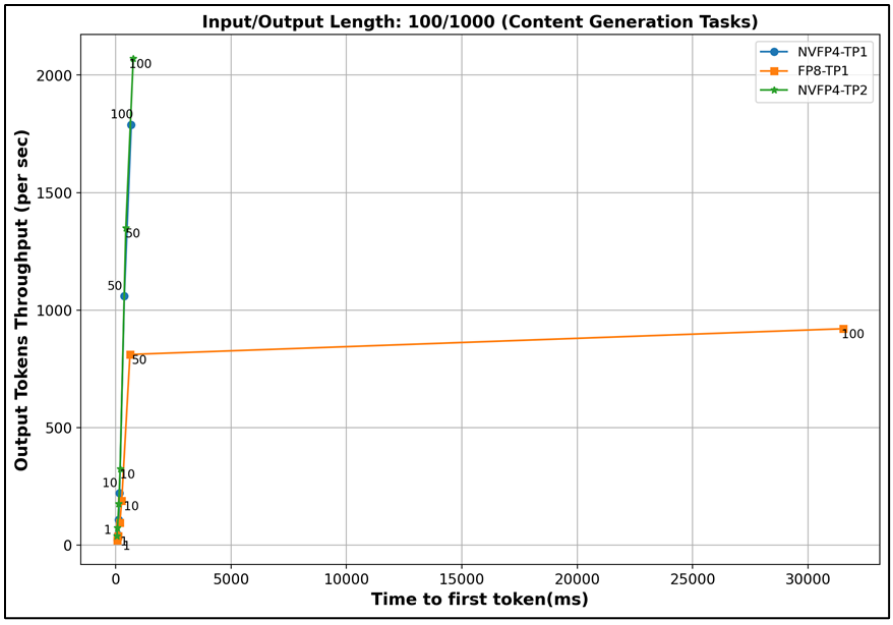
The evaluation results indicate that the LLM configured with NVFP4 quantization consistently achieves higher throughput and lower latency across all task profiles compared to the same model running with FP8 quantization. Each data point in the performance charts is annotated with its corresponding concurrency level.

Additionally, the impact of tensor parallelism (TP2)—spanning two GPUs—shows a noticeable improvement in throughput only for workloads involving large output sequence lengths. For smaller sequences, the benefits of TP2 are minimal.

Result 3—Input/Output Length: 1000/200 (Summarization Tasks)



Result 4—Input/Output Length: 100/1000 (Content Generation Tasks)



VLM Inferencing Benchmark— Time-to-First Token (TTFT)

Test configuration

Parameters	Settings
Model	Cosmos-Reason1-7B
Concurrencies	1, 2, 5, 10, 50, 100, 180, 250
Deployment Modes	FP8, BF16, FP8 (TP2), BF16 (TP2)
Input/Output Sequence Length	1000
Input Image Size	1120 x 1120 pixels

Analysis Of Quantization And Parallelism

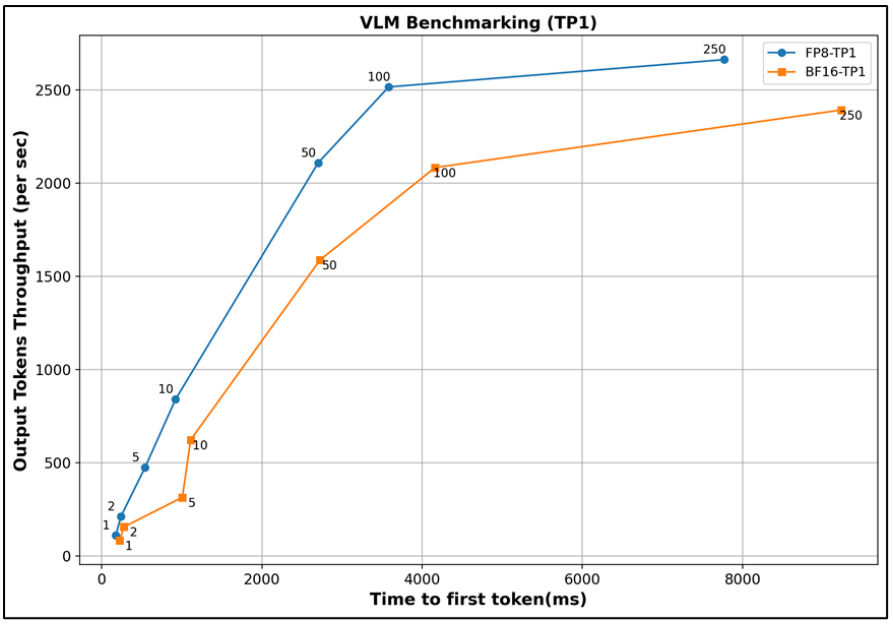
The results demonstrate that the VLM configured with FP8 quantization consistently delivers higher throughput and lower latency compared to the same model operating with BF16 precision. This performance advantage is evident across the evaluated task profiles.

A similar trend is observed when enabling tensor parallelism (TP2) across two GPUs, where throughput and latency metrics improve relative to single-GPU execution.

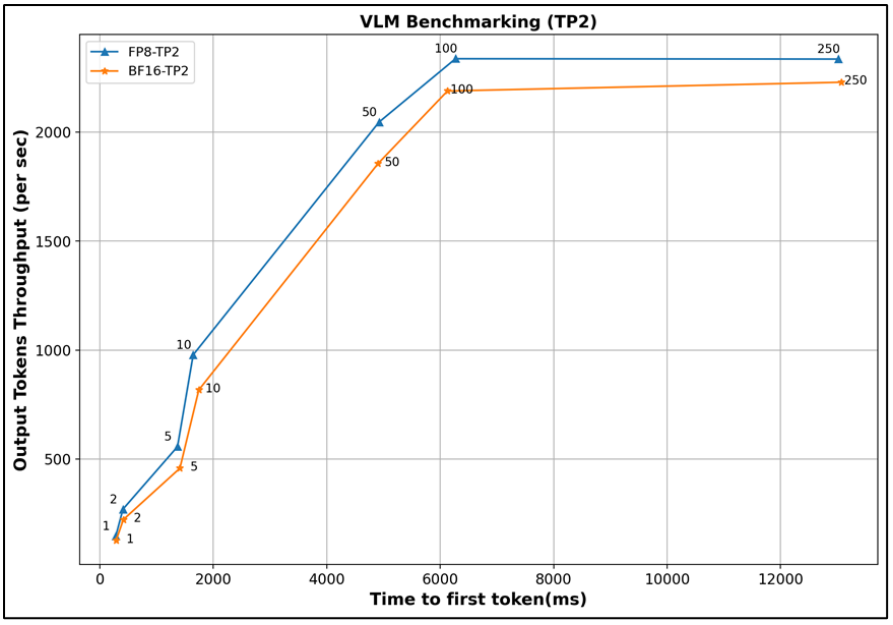
Benchmarks Ran On The NVIDIA VSS Blueprint

- Benchmark 1
  Time to Process Video Files
- Benchmark 2
  Time to Process Video Files (Captions Only)
- Benchmark 3
  Maximum number of Concurrent Video Streams
- Benchmark 4
  Maximum number of Concurrent Video Streams (Captions Only)

Result 1—VLM (Cosmos-Reason1-7B) Benchmarking (TP1)



Result 2—VLM (Cosmos-Reason1-7B) Benchmarking (TP2)



Use Case

NVIDIA VSS Blueprint

The NVIDIA VSS Blueprint provides a reference architecture for video search and summarization workloads, integrating multimodal models for captioning and semantic summarization.

In this PoC, the blueprint was deployed on Rafay MKS in a fully local environment, ensuring all components run on-premises for optimal performance and security.

Blueprint components

Components	Settings
VLM	Cosmos-Reason1-7B (FP16)
LLM	Llama 3.3 70B Instruct (NVFP4)

Benchmark 1—Time Taken to Process Video Files

Evaluates the efficiency of video summarization by measuring the time required to process high-definition video files under a standardized configuration.

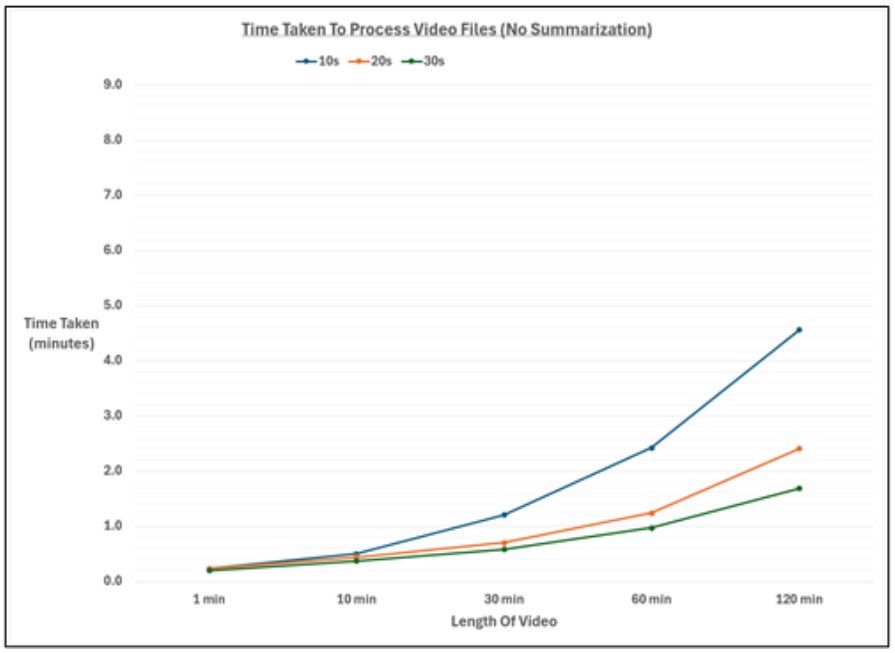
Test configuration

Parameters	Settings
Video Resolution	1920 x 1080 pixels
Video Frame Rate	30 FPS
Video Codec	H.264
Video Chunk Size	10, 20, 30 seconds
CA RAG	GraphRAG

Benchmark 1 Result

The smaller the video chunk (or longer video duration), the larger the number of frames to be processed from a single video file, amounting to a longer processing time.

Result



Benchmark 2—Time Taken To Process Video Files (Captions Only)

Focuses on the VLM as captions generated are not added into a Knowledge Graph database. There is one VLM instance.

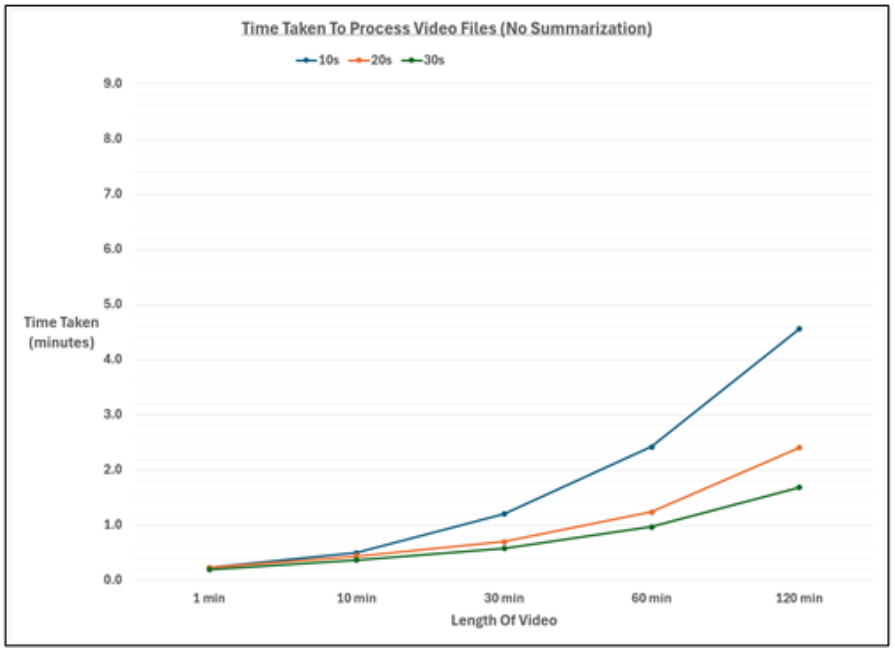
Test configuration

Parameters	Settings
Video Resolution	1920 x 1080 pixels
Video Frame Rate	30 FPS
Video Codec	H.264
Video Chunk Size	10, 20, 30 seconds

Benchmark 2 Result

The times taken to process the video files in this Benchmark are considerably shorter compared to Benchmark 1 since summarization is being excluded.

Result



Benchmark 3—Maximum Number of Concurrent Video Streams

The VLM generates captions that describes the contents and events happening in the video.

Test configuration

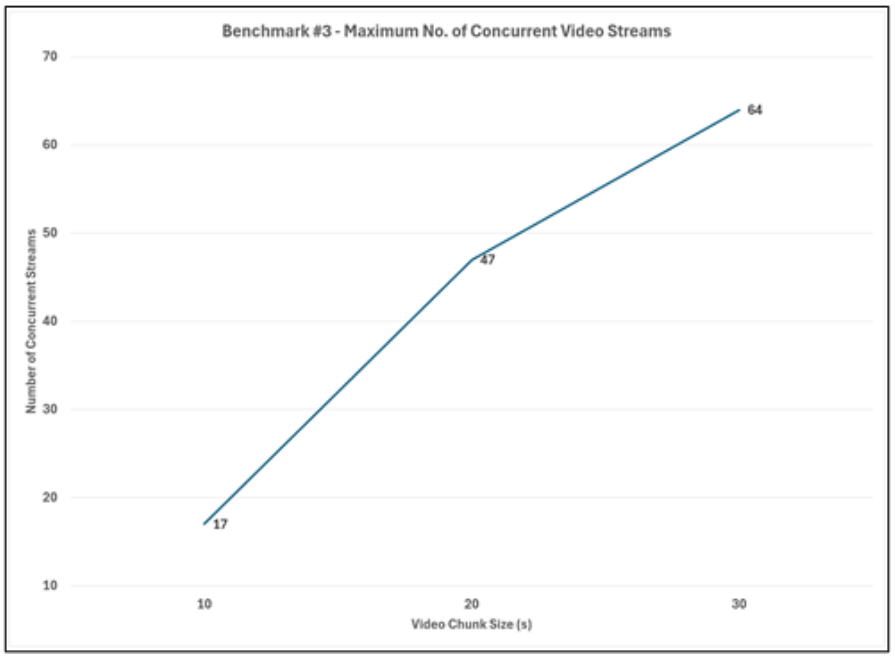
Parameters	Settings
Video Resolution	1920 x 1080 pixels
Video Frame Rate	30 FPS
Video Codec	H.264
Video Chunk Size	10, 20, 30 seconds
Summary Duration	60 seconds
CA RAG	GraphRAG
VLM Max Token	1024

Benchmark 3 Result

The smaller the video chunk, the larger the number of frames to be processed, amounting to a longer processing time.

Note:
A new knowledge graph database was used for every test run.

Result



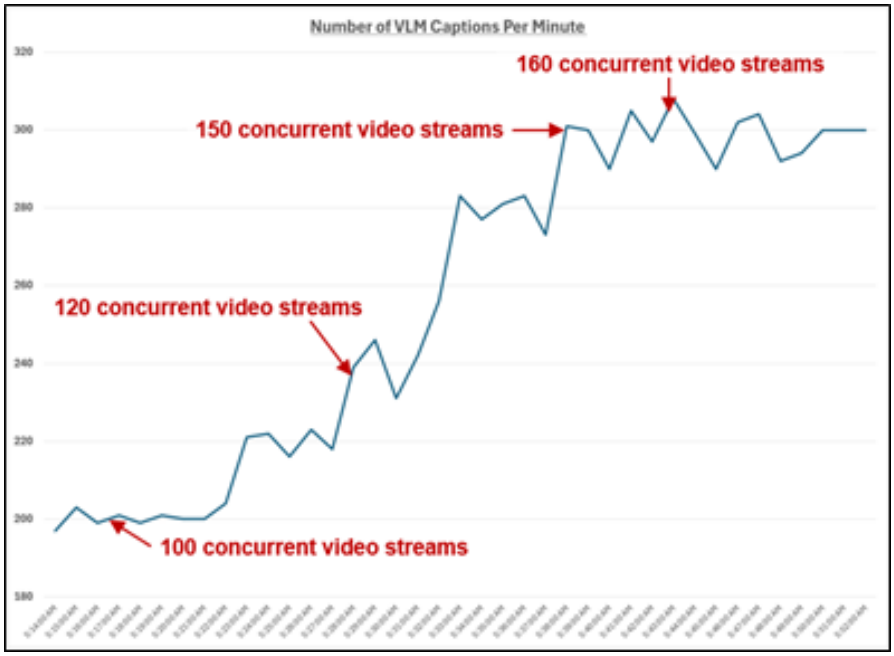
Benchmark 4—Maximum Number of Concurrent Video Streams (Captions Only)

Focuses on the VLM as captions generated are not added into a Knowledge Graph database. There are two VLM instances.

Test configuration

Parameters	Settings
Video Resolution	1920 x 1080 pixels
Video Frame Rate	30 FPS
Video Codec	H.264
Video Chunk Size	30 seconds
Model Reasoning	Enabled
Number of VLM Instances	2
Video Bitrate	2.5 Mbps (approximately)

Result—Number of Captions Generated by VLM Per Min



Benchmark 4 Result

It is seen that the number of captions generated peaked at approximately 150 concurrent video streams.

Conclusion

This proof-of-concept delivers a scalable, efficient, and domain-optimized AI solution for video search and summarization. Powered by the latest NVIDIA RTX PRO 6000 Server Edition GPUs, the AIB v2 delivers the performance headroom needed for modern intelligent video analytics powered by vision-language models (VLMs). In our PoC testing, the AIB v2 handled multiple simultaneous video streams, turning raw video into timely, high-value insights.

The server can be deployed as a single-node cluster (as in this PoC) or integrated as a compute node within a larger cluster, providing flexibility in infrastructure architecture. Its power-efficient design also makes it well-suited for edge deployments where available power capacity is limited.

In essence, this solution is not just an AI enhancement—it is a strategic enabler for intelligent video analytics, delivering measurable efficiency gains and operational value across diverse surveillance ecosystems.

Deployable Out-Of-Box

Optimized for maximum GPU compute, delivering exceptional performance and power efficiency for demanding LLM and VLM workloads, while integrating seamlessly into your existing data center environment.

Appendix A

This appendix provides a list of key technical terms relevant to the topics discussed in this document. Each term is accompanied by a brief definition to enhance understanding. Additionally, we have included URL links to reputable resources that offer further explanations and insights into these terms. This section serves as a valuable reference for readers seeking to deepen their understanding on the information presented in this document.

Key terms

Model Customization	The process of adapting a pretrained foundation model to perform a specific task or cater to a particular domain.
Model Inferencing	The process of using a pretrained model to generate predictions, make decisions, or produce outputs based on specific input data and contexts.
Parameter-Efficient Fine-Tuning (PEFT)	A technique that adapts large pre-trained models to new tasks by training only a small subset of parameters, reducing computational cost and memory usage while preserving model performance.
Video Search and Summarization (VSS)	The process of analyzing video content to enable fast retrieval of relevant segments and generating concise summaries for efficient review and decision-making

References

Customization:

- PEFT in NeMo 2.0 — NVIDIA NeMo Framework User Guide
- SQuAD: 100,000+ Questions for Machine Comprehension of Text (Dataset used for Model Customization)

Inferencing:

- LLM Inference Benchmarking Guide: NVIDIA GenAI-Perf and NIM | NVIDIA Technical Blog
- Benchmarking — NVIDIA NIM for Vision Language Models (VLMs)
- Using AIPerf to Benchmark — NVIDIA NIM LLMs Benchmarking