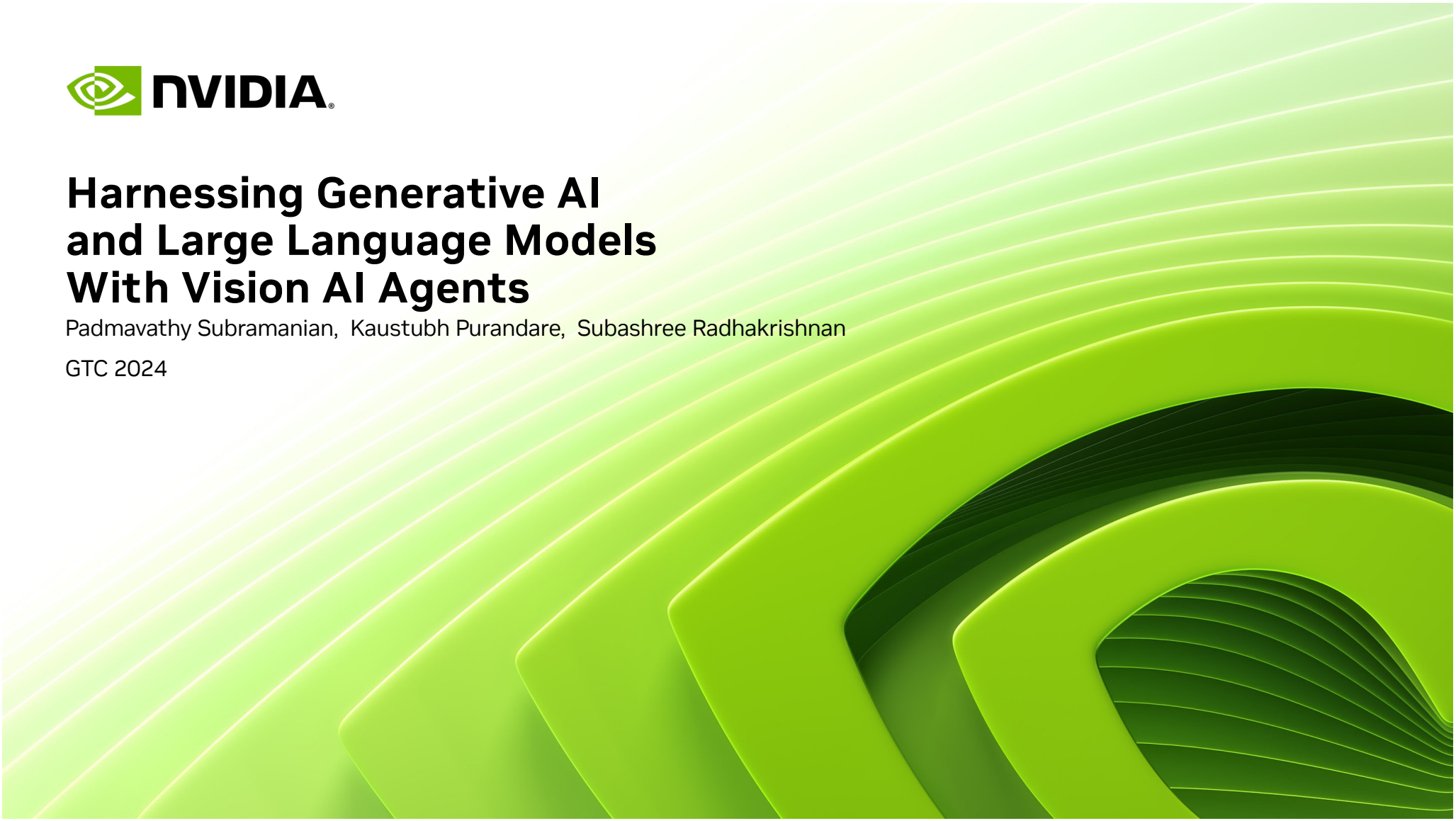


Harnessing Generative AI and Large Language Models With Vision AI Agents

Padmavathy Subramanian, Kaustubh Purandare, Subashree Radhakrishnan

GTC 2024

Agenda

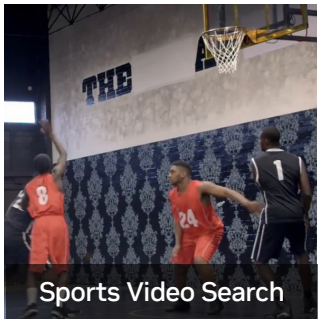
- Why Multi-modal AI
- Introducing Visual Insight Agent
- Architecture
- Workflows
- Vision-Language Models – Customization
- VLM Challenges
- CTA

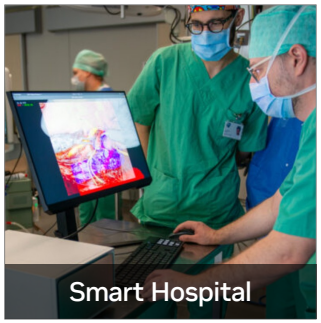
Why Multi-modal AI?

Multi-modal models autonomously provide video understanding and insights

Automated Inspection


Process Automation


Sports Video Search


Smart Hospital

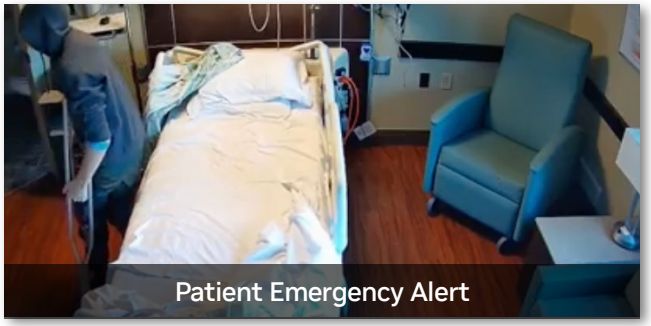
Challenges

- Petabytes of videos, images
- Live streams
- Scene understanding
- Proactive alerts

Requirements

- Custom data
- Accuracy
- Context
- Video, audio, OCR
- Low latency

Are Pallets Being Moved Out?


Patient Emergency Alert

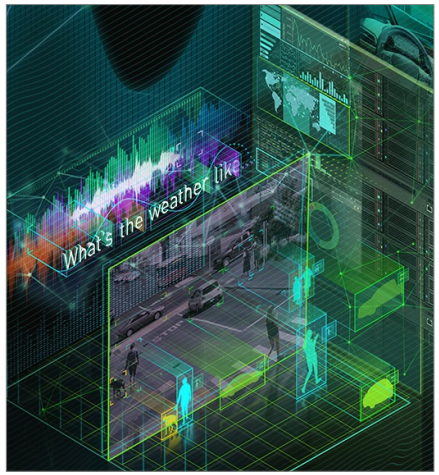
- AI automation for every industry
- Multiple sensors
- Petabytes of multi-modal data

- Cloud native microservices
- Ability to customize models
- Ease of creating vision AI agents

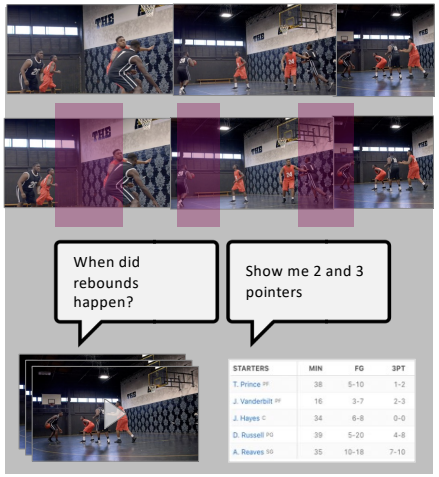
- Enterprise operational excellence
- Natural language interaction
- Contextual insights, proactive alerts

Generative AI use cases across Industries

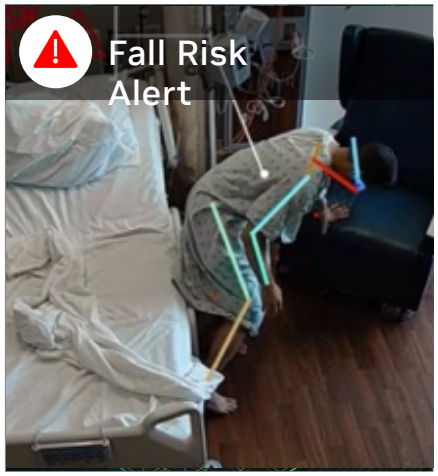
Warehouse and Logistics | Intelligent Analytics | Smart Space | Manufacturing | Healthcare | Media & Entertainment | Retail | Public Safety | Intelligent Traffic Systems



Contextual Understanding
Generalization



Spatio-temporal Localization
Insights correlating space and time, Q&A



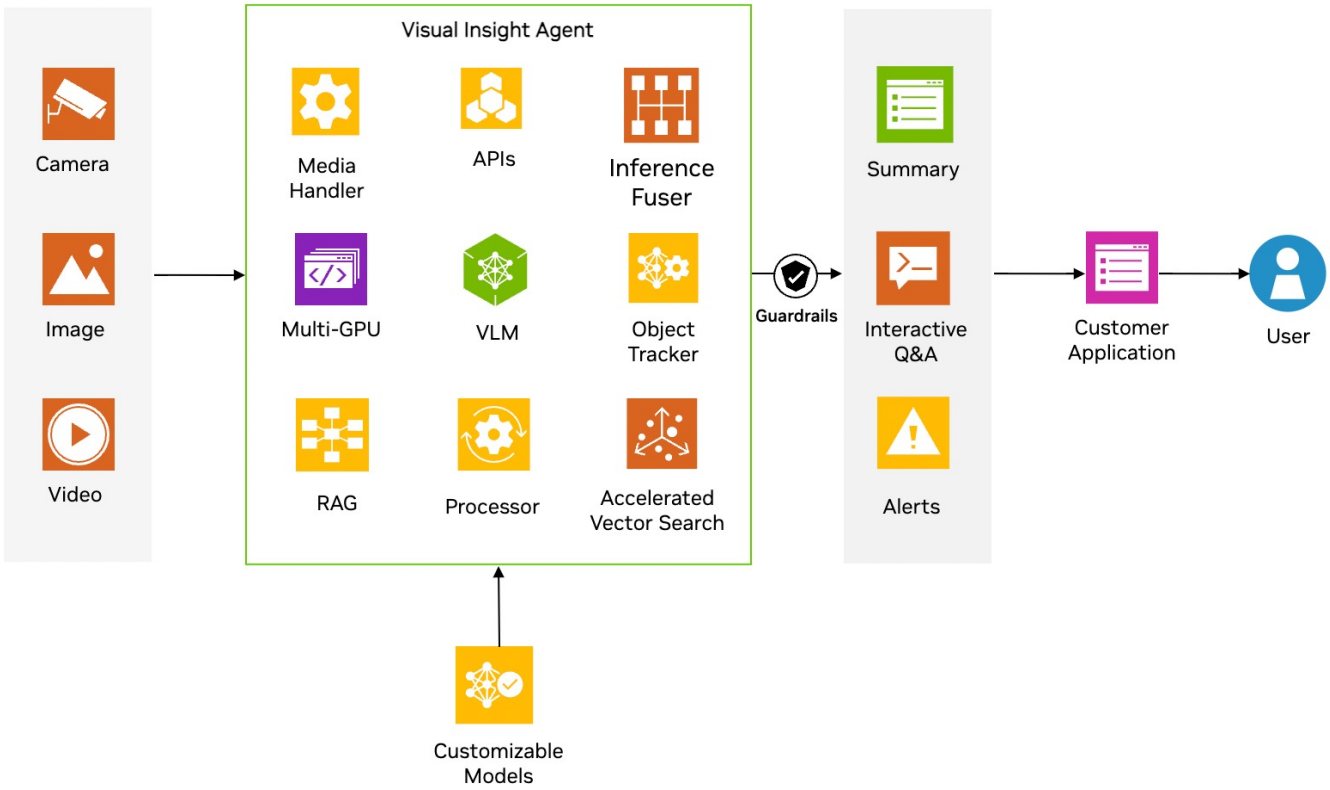
Video Search
Highlights, proactive alerts

Video/Image Summarization
Complex reasoning

NVIDIA

Introducing Visual Insight Agent

AI workflows to build Vision AI agents with VLMs



Build next wave of Vision AI agents using VIA

Leverage VIA AI workflows with disaggregated, accelerated inference for insights using VLMs

Build your AI agent using VIA Workflows!

Customize VLM with your data
— Optimize with TRT-LLM

Microservice for ingestion,
data disaggregation, scheduling

Microservice for inferencing,
object tracking.

Microservice for inferred multi-modal
data, metadata

Accelerated vector search
Spatio-temporal understanding

Accelerate inferencing on large amounts of videos, images and live streams!

High performance — Parallelism
techniques

Latest chunking algorithm
— Aggregation of inference

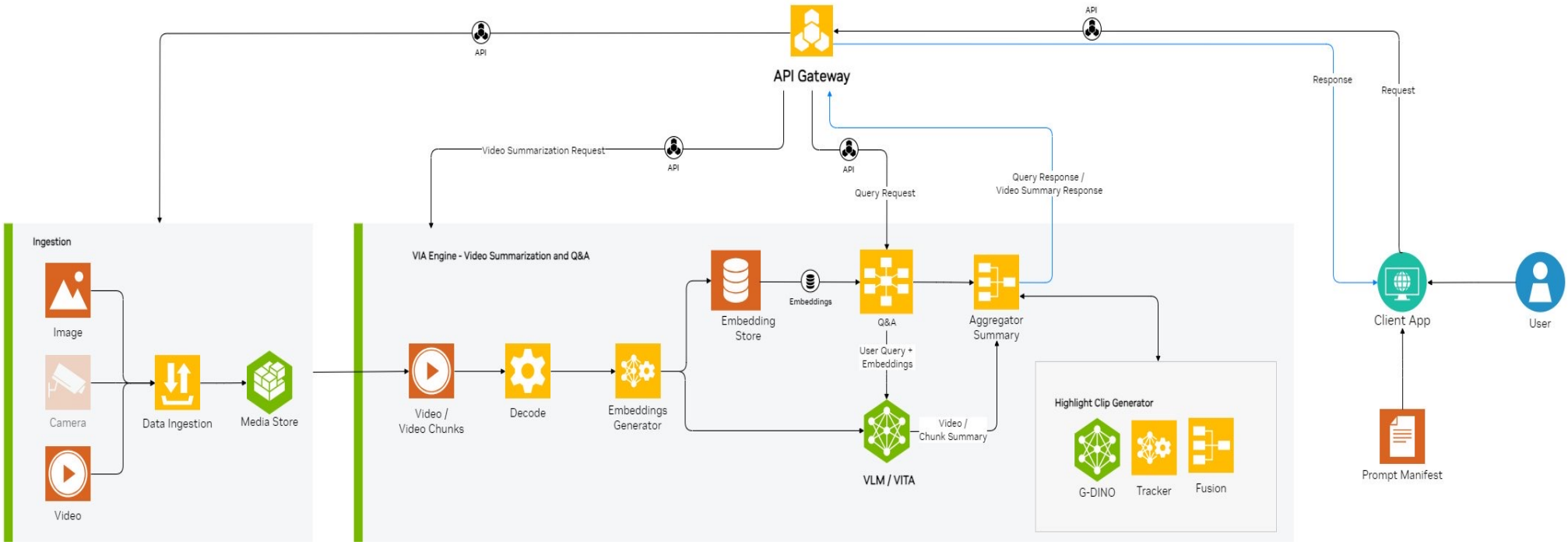
Multi-GPU multi-node
disaggregation & acceleration

Low latency — High throughput
Intra/inter-node communication

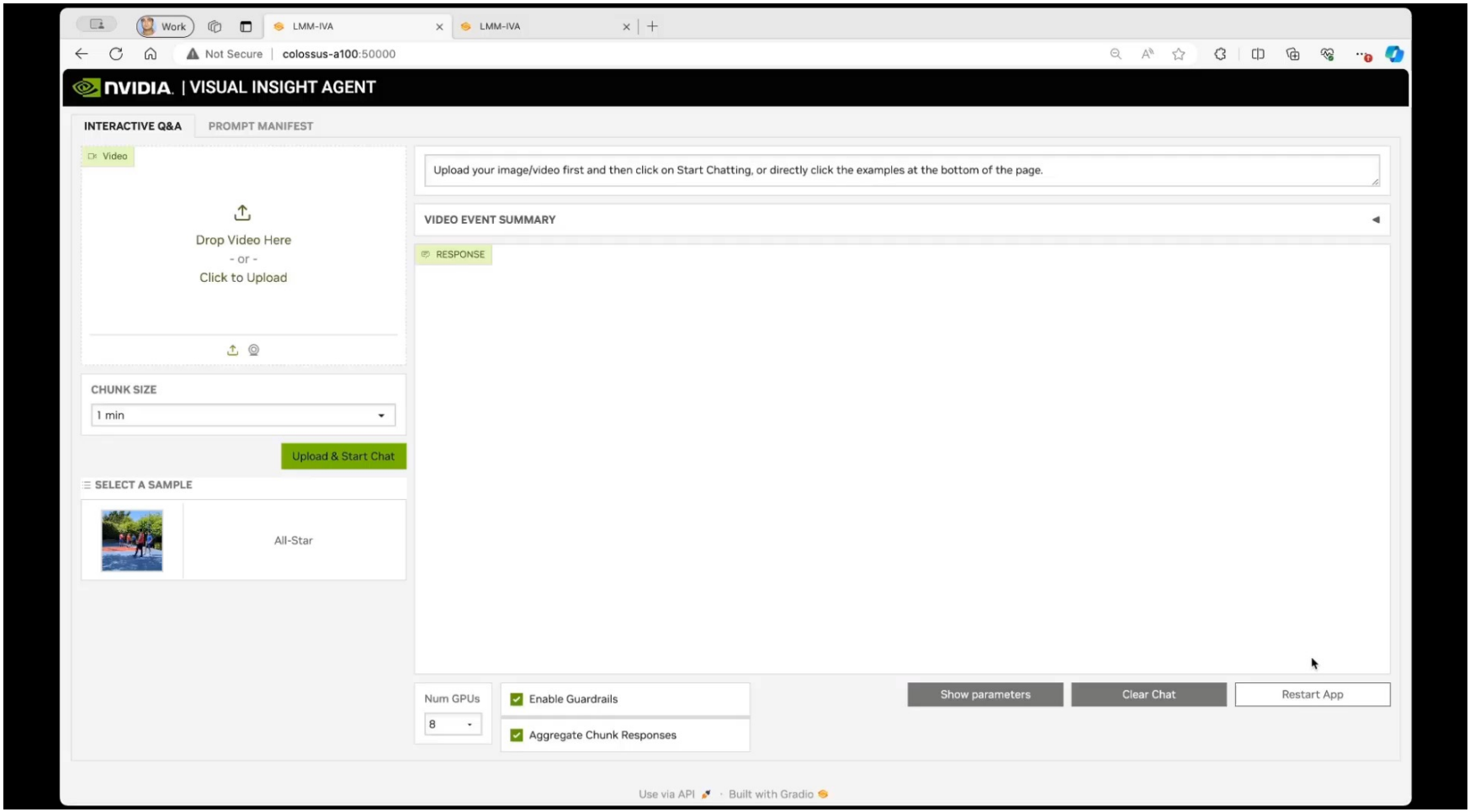
Create AI agent with VLM, RAG,
Guardrail, Analytics & more!

VIA – Workflows for building Vision AI agents

VIA Architecture

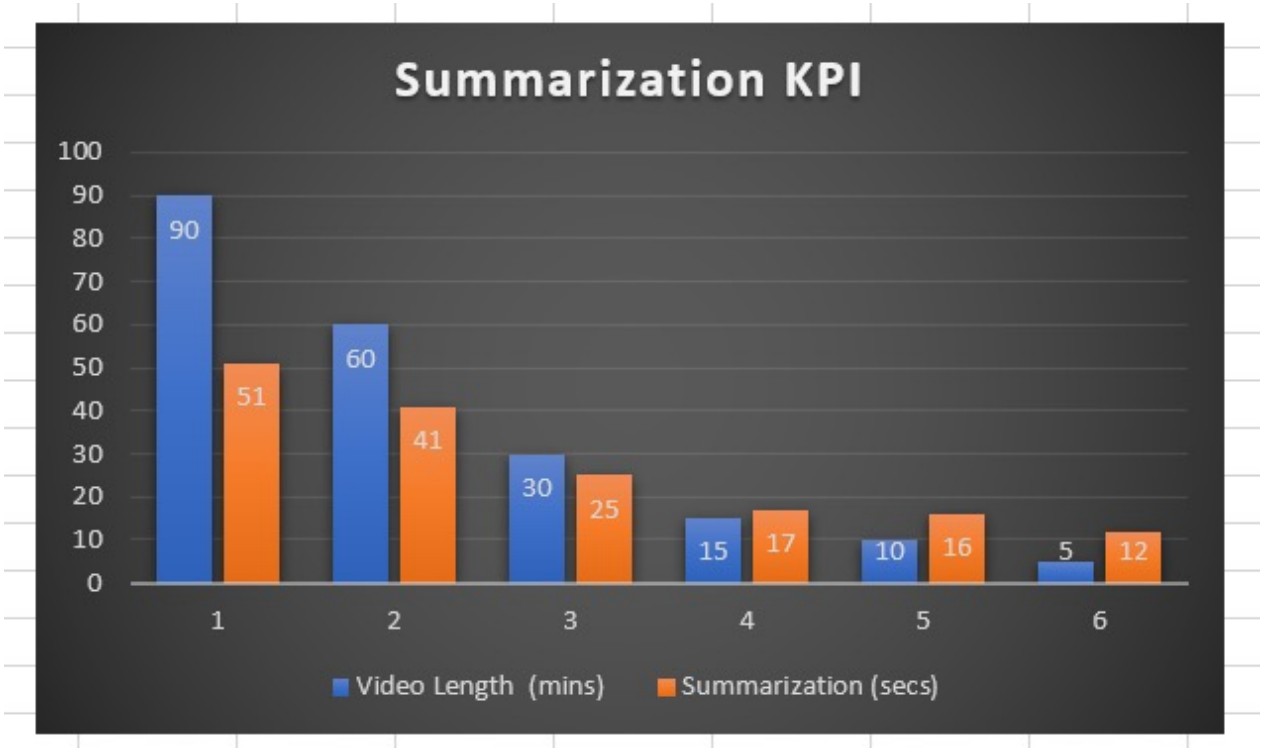


Summarization Workflow Demo

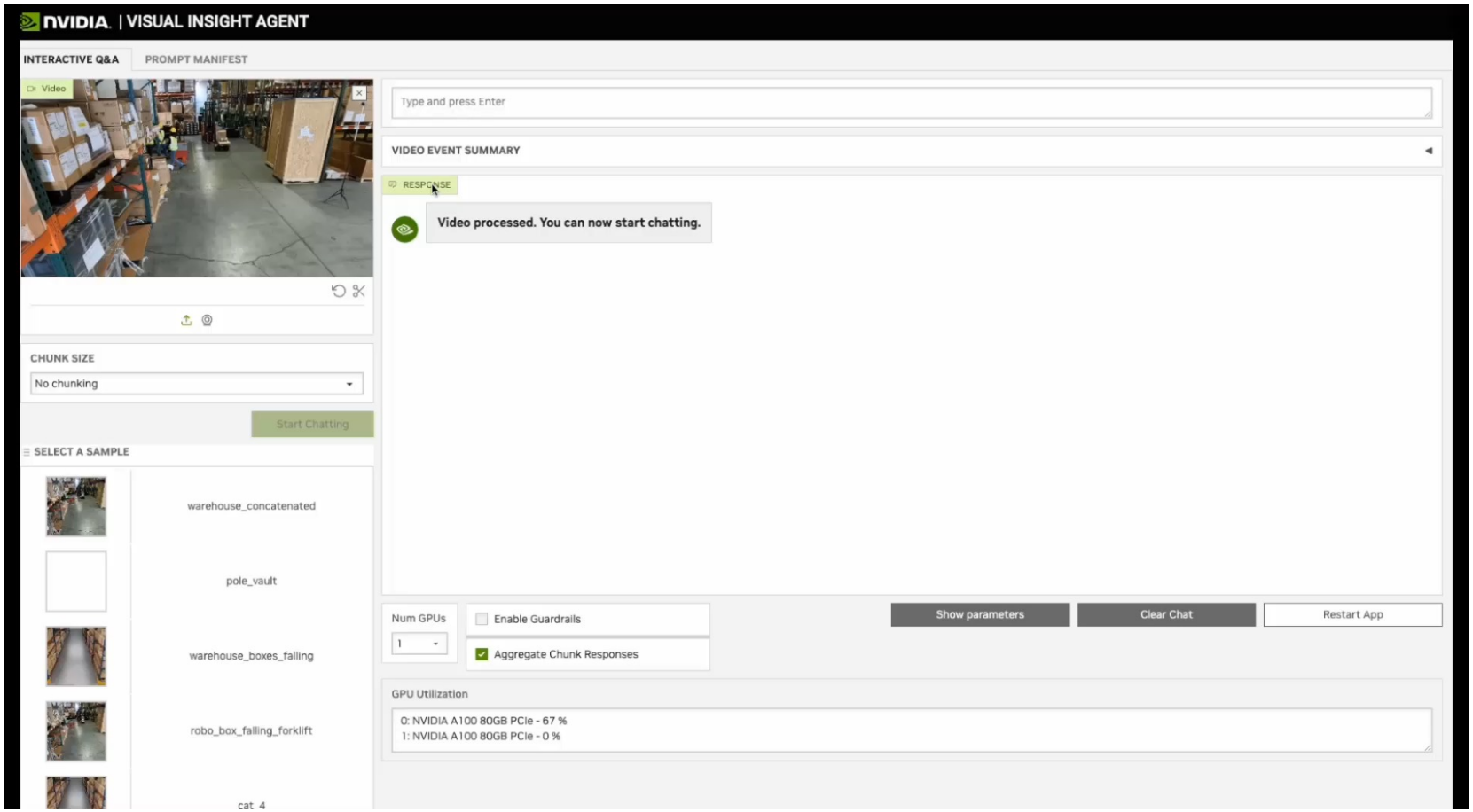


Performance on Long Video

End to End



Interactive Q&A Workflow Demo



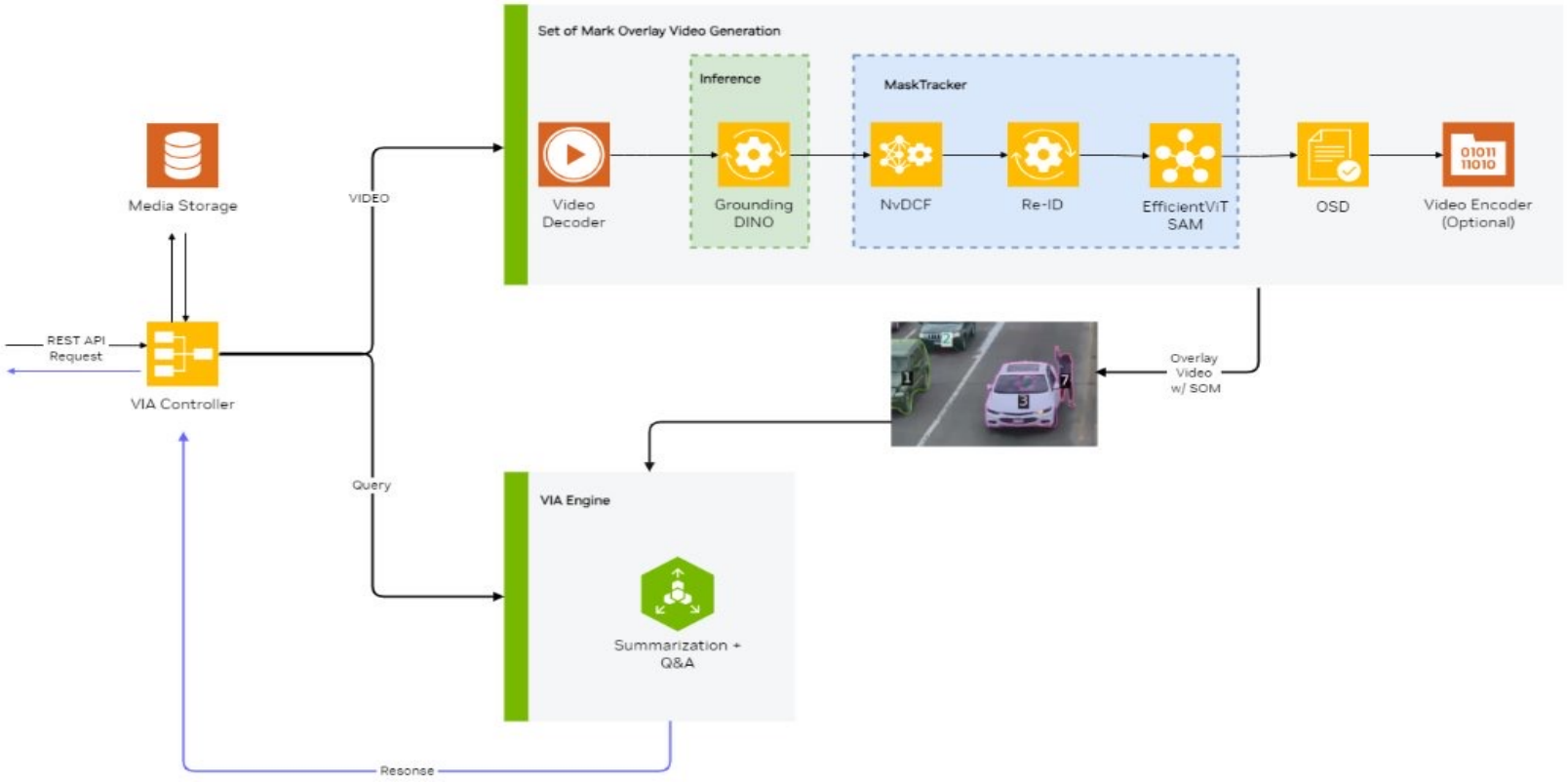
Spatial Grounding on Video

Key Challenges

- Accuracy Challenges - To help VLM to better understand and describe the scene in the video with localization info.
- Persistent ID over many frames throughout the entire video.
- Performance - spatial grounding is expensive along segment anything for every frame, especially for long Videos

Interactive Q&A with Spatial Grounding on Video

Workflow Architecture



Improving Spatial and Contextual Awareness using SoM Prompting

AI framework for Image Insights - Grounding DiNO + Efficient SAM with ViLA

Prompt

Person

Grounding-DINO

EfficientViT SAM

Prompt

There are 1-8 numeric IDs that are person class. Which numeric ID is in most danger in this image

VILA

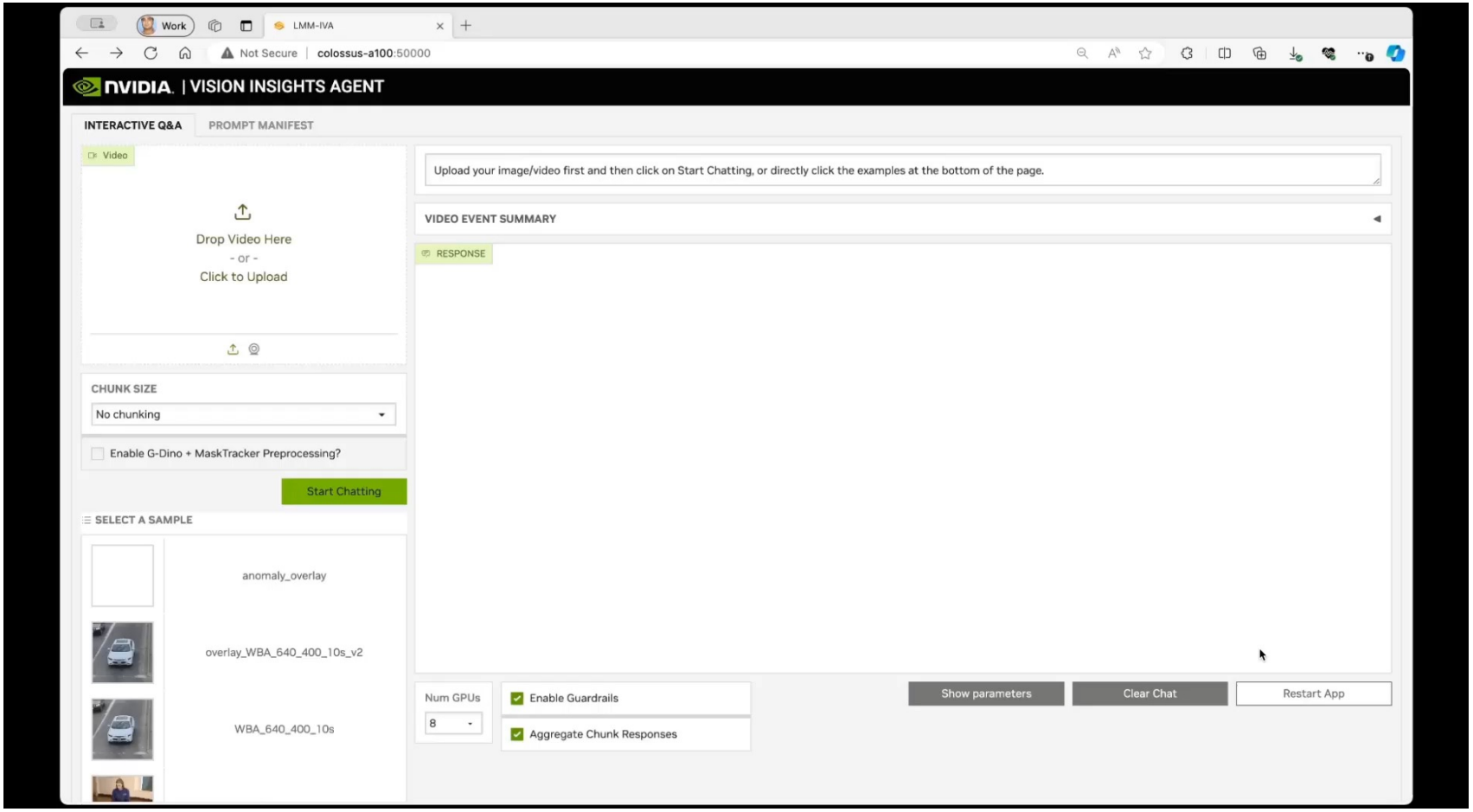
Answer

The firefighter wearing the number 4 is in the most danger in this image. He is standing closest to the fire, which is actively burning and producing thick smoke. This puts him at a higher risk of being affected by the heat, smoke, and potential hazards associated with the fire.

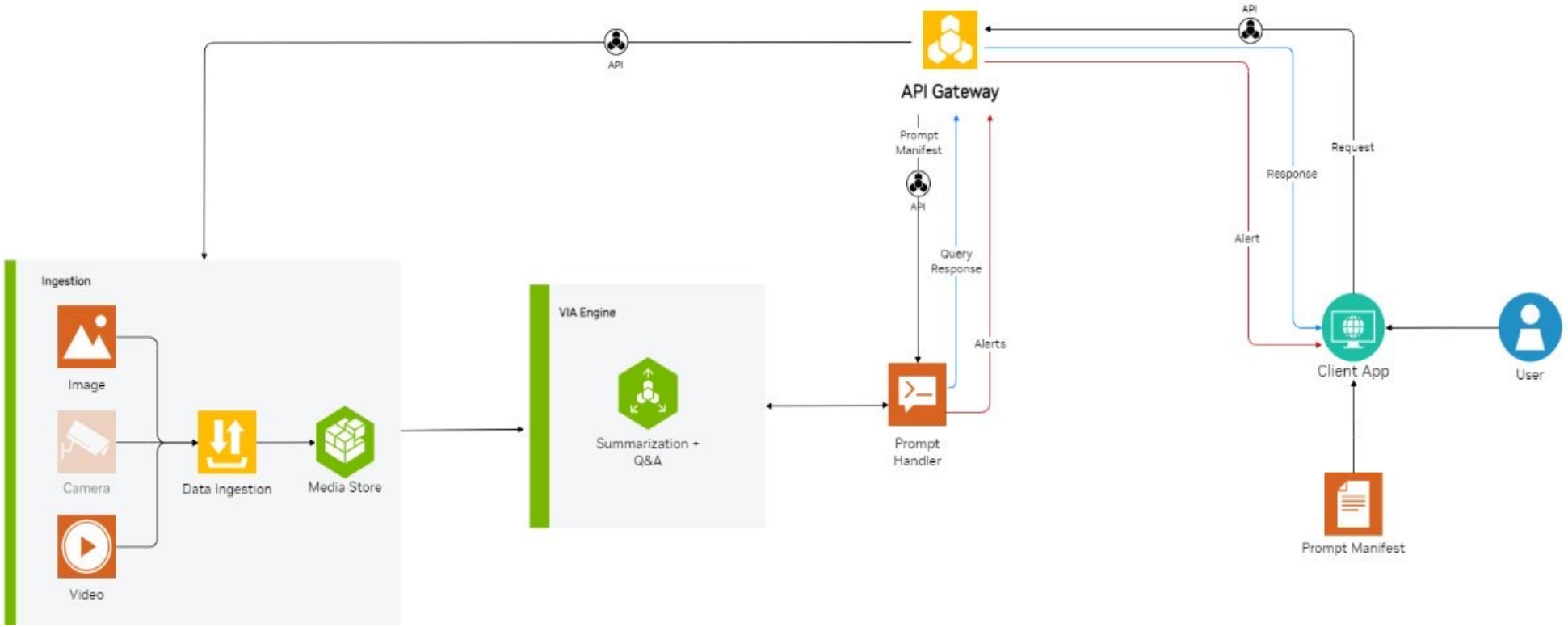
SoM - Set of Mark Visual Prompting

VIA Workflow with Spatial Grounding on Video

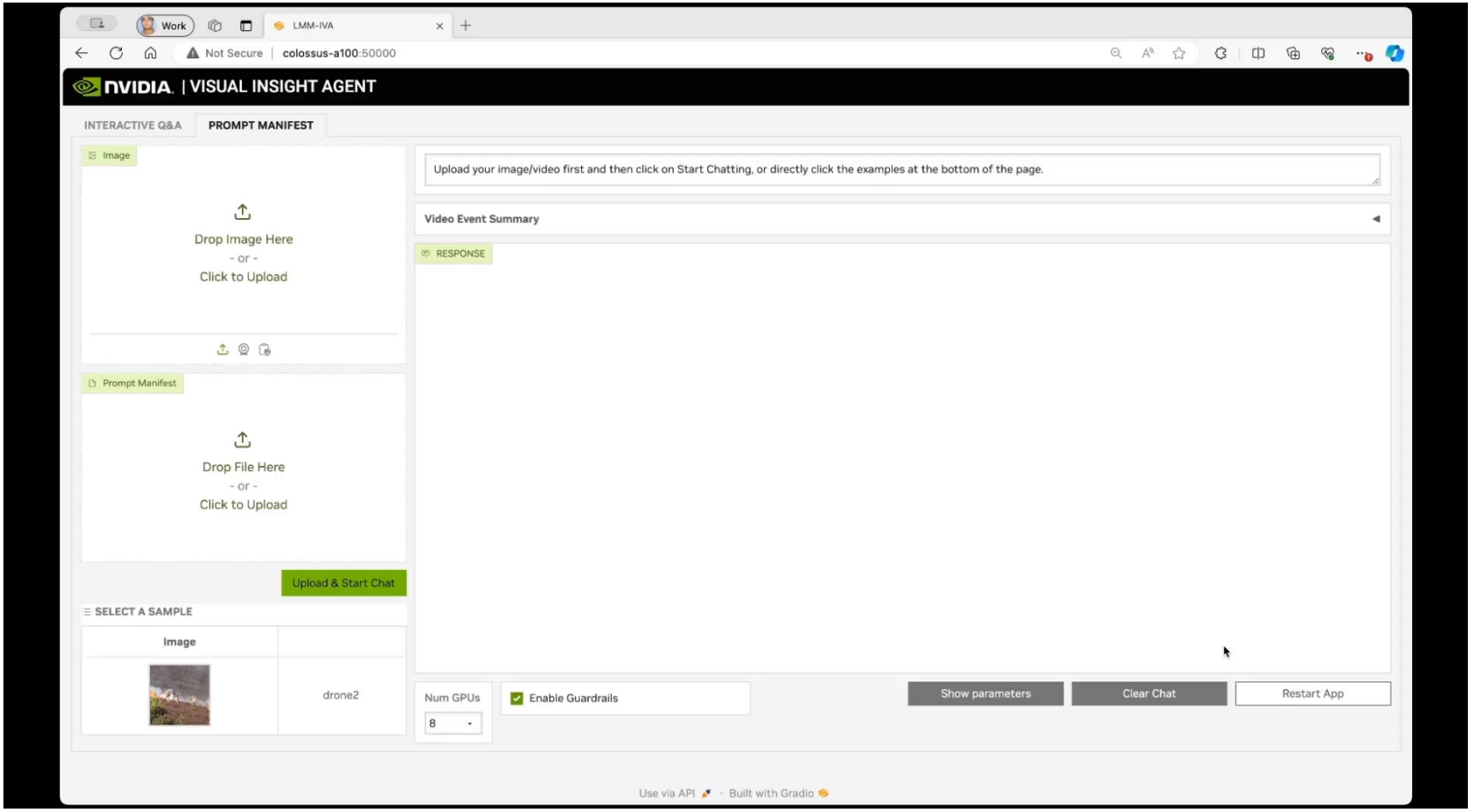
Video Demo



Autonomous Alerts Workflow Architecture



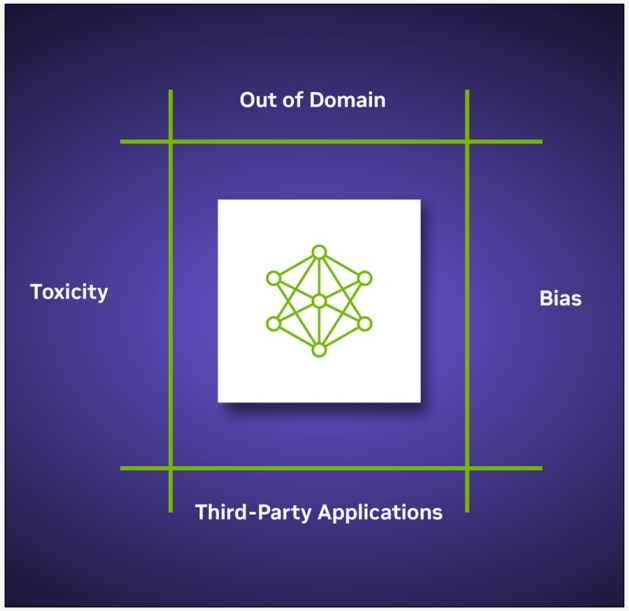
Autonomous Alerts Workflow Demo

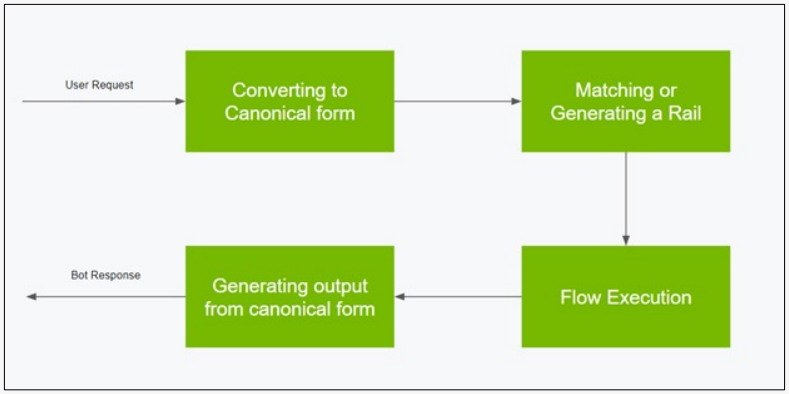


Mitigating Hallucinations with Guardrails

Building trustworthy, safe, and secure LLM conversational systems

- When LLMs are not supplied with factual actual information, they hallucinate and provide faulty, but convincing responses.
- RAG reduces the likelihood of hallucinations by providing relevant facts
- NeMo Guardrails helps enterprises keep applications built on LLM aligned with their safety and security requirements.
- Today, NeMo Guardrails supports three broad categories of guardrails:
  - Topical
  - Safety
  - Security



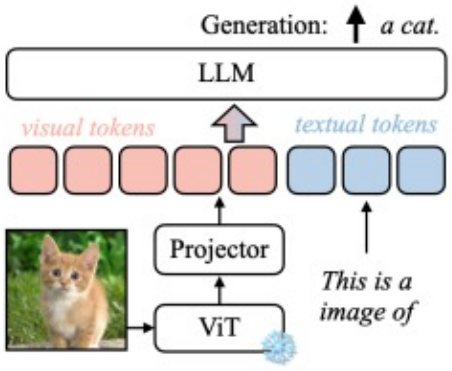


Vision-Language Models

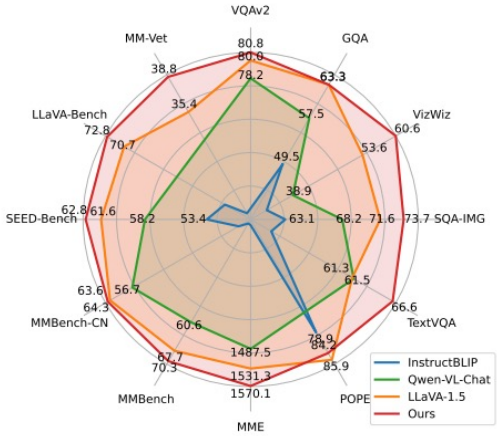
Vision-Language Model for Images

ViLA (Visual Language Assistant)

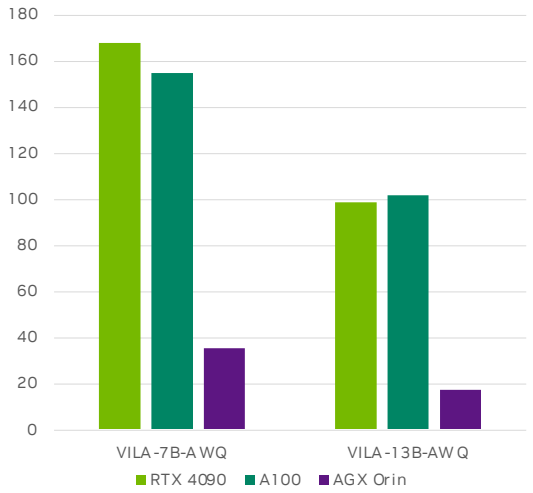
Augmenting LLMs With Visual Input



SOTA Accuracy for Visual Language



Optimizing Inference With 4-bit AWQ



https://arxiv.org/pdf/2312.07533.pdf

ViLA Capabilities

Pred: Home to the greatest pizza.

prediction


Home to the best burgers and fried chicken.


Home to unbeatable fish and chips.


Home to outstanding ramen.

context

In-Context Learning

Q: Can the vehicle proceed through the traffic now?

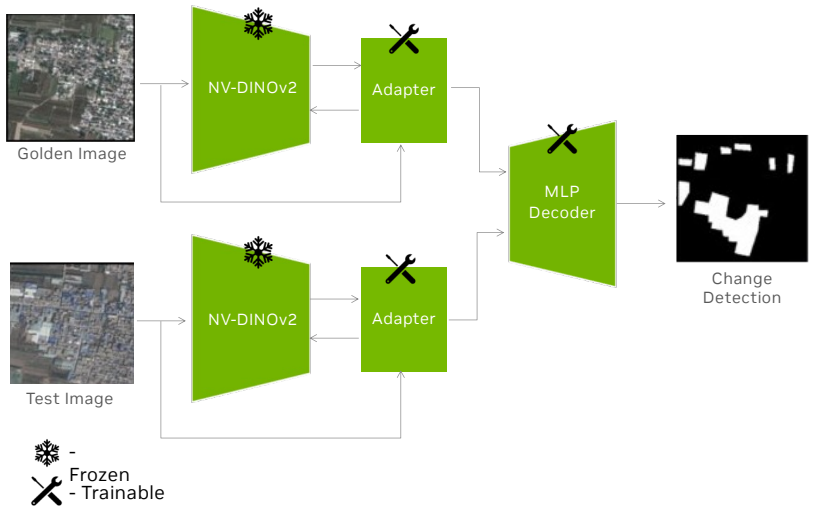
A: Based on the image, the vehicle cannot proceed through the traffic yet. There are multiple people and bicycles in the crosswalk, and the traffic light is red. The vehicle must wait for the traffic light to turn green before proceeding.

Zero-shot Generalization and Complex Reasoning

Prompt: <img>\nI am getting just two custom pizza for me and my friend. How much should I pay in total according to the image + ***Let's think step-by-step***.
Answer: The image shows a chalkboard menu with a special for today, which is to create your own pizza. The price for this special is $9.99. Since you are getting two custom pizzas, you should pay $9.99 x 2 = $19.98 for the two pizzas. So, the answer is $19.98.

Chain of Thought

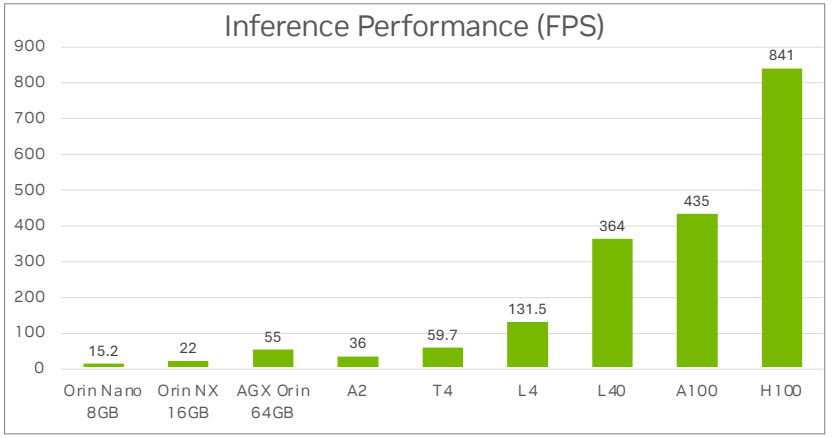
ViLA for Change Detection



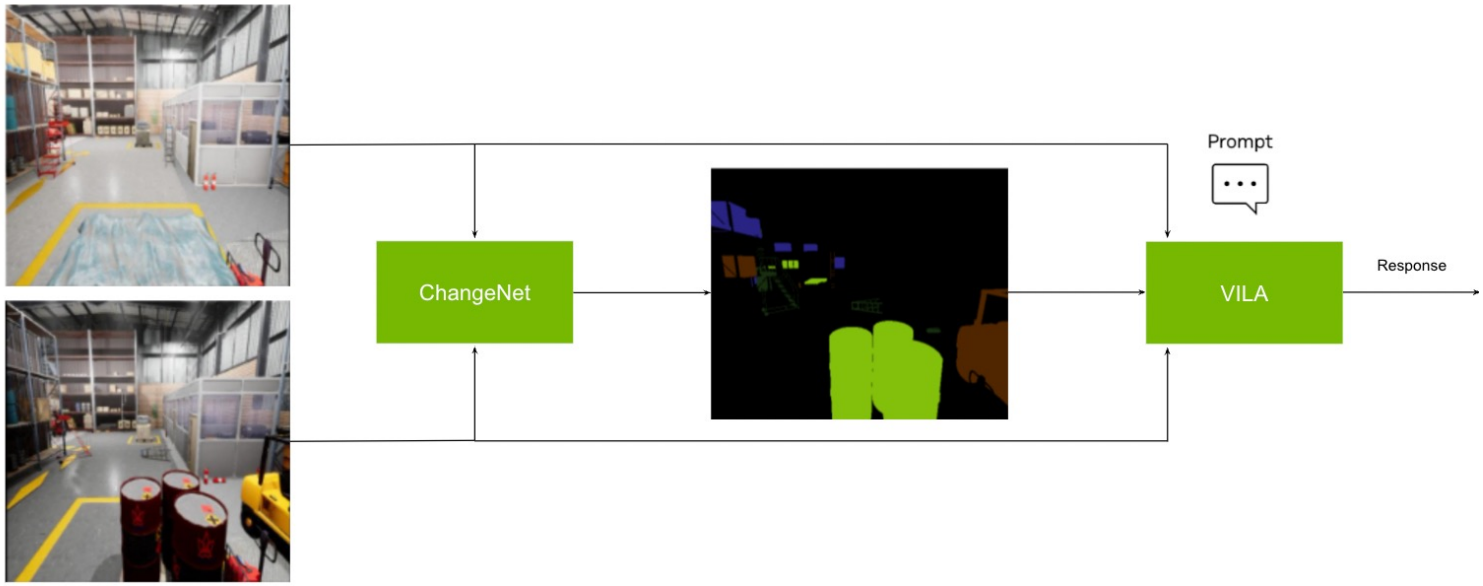
Fine-Tuning Configuration

```
model:
  # pretrained: null
  backbone:
    type: "vit_large_nvdinov2"
    pretrained_backbone_path: /path/to-nvdino.ckpt
    freeze_backbone: True
  decode_head:
    feature_strides: [4, 8, 16, 32]
```

Dataset	Model	Accuracy
Multi-class Satellite Imagery	Multi-class Segmentation	98%
Industrial Defect Detection	Classification	100%
Indoor Warehouse Scene	Multi-class Segmentation	93.5%
Outdoor scene - Binary	Binary Segmentation	98%



ViLA for Change Detection

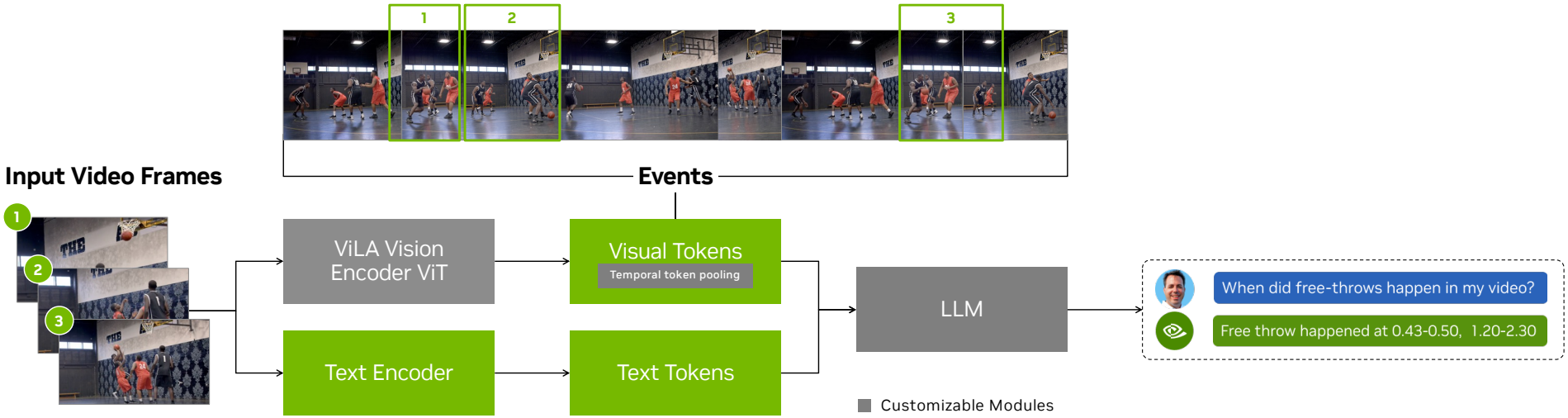


The warehouse has undergone significant changes.

The most noticeable transformation is the addition of a large number of barrels.

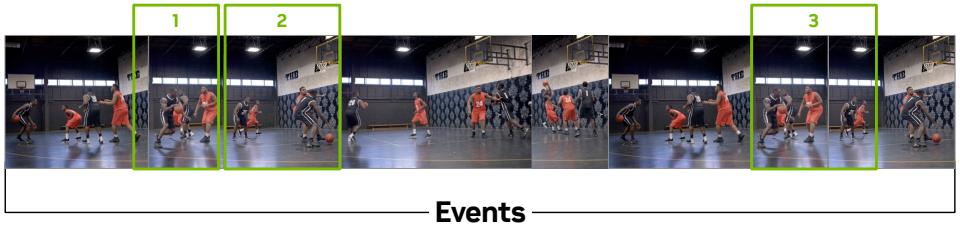
Vision-Language Models for Videos

LiTA (Language-instructed Temporal Assistant)



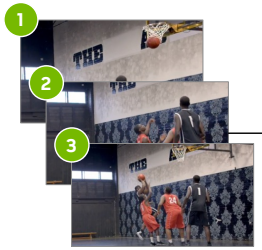
Capabilities		
• Spatial & temporal localization	• OCR	• Interactive Q&A
• Complex reasoning	• Inference metadata out	• Timestamps, events, sequence
	• Video summarization	

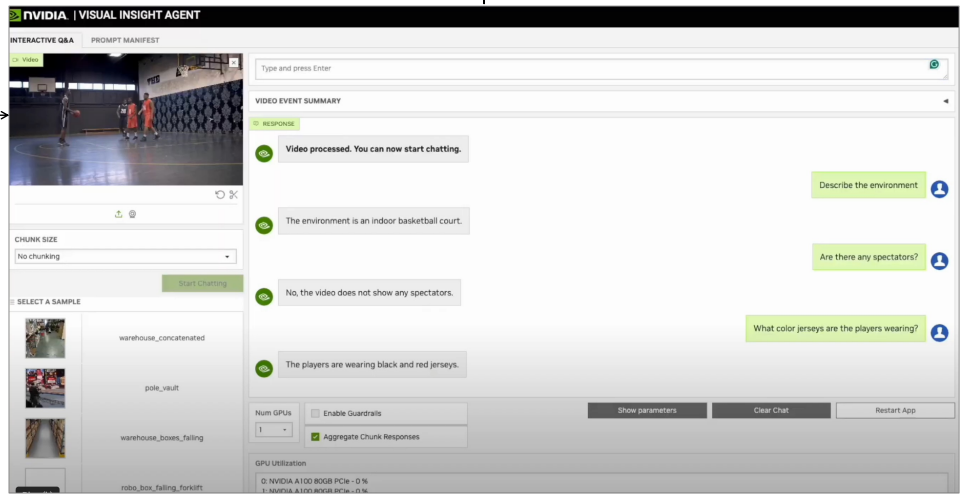
LiTA Capabilities



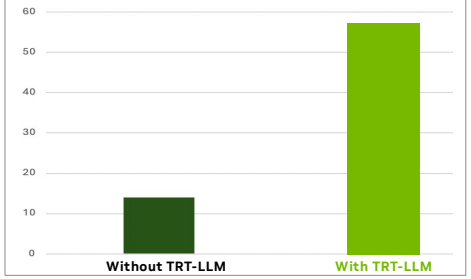
Events

Input Video Frames

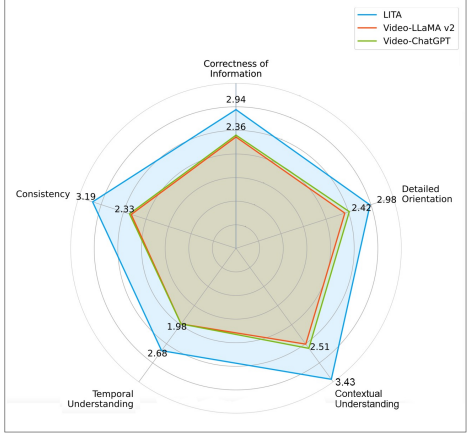




LiTA TRT-LLM Perf with TRT-LLM

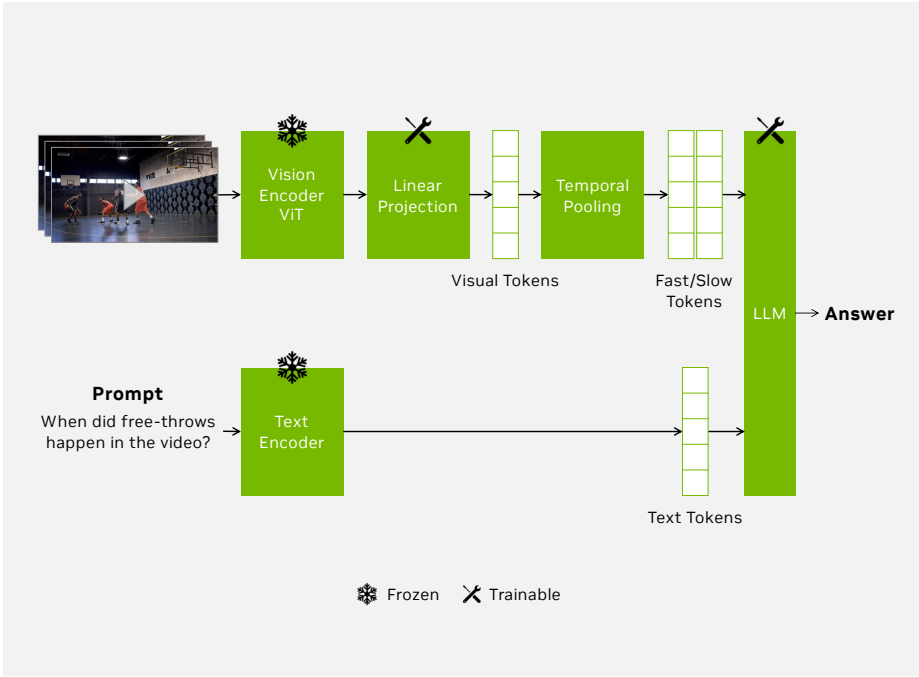


LiTA SOTA Vision-LLM Benchmark



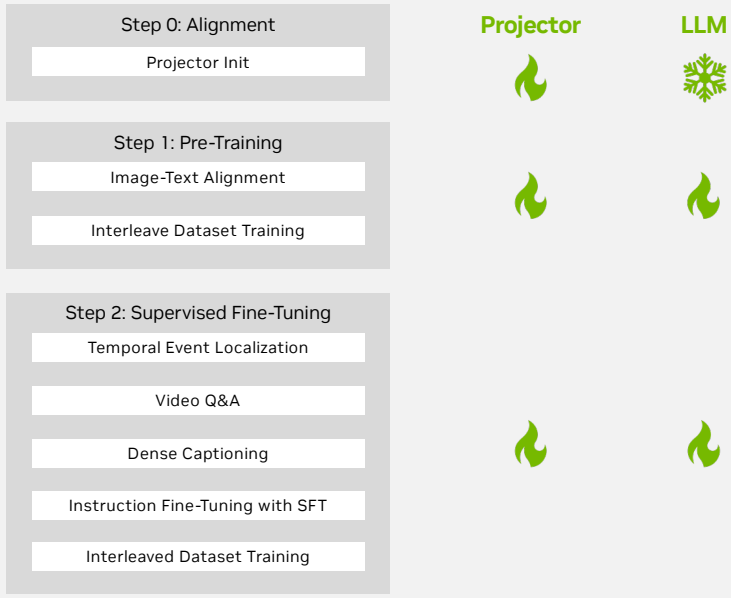
NVIDIA

Fine-Tuning LITA for Custom Use Case



Supervised Fine-Tuning with LoRA

Model Size: 13.3B parameters



Training Workflow

Trainable Parameters: 50M (with LoRA)

*LoRA: Low-Rank Adaptation of LLMs https://arxiv.org/pdf/2106.09685.pdf

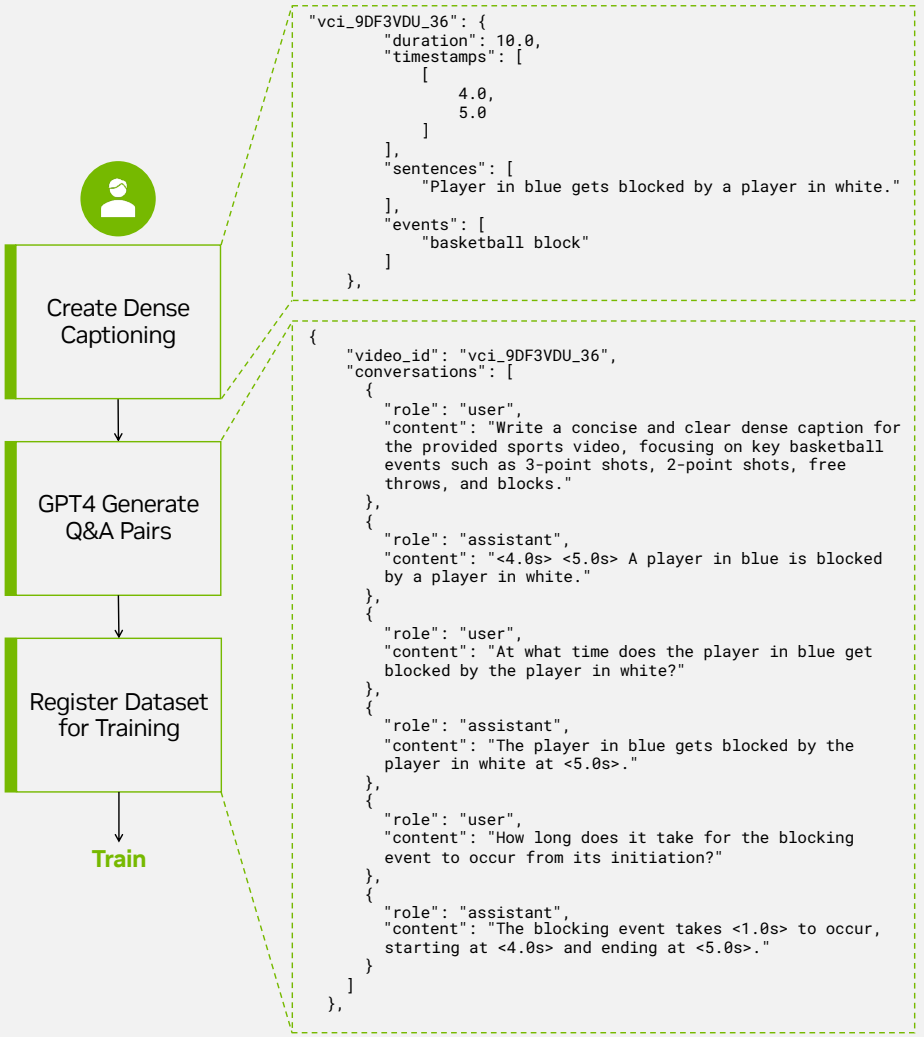
Steps to Customize LiTA

Dataset Recommendation

- 1000 videos for good fine-tuning
- 1 – 2 min video clips
- Events between 1% – 10% of clip duration

Training KPIs

- Sports (Basketball) action dataset
- Trained ~1K videos in 16 hours on 8xA100



After Fine-Tuning LiTA

Example: basketball use case

Question: What type of shot is this?
Original model: This is a basketball shot.
Fine-tuned model: This is a 2-pointer shot from the short corner.

Challenges with finetuning VLM for temporal localization

- Dynamic input number of frames
- Long context video
- Improving Spatial Reasoning
- Model resolution for OCR
- Improving temporal localization
- Adding Audio modality
- Fine-grained temporal localization data
- Inference performance

Apply for Early Access

https://developer.nvidia.com/visual-insight-agent-early-access

VIA workflows will be offered:

- To ecosystem ISVs, SDPs and enterprises partners in various verticals
- With workflows illustrating the ability to customize its microservices and build AI agents.

VIA will:

- Have AI workflows that include microservices to accelerate building next wave of vision AI agents
- Enable workflows for video summarization, interactive Q&A and automated alerts by leveraging VLMs

Leverage VIA AI workflows:

- Finetune and customize large multi-modal models
- Vision AI foundation models
- GPU training performance
- Acceleration in ingestion and streaming of multimedia
- Cloud-native deployment at enterprise scale